Pierre Marquis
Odile Papini
Henri Prade
Editors

A Guided Tour of Artificial Intelligence Research

1 Knowledge Representation, Reasoning and Learning

Springer

A Guided Tour of Artificial Intelligence Research

Pierre Marquis · Odile Papini · Henri Prade
Editors

A Guided Tour of Artificial Intelligence Research

Volume I: Knowledge Representation, Reasoning and Learning

Springer

Editors
Pierre Marquis
CRIL-CNRS, Université d'Artois and
Institut Universitaire de France
Lens, France

Odile Papini
Aix Marseille Université, Université de
Toulon, CNRS, LIS
Marseille, France

Henri Prade
IRIT
CNRS and Université Paul Sabatier
Toulouse, France

ISBN 978-3-030-06163-0 ISBN 978-3-030-06164-7 (eBook)
https://doi.org/10.1007/978-3-030-06164-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland


PREFERENCES
MODALITIES
DECISION
FUZZY
UNCERTAINTY
CASES
DEFAULTS
NEURAL NET
REINFORCEMENT
REASONING
SPACE
ACTION
NORMS
DATA
CAUSALITY
ANALOGY
KNOWLEDGE
TIME
EXPLANATION
PROBABILITY
BELIEF
STATISTICS
LEARNING
LOGIC
ARGUMENTATION
POSSIBILITY
Emilie Prade

General Presentation of the Guided Tour of Artificial Intelligence Research

Artificial Intelligence (AI) is more than sixty years old. It has a singular position in the vast fields of computer science and engineering. Though AI is nowadays largely acknowledged for various developments and a number of impressive applications, its scientific methods, contributions, and tools remain unknown to a large extent, even in the computer science community. Notwithstanding introductory monographs, there do not exist treatises offering a detailed, up-to-date, yet organized overview of the whole range of AI researches. This is why it was important to review the achievements and take stock of the recent AI works at the international level. This is the main goal of this *A Guided Tour of Artificial Intelligence Research*.

This set of books is a fully revised and substantially expanded version, of a panorama of AI research previously published in French (by Cépaduès, Toulouse, France, in 2014), with a number of entirely new or renewed chapters. For such a huge enterprise, we have largely benefited the support and expertise of the French AI research community, as well as of colleagues from other countries. We heartily thank all the contributors for their commitments and works, without which this quite special venture would not have come to an end.

Each chapter is written by one or several specialist(s) of the area considered. This treatise is organized into three volumes: The first volume gathers twenty-three chapters dealing with the foundations of knowledge representation and reasoning formalization including decision and learning; the second volume offers an algorithm-oriented view of AI, in fourteen chapters; the third volume, in sixteen chapters, proposes overviews of a large number of research fields that are in relation to AI at the methodological or at the applicative levels.

Although each chapter can be read independently from the others, many cross-references between chapters together with a global index facilitate a nonlinear reading of the volumes. In any case, we hope that readers will enjoy browsing the proposed surveys, and that some chapters will tease their curiosity and stimulate their creativity.

July 2018

Pierre Marquis
Odile Papini
Henri Prade

Contents

Preface: Knowledge Representation, Reasoning and Learning

Artificial Intelligence (AI) aims to provide machines with abilities to perform "intelligent" tasks in the sense that they are considered as such by humans. These tasks take advantage of pieces of information of different nature. They may be machine- or human-originated, factual or generic, structured or unstructured. Those pieces of information consist of data issued from sensors, observations, rules, pieces of belief, pieces of knowledge, preferences, norms, etc. They are often imperfect: incomplete, imprecise, uncertain, or contradictory. They may involve time, space, as well as multiple agents. They are about a world which can be static or dynamic. All those pieces of information must be acquired, learnt, updated. They fuel various inference and decision machineries. Modeling and representing the available information in suitable settings, and formalizing learning, reasoning, and decision processes are fundamental issues in AI, as acknowledged in the foreword of this volume.

This first volume of this guided tour of AI research focuses on these issues and aims to present in 23 chapters the main approaches to knowledge representation, reasoning, and learning developed in AI.

First dreamed before being envisioned, AI was not born ex nihilo. A historical perspective (Chapter "Elements for a History of Artificial Intelligence") sketches a panorama, from the Antiquity to the 1980s, of both the dreams and genuine contributions coming from different scientific fields or cultural productions, and that led to the final emergence of AI.

The limitations of classical logic for belief and knowledge representation have motivated the introduction of new logical formalisms, which are presented in Chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning". Chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" first addresses the issues of imprecision, uncertainty, gradualness, and granularity and then contrasts probability and possibility theories. Chapter "Representations of Uncertainty in AI: Beyond Probability and Possibility" presents extensions of these two frameworks by focusing on evidence theory based on belief functions and on imprecise probabilities. Chapter "Qualitative Reasoning" deals with time and space. The representation of ontologies in the setting of

description logics is the topic of Chapter "Reasoning with Ontologies". Preference representation and norm representation are respectively dealt with in Chapter "Compact Representation of Preferences" and in Chapter "Norms and Deontic Logic". Chapter "A Glance at Causality Theories for Artificial Intelligence" discusses the handling of causality in different approaches. Case-based reasoning, interpolative reasoning, and analogical reasoning are presented in Chapter "Case-Based Reasoning, Analogy, and Interpolation".

The next two chapters are devoted to formal models for machine learning, more precisely *Statistical Computational Learning* (Chapter "Statistical Computational Learning") and *Reinforcement Learning* (Chapter "Reinforcement Learning").

Specific chapters are dedicated to reasoning in the presence of contradiction (Chapter "Argumentation and Inconsistency-Tolerant Reasoning"), and reasoning with several sources of information: revising and merging (Chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion"), as well as updating (Chapter "Reasoning About Action and Change").

Several chapters focus on decision making: multicriteria decision (Chapter "Multicriteria Decision Making"), *Decision Under Uncertainty* (Chapter "Decision Under Uncertainty"), and collective decision (Chapter "Collective Decision Making").

The two next chapters respectively deal with agents' cognitive aspects, such as trust or emotion (Chapter "Formalization of Cognitive-Agent Systems, Trust, and Emotions") and with multiagent systems and interactions (Chapter "Negotiation and Persuasion Among Agents"). Diagnosis on the one hand and validation and explanation on the other hand are surveyed respectively in Chapter "Diagnosis and Supervision: Model-Based Approaches" and Chapter "Validation and Explanation". The last chapter of the volume mainly addresses knowledge engineering concerns. The afterword emphasizes the importance of supplementing inference systems with trust information.

Lens, France — Pierre Marquis
Marseille, France — Odile Papini
Toulouse, France — Henri Prade

Foreword: Knowledge Representation and Formalization of Reasoning

The first volume of the Artificial Intelligence (AI) guided tour describes how one can enable a computer system to reason. As this book is a guided tour, it considers the many topics developed by AI researchers. Naturally, this is essential for all those who want to make progress in our field; however, this does not imply that reading these books should be restricted to AI scientists. In fact, they do not describe a succession of achievements, but the general principles that allow the realization of the most noteworthy results. This overall perspective gives useful ideas to all scientists, although they do not consider themselves as belonging in the mainstream of Artificial Intelligence, even if they do not try to solve problems on a computer.

AI is interested in its main goal: solving problems that were previously only solved by living beings, and particularly by human ones. However, it turns out that, while doing so, AI has often discovered methods and ideas useful for other scientific disciplines. Thus, in several domains of Cognitive Science, new approaches came from ideas widely used in AI: For instance, apart from the statistical description of the life of an anthill, some have begun to model every ant, considered as a small automaton. In this book, cognitive scientists will certainly be interested in the study of trust and emotions for a cognitive agent. For their part, philosophers and logicians will read the chapter on deontic logics, which specify in a rigorous, unadorned language, concepts such as obligatory, permissible, optional, and ought. They could also see how it is possible to reason even when there is contradictory information. The discovery of a contradiction is not always a total disaster, since it can come from a small amount of information, where one can clean up the mess. Belief revision can restore the consistency of knowledge when new data are inconsistent with what was already known.

Economists and sociologists will look with interest at the methods for collective decision, where several agents must cooperate to find a common decision; these methods are very often helpful, for choosing the president of a political party as well as for the choice of a restaurant by a group of friends. For making such decisions, IA researchers have experimented with various kinds of votes and auctions; they defined equity in the sharing of resources. They have found methods and

concepts useful when there are several artificial agents, but these results can also be used when a group of humans is taking a collective decision. However, I believe that computer specialists will be the most interested community: For them, nearly all the chapters of this book will be very useful. Indeed, the distinction between computer and AI scientists is often very tenuous: AI researchers become computer specialists when they implement a system that must obtain excellent results. Conversely, without realizing it, computer specialists may use AI methods when they are developing a system for solving a particular problem: It is natural to wonder which methods a human being is using for solving it, and to implement them in the system.

Several areas of computer science have been developed for the first time by AI researchers: We consider them because human beings manage to obtain good results when they use them. For us, the natural approach is to ask: Why can't a computer program do the same? For other people, it seems impossible or too difficult for the present state of the art. Being the first to consider a problem, we cannot be prevented to find new methods for solving it! It happens that these methods may be useful in new applications, and the computer scientists should put them in their tool box, so that they will think about using them when an opportunity presents itself.

In particular, the beginning of this volume title mentions an important problem for computer specialists: "Knowledge Representation." Several aspects of this issue are considered, such as the representation of preferences and of uncertainty. Ontologies are very important for semantic Web applications: They provide a formal knowledge representation for their domain, and they allow their management, acquisition, retrieval, etc. Knowledge engineering gives methods for finding knowledge for a particular problem, especially by a collaboration with human experts. These chapters can help a computer scientist to find ideas for developing future systems.

The second part of the title is "Formalization of Reasoning." This capacity is important: If a system can reason, it is more general. Indeed, it is no longer necessary to anticipate all the possible situations: By reasoning, a system can automatically find the right action in an unexpected situation. General systems offer a dual advantage: Firstly, fewer programs must be written; a general system does the job of several specific systems. Secondly, results are obtained far more quickly: A general system may be adapted to a new application without the need for writing more programs. Many kinds of reasoning are presented; I will choose some of them and show that they are important and go beyond AI.

A powerful AI method is reasoning about action and change: In examining the changes that should be made in the present state for reaching the target, one finds the actions that are more likely to succeed. To do that, one models the behavior of a human expert, and expert systems have often led to useful results. A by-product of this approach is to help human specialists to improve their own way for solving these problems. Such knowledge is given in a declarative form, regardless of how it is used; in this way, it is easier to understand or modify it. This allows a high level of generality: This knowledge can be used in various contexts, even when they were

not known by the expert. Moreover, such a system can explain its results: It indicates the steps of its reasoning. If we agree on the rules, one must agree in the result. This is essential for the users: They want to make sure that a surprising decision is not the consequence of a bug.

Furthermore, it is often necessary to take a multicriteria decision, depending upon numerous factors, where several objectives must be simultaneously met. This is very difficult; usually, there is no perfect solution: One must find a sensible compromise between the requirements. The situation is particularly sensitive when the consequences of a decision depend on events that did not even take place when this decision must be taken. This happens, for example, when a medical doctor, in a serious and urgent situation, must prescribe a treatment without the test results. In particular, diagnostic, for a disease as well as for a mechanical failure, is an important kind of decision. To do that, one must use a model to the system at fault in order to understand why an unwanted event happened.

It is very difficult to find knowledge useful for solving a problem; therefore, AI has always been interested in learning. Many methods have been successfully experimented; among them, statistical learning discovers regularities in the domain and uses them for building an efficient solving method. On the contrary, case base learning compares the present situation with a similar one already experienced; it tries to adapt the previous solution to the new problem. Learning ability is a key step toward the realization of general systems.

In promoting the use of declarative knowledge, will not AI lead to the extinction of computer specialists? On the contrary, reading this book should help them, and not only AI researchers, to manage their domain better: Computer scientists will develop more efficient systems more easily. They will implement them so that knowledge will be easier to give, to understand, to modify, and sometimes will even be learned by the system itself.

Jacques Pitrat
(1934-2019), formerly with
CNRS-LIP6
Université Paris 6
Paris, France

Elements for a History of Artificial Intelligence

Pierre Marquis, Odile Papini and Henri Prade

Abstract Artificial intelligence (AI) is a young scientific field, which like other domains of information processing sciences, was born in the middle of the XXth century, with the arrival of the first computers. However, much more long-standing concerns have contributed to its final emergence. They can be broadly articulated around two main issues: the formalization of reasoning and learning mechanisms and the design of machines having autonomous capabilities in terms of computation and action. Over time, such machines have been first dreamed, before being designed and made real. The progressive achievements have fed the imagination of philosophers, but also writers, movie makers, and other artists. This is the reason why in the few elements of the great historical epic that we sketch here, references to all sectors of human creativity are involved.

P. Marquis
CRIL-CNRS, Université d'Artois and Institut Universitaire de France, Lens, France
e-mail: marquis@cril.univ-artois.fr

O. Papini
Aix Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
e-mail: odile.papini@univ-amu.fr

H. Prade (✉)
IRIT, CNRS and Université Paul Sabatier, Toulouse, France
e-mail: prade@irit.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_1

1 Introduction

Artificial intelligence was not born ex nihilo during a series of meetings in the summer of 1956 within the framework of a research program with this name,[1] involving 10 participants,[2] organized at Dartmouth College (Hanover, New Hampshire, USA). AI has immediate roots in cybernetics (Wiener 1949) and in computer science, but its emergence is the result of a long and slow process in the history of humanity, which can be articulated around two main trends: the formalization of reasoning and learning mechanisms on the one hand and the design of machines having autonomous capabilities in terms of computation and action on the other hand.

There exist a number of valuable books focusing on different aspects of the modern history of AI (Anderson 1964; McCorduck 1979; Rose 1984; Pratt 1987; Kurzweil 1990; Crevier 1993; Nilsson 2010).[3] But what is said about what may be called the "prehistory" of AI (corresponding roughly speaking to the time period before the advent of the first computers), is usually very sketchy and sparse, with the mention of a few names: Aristotle and his *Organon* (Jones 2012) for the (Western) Antiquity, Ramon Llull and his *Ars Magna* (1305) (Fidora and Sierra 2011) for the Middle Ages, Thomas Bayes (1763) and George Boole (1854) for the modern times before the last century.

In order to structure this historical panorama, several main periods can be roughly distinguished: the first one goes from Antiquity to the XVIth century, followed by a transition period toward modernity in the XVIIth century and then in the Age of Enlightenment in the XVIIIth century, before the mathematization of logic in the XIXth century, and then the birth of computer science, from computability theory to cybernetics, in the first half of the XXth century, and finally the development of AI in the second half of the XXth century.

Although various rarely referred works are mentioned, and some which have probably never been related to the origins of AI, this panorama should be considered as a sketch and a draft. The ambition of this chapter is only to provide some (often forgotten or ignored) elements that may be considered as parts of the slow emergence of AI concerns during the last centuries. Further in-depth analysis would be certainly of interest. Even if quite a number of names are cited in the following, it is very likely that names are missing. This is unavoidable with such an attempt.

[1] The application for getting a financial support, written in the summer of 1955, and entitled "A proposal for the Dartmouth summer research project on artificial intelligence" (where the name of the new research area was already coined!), was signed by John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon (McCarthy et al. 2006).

[2] Apart from the 4 signees of the project appearing in the above footnote, they were Trenchard More, Allen Newell, Arthur Samuel, Oliver Selfridge, Herbert A. Simon, and Ray Solomonoff. Interestingly enough, it is worth noticing that these 10 participants were already carriers of the large variety of research directions that can still be observed in AI.

[3] A year between parentheses is at the same time a publication date and indicates a reference to a publication of the author cited just before in the text. Exceptionally for the works from Antiquity to Middle Ages, the year used in the references is one of a modern edition and not the one of the first publication.

Beyond very famous names, most often English or American, the French and German domains are somewhat favored, while undoubtedly other names of equivalent importance from other countries and other languages could be cited. Moreover, we only indicate the general concerns of the authors cited and the references to their work, without discussing their contribution in detail (which would have required a much longer chapter). It should be clear that several works mentioned, which may be a posteriori related to AI, were only small parts of the production of their authors, involved in very different scientific fields. As will be seen below, the pieces of work from which AI emerged were progressively built up over time through a succession of espistemological ruptures, progressively enriched by new scientific knowledge, techniques, and technologies.

2 The First Steps: From Antiquity to the XVIth Century

Discussions about patterns of reasoning may be encountered in all the great philosophies. In the following, we mainly focus on Greek philosophy, but Chinese, Indian, Hebraic, Arab and Persian philosophies should not be ignored. For detailed studies the reader is referred to the works by Chad Hansen (Chen Hang Sheng) (1983), Gillon (2010), Sarukkai (2018), Gabbay and Woods (2004a), Abraham et al. (2010, 2013, 2016), Schumann (2012, 2017), Rescher (1963, 1964), and Akrami (2017).

Antiquity

By the scope and the range of the topics addressed in his *Organon*, which gathers six treatises on logic (Jones 2012), Aristotle (384–322 BC) appears as the father of logic, although other Greek philosophers, like the Stoics also contributed to the first developments of logic (for example, Chrisippus (c. 279–c. 206 BC)). Indeed, Aristotle contributed to the first analysis of human reasoning, namely in stating the different forms of syllogisms and in discussing their validity conditions in the *First Analytics* and the *Second Analytics* (see for example (Gochet et al. 1988, 1989; Blanché 1970)). In *On Interpretation*, he identified the vertices of the logical square of opposition generated from the application of an internal negation and an external negation over a statement. This square, drawn later in a comment of Aristotle's text ascribed (but maybe wrongly attributed) to the Numidian Latin writer and philosopher Apuleius of Madaura (c. 124 c. 170) (Londey and Johanson 1984, 1987; Gombocz 1990), visualises the oppositions between statements corresponding to its four vertices (Parsons 2017). In the *Topics*, Aristotle studied dialectics and describes elements of argumentation for dealing with uncertain knowledge. Besides, he was also interested in analogy and analogical proportions. His student, and follower as head of the Lyceum, Theophrastus (c. 371–c. 287 BC) was bearer of multiple talents: he was one of the first botanists, his book *Characters* portrays thirty moral types, and he was a logician. He developed elements of modal logic (Bocheński 1947), worked on hypothetical syllogisms (Barnes 1983) and seems to have been the first one to propose that the trust in the conclusion of a logical chain corresponds to the

trust in the weakest link (Rescher 1976). Let us also cite Galen of Pergamon (129–c. 200), the famous Greek physician, surgeon and philosopher who used hypothetical syllogistic after Aristotle and his school (Bobzien 2004).

In the same time period, Chinese philosophers belonging to Confucius (551–479 BC)'s school, such as Meng Tzeu (Mencius) (c. 385–c. 301 BC), who was a contemporary of Aristotle, used argumentation and analogy, but were not considering logical issues nor discussing reasoning patterns such as syllogisms, with perhaps the exception of Gongsun Long (c. 325–250 BC), interested in paradoxes, who was belonging to another school founded by Mozi (c. 468–c. 391 BC). See Guo (2017) on the discrepancy between Greek and Chinese philosophers with respect to reasoning issues. The Roman–Greek world had also supporters of analogical reasoning, like the Epicurean philosopher Philodemus of Gadara (c. 110–prob. c. 40 BC) also interested in inductive reasoning (De Lacy and De Lacy 1941).

Logicians and Theologians in Medieval Time

Augustine of Hippo (354–430), in his treatise *Contra Academicos* (386) against the skepticism of the neoplatonic Academy, asserts the existence of knowledge; hence he was certain of logical truths such as "either there is only one world or there is not only one", mathematical facts such as "it is necessarily true that three times three equals nine", as well as reported perceptions "I do not know how an academician may rebut a man who says: "I know that this seems white to me; I know that this tone sounds pleasant to my ears, this has a pleasant smell for me, this tastes wonderful, this is still cold for me" " (Augustine of Hippo 1995; Smalbrugge 1986; Curley 1996). The Aristotelian legacy of syllogisms and of the square of opposition was reworked during the Middle Ages from Boethius (c. 477–524) to William of Sherwood (c. 1200–c. 1272) (1966). This English logician who taught at the University of Paris and was the author of one of the four main textbooks on logics in his century; the three others were authored by Lambert of Auxerre (or Lambert of Lagny) (XIIIth century) (2015), Peter of Spain (XIIIth century) (2014), and Roger Bacon (c. 1219/20–c. 1292) (2009). Jean Buridan (1292–1363), professor at Sorbonne in Paris, seems to be one of the very first who attempted to separate logic from theology (see (Read 2012) for a presentation and a discussion of his work on syllogisms, and (Hughes 1982) on self-reference). Indeed, the interest for logic at that time was closely related to various theological issues, among which the proofs of the existence of God, such as the one of Anselm of Cantorbery (1033–1109) (2001) (this ontological proof was shortly challenged by Gaunilo of Marmoutiers (c. 990–c.1083) who dared substitute the "Lost Island" for God in the proof and introduced a distinction between "(rational) thinking" ("cogitare") and "knowing by intelligence" ("intelligere")).[4] The main scholastic philosophers, Pierre Abélard (1079–1142), Thomas Aquinas (1224–1274),

[4] Arguments for and against the existence of God, and their discussions have a very long history from Anselm to Kurt Gödel (1995) (his proof is stated in second-order modal logic); see the monograph by Sobel (2004) for a complete exposition of these arguments. Thomas Aquinas (1975, 2006) himself disputes Anselm's ontological argument (see (Sousa Silvestre 2015) for a modern logical discussion) and proposed five other proofs (Aquinas 1975, 2006); see also Mavrodes (1963), Wade (1967) on logical accounts of God omnipotence paradoxes.

Jean Duns Scot (1266–1308) are logicians and theologians. As Thomas Aquinas, their interests also included issues connected with argumentation or even analogy.

Aristotle was the main non-Christian philosopher to influence the above medieval philosophers, in particular thanks to the comments of Muslim philosophers such as the Persian polymath Avicenna (Ibn Sīnā) (980–1037) and the Andalusian thinker Averroes (Ibn Rushd) (1126–1198). Originally Arabic logic was developed by a school centered at Baghdad, where Abū Bishr Mattã ibn Yūnus (c. 870–940) was the first logician to write in Arabic, to translate Aristotle, and to write commentaries. Works on modalities, in particular temporal ones, have been developed early by Arabic logicians (Rescher 1967) continuing the works initiated by Al-Fârâbî (872–950), himself influenced by Aristotle, and "perhaps the most important logician of Islam" (Rescher 1963). Let us also mention Yahyā ibn 'Adī (893–974), a Nestorian Christian like Abū Bishr, who studied logic and philosophy with both previously mentioned scholars, and translated a number of works of Greek philosophy into Arabic; his teaching was especially influential ("virtually half of the Arabic logicians of the Xth century are his pupils" Rescher 1963). It has been recently advocated that the idea of diagrammatic reasoning can be already found in Abū al-Barakāt (c.1080–1164/1165)'s writings, as well as the definition of model-theoretical consequence, which as its roots before in works by Ibn Sīnā and even before by Paul the Persian (VIth century),[5] see Hodges (2018) for details.

In old Indian logic (Gillon 2010; Sarukkai 2018), the Nyāya school is known for its development of the first elements of logic. Later, Dignāga (c. 480–c. 540), a Buddhist scholar, renewed inferential reasoning and provided a list of valid arguments, while Bhāviveka, (c. 500–c. 578), another Buddhist scholar, is usually credited as being one the first to use formal syllogisms.

The argumentation issue is one of the main topics in the protean work of the Catalan mystic, writer, theologian and philosopher, Ramon Llull (1232–1316), who designed a "logical machine" for arguing, his *Ars Magna*, in order to establish the truth of statements from rules of combinations of symbols (Fidora and Sierra 2011; Crossley 2005).

Finally, let us also mention Guillaume d'Ockham (c. 1285–1347), logician and philosopher, whose principle of parsimony (known as "Ockham's razor") expresses that the simplest sufficient hypotheses for explaining a situation must be the most likely ones. One may also cite the fourteenth century author Roger Swyneshed, credited with logical works on insolubles (Spade 1979; Spade and Read 2018) and on obligations (*De insolubilibus* and *De obligationibus*), as well as some of the members of the "Calculators's group" at Merton College in Oxford, who were primarily interested in kinematics and mechanics, such as William Heytesbury (before 1313–1372/3), also known for his *Rules for Solving Sophismata* (1335), or Thomas

[5]Paul the Persian is an East Syrian theologian and philosopher who worked at the court of the Sassanid king Khosrow I (501–579), and wrote several treatises and commentaries on Aristotle (Teixidor 2003), which had some influence on medieval Islamic philosophy.

Bradwardine (c. 1290–1300–1349) who also contributed to the medieval *insolubilia* (insolubles) literature (Read 2010) (dealing with the liar paradox, among other topics), or Richard Swineshead (mid 1300s) known for *The Book of Calculations* where logic and mathematics began to move physics outside natural philosophy. Let us also mention Paul of Venice (c. 1369–1429) who dealt with the problem of the meaning and truth of sentences in his *Logica Magna* (Conti 2017), while Charles de Bovelles (c. 1475–after 1566), a French mathematician and philosopher, was the author of an *Ars oppositorum* (1510), a treatise of logic where the idea of opposition plays a central role.

Hence, for nearly 1900 years, logics and the art of reasoning mainly remained on the path set by Aristotle, mainly motivated in the Middle Ages by theological concerns. We have limited this brief overview of medieval logic to the mention of the main authors and to the concerns on which their speculations were based, without attempting to detail their contributions in the context of AI today. For introductory overviews, the reader is referred to Uckelman (2017), Hubien (1977), and for more detailed surveys, the reader should consult (Dutilh Novaes and Read 2016; Gabbay and Woods 2008b; Dubucs and Sandu 2005; Busquets 2006).

Dreams and Machines

Besides, artificial creatures populate our collective psyche since Antiquity, as it is reflected in many myths, fairy tales and literary works in most of the cultures. We limit ourselves to some great figures, for further details the reader may consult (Cohen 1968; Chassay 2010). Homer (1984) describes, in the book XVIII of the *Iliad* the god Héphaïstos' creations, in particular, twenty autonomous tripods equipped with gold wheels for carrying the products of his forgery, or the two golden servants able to assist him. Apollonios of Rhodes (IIIrd century BC) in the *Argonautica* (II, 4) (1959) relates the creation of a giant bronze statue, named *Talos*, by Héphaïstos, which he offers to Minos in order to defend Crete against invaders. Ovid (43 BC–17/18 AD) reports the myth of *Pygmalion* in the book X of *Metamorphoses* (Ovid 1998). Pygmalion carved an ivory statue of a woman *Galatea*, to which the goddess *Venus* gives life. These few examples illustrate recurring themes that are later found in the literature of the XIXth and XXth centuries, long before the first robots: the desire for escaping from the labour servitude, the satisfaction of love or erotic fantasies, the use of artificial creatures for warlike use. Hence, these legends and myths, rooted into the reality of their time, imagine the production of prophetic and articulated statues or masks. For example, the *Bible* refers to *teraphim* oracle figurines that Nabuchodonosor consults for querying the fate *Ezechiel* (XXI, 26). All these creatures refer to the fabulous; however other kinds of mechanisms and machines are designed for utilitarian or ludic ends, such as the invention of the pulley and the screw ascribed to Archytas (IVth century BC), or the clepsydras from Ctesibios (IIIrd century BC) and Philo of Byzantium (end of IIIrd century BC), considered as the first fully automatic devices, or even the pneumatic machines and the automata from Hero of Alexandria (1st century). Let us also mention Petronius (14–66) who in the Satyricon (book XXXIV) (1969) describes a silver articulated skeleton able to take up various positions. Later,

Ismail al-Jazari (1136–1206), a mechanical inventor, built musical automata powered by water.

Regarding the idea of mechanizing reasoning, we can hardly mention the logical machine proposed by Ramon Llull (1232–1315), described in *Ars Magna* (1305) (2011). This machine, consisting of paper disks swivelling on an axis, is a tool intended for reasoning assistance in order to answer theological queries, in particular with the goal ... of converting Muslims to christianity on a rational basis. In the field of arithmetics, it is worth mentioning the Chinese abacus which appears in its final form in the XIIth century.

Literature

The Medieval chivalric romance, permeated with the fabulous, often refers to animate statues. For example, in the *Roman de Tristan* from Thomas of Britain (XIIth century) (1969), Tristan, thanks to the giant Moldagog, erects wonderful statues among which is the statue of Yseult. During this period some narratives witness of the construction of automata, however they are considered as devilish and sacrilege by the ecclesial authorities. An important myth that transcends the centuries is the one of Golem, an artificial creature created from clay. It yet appears in the *Talmud* and is pointed out in the *Bible* (Psalm 139 : 16). In the Middle Ages, an ashkenazi esoteric text, the *Sefer Yetsirah* gives a detailed description of the creation of a Golem, however it is only from the XVIth century onwards that a Golem becomes a servant who discharges his creator of heavy works until he gets far beyond his control. Several versions of the myth circulated in central Europa. According to the Polish version reported by the German storyteller Jacob Grimm in the *Zeitung für Einsiedler* ("Newspaper for Hermits") in 1808, the Rabbi Chelm would have given life to a clay Golem writing the word truth in hebrew on his forehead. According to the version from Prague, the Rabbi Loew would have given life to a Golem by putting on his mouth a paper on which is written the name of God. The latter will be popularized in the XXth century by the eponymous novel.

3 The XVIIth Century: Preliminary Steps Towards Modernity

The time period that starts with the beginning of the XVIIth century exhibits a slow transition towards the birth of modern logic 250 years later with the funding works of George Boole, as well as it shows the first developments of probability. This is also the time of the emergence of the first machines.[6] As already said, we try to indicate names (and facts) having a significant relation with some concerns of AI, some being better known for other issues not related to AI, others being just forgotten.

[6] A version of this section and of the next two sections has already appeared in Marquis et al. (2014).

Treatises of Logic

In philosophy of knowledge, Francis Bacon (1561–1626) promotes the inductive method based on observation for scientific discovery (Bacon 1605), at the beginning of the XVIIth century. Besides, in 1603 the first treatise of logic in French is published (Dupleix 1603). It is written by Scipion Dupleix (1569–1661), a preceptor of a son of the king Henri IV. His course of philosophy also includes a *Physique*, a *Métaphysique*, and an *Ethique*. His *Logique* is a vast compilation of previous knowledge, and deals, among other issues, with the square of oppositions, modalities, syllogisms, incomplete syllogisms (patterns of default reasoning called enthymems), sorites, and argumentation, all topics inherited from Aristotle and from his followers during the Antiquity and the Middle-Age. In the middle of the XVIIth century, *Le Philosophe François* (De Ceriziers 1650) written by René de Ceriziers (1603–1662), includes a large section devoted to logic where argumentation is developed in great detail. This is also the time where the modern history of legal reasoning (Kalinowski 1982) starts. Let us also mention the *Essai de Logique* (Mariotte 1678) by the physicist Edme Mariotte (c.1620–1684), which discusses issues about proofs in geometry, reasoning about the physical world, and deontic reasoning.[7]

Prescience

Thomas Hobbes (1588–1679) seems to be the first to explicitly link the symbolic manipulation of terms in logic to the idea of mathematical calculation. Indeed, he wrote "*Per ratiocinationem autem intelligo computationem.*" (or in English one year later " *By ratiocination I mean computation*.")[8] in his *De Corpore* (Hobbes

[7]Some other authors would be also worth mentioning, such as the Flemish philosopher Arnold Geulincx (1624–1669), author of treatises of logic entitled *Logica fundamentis suis restituta* (1662) and *Methodus inveniendi argumenta* (1663).

[8]The text continues with "*Now to compute, is either to collect the sum of many things that are added together, or to know what remains when one thing is taken out of another. Ratiocination, therefore, is the same with addition and subtraction*;" (or in Latin: "*Computare vero est plurium rerum simul additarum summam colligere, vel una re ab alia detracta cognoscere residuum. Ratiocinari igitur idem est quod addere and subtrahere*"). One page after one reads: "*We must not therefore think that computation, that is, ratiocination, has place only in numbers, as if man were distinguished from other living creatures (which is said to have been the opinion of Pythagoras) by nothing but the faculty of numbering; for magnitude, body, motion, time, degrees of quality, action, conception, proportion, speech and names (in which all the kinds of philosophy consist) are capable of addition and subtraction.*" (or in Latin: "*Non ergo putandum est computationi, id est, ratiocinationi in numeris tantum locum esse, tanquam homo a caeteris animantibus (quod censuisse narratur Pythagoras) sola numerandi facultate distinctus esset, nam and magnitudo magnitudini, corpus corpori, motus motui, tempus tempori, gradus gradui, actio actioni, conceptus conceptui, proportio proportioni, oratio orationi, nomen, nomini (in quibus omne Philosophiae genus continetur) adjici adimique potest.*"). In fact, the anecdote reported does not concern Pythagore, but Platon, see Hobbes of Malmesbury (1655) note p. 13. Moreover, as early as 1651 (Hobbes of Malmesbury 1651) in chapter V (*Of Reason and Science*) of *Of Man*, the first part of his *Leviathan*, Hobbes had given a preliminary version whose beginning was "When a man 'reasoneth' he does nothing else but conceive a sum total, from 'addition' of parcels, or conceive a remainder, from 'subtraction' of one sum from another; which, if it be done by words, is conceiving of the consequence of the names of all the parts, to the name of the whole; or from the names of the whole and one part, to the name of the other part."

of Malmesbury 1655), whose reputation was unfortunately somewhat damaged by the inclusion of a tentative proof of the squaring of the circle, even if Hobbes will acknowledge its falsity later.

It is also interesting to mention here a passage of the fifth part of the "*Discours de la Méthode*" (Descartes 1637), where René Descartes (1596–1650) who advocates a conception of animals as beings with a complete lack of reason, similar in that respect to machines, shows a remarkable prescience with respect to the discussion about how to distinguish humans from machines and the Turing test.[9]

Reasoning and Probability

The *Logic of Port-Royal* (1662) by Antoine Arnauld (1612–1694) and Pierre Nicole (1625–1695), initiates a theory of sign and representation for about two centuries, and is a landmark in the history of philosophy of language and in logic, which however still remains here somewhat connected to issues in theology. The book is organized in four main parts corresponding respectively to the faculties of conceiving, of judging, of reasoning (deductively through syllogisms), the last part discussing methodological questions. The mathematics are here the reference that should be transposed to the study of language statements, and reasoning. Lastly, the idea of probability is here, apparently for the first time, associated not with the combinatorics of games of chance, but with the evaluation of the confidence that can be attached to testimonies. Jacques Bernoulli (1654–1705), in his *Ars conjectandi* published only in 1713, proposes distinct calculi for these two types of uncertain situations (Shafer 1978). Close to Port-Royal people, let us recall that Blaise Pascal (1623–1662) is, among many contributions, both a pioneer of probabilities (in communication through letters on this topic with Pierre de Fermat (c. 1605–1665)), and the inventor (in 1642) of a mechanical computation machine called *Machine arithmétique* able to perform additions and subtractions.

This is the Dutch mathematician and physicist Christian Huyghens (1629–1695) who publishes the first treatise on the probability calculus (Bessot et al. 2006) and

[9] "I worked especially hard to show that if any such machines had the organs and outward shape of a monkey or of some other animal that doesn't have reason, we couldn't tell that they didn't possess entirely the same nature as these animals; whereas if any such machines bore a resemblance to our bodies and imitated as many of our actions as was practically possible, we would still have two very sure signs that they were nevertheless not real men. The first is that they could never use words or other constructed signs, as we do to declare our thoughts to others. We can easily conceive of a machine so constructed that it utters words, and even utters words that correspond to bodily actions that will cause a change in its organs (touch it in one spot and it asks "What do you mean?", touch it in another and it cries out "That hurts!", and so on); but not that such a machine should produce different sequences of words so as to give an appropriately meaningful answer to whatever is said in its presence - which is something that the dullest of men can do. Secondly, even though such machines might do some things as well as we do them, or perhaps even better, they would be bound to fail in others; and that would show us that they weren't acting through understanding but only from the disposition of their organs. For whereas reason is a universal instrument that can be used in all kinds of situations, these organs need some particular disposition for each particular action; hence it is practically impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way our reason makes us act." (Transl. J. Bennett).

introduces the notion of expectation in an uncertain situation, among multiple scientific contributions including the improvement of clocks. Then this will be followed by the works of Abraham de Moivre (1667–1754) who will propose the first definition of statistical independence (De Moivre 1718). However, it seems that people had started to be interested in questions about uncertainty and risk (Meusnier and Piron 2007) before the beginning of the XVIIth century. Indeed, it has been discovered that a Latin text, the *De Vetula* (in English *On the Old Woman*), a pseudo-Ovidian poem attributed to the philosopher and poet Richard de Fournival (1201–1260), which was widely circulated in its time, was containing probability calculations on the throw of three dice (Bellhouse 2000); see more generally (Bru and Bru 2018) and (Shafer 2018) for the history of the counting of chances in dice games and the estimation of fair price when dividing the stakes in a prematurely halted game, which started much before Fermat and Pascal.

4 The XVIIIth Century: The Age of Enlightenment

Progresses in logic and probability remain slow in the XVIIIth century, although philosophers develop concerns about human understanding.

Philosophers

At the transition between the two centuries, Gottfried Wilhelm Leibniz (1646–1716) has not only been the philosopher that everybody heard about, and one of the fathers of the infinitesimal calculus (without mentioning many other works in mathematics, in physics, and in history). Indeed he also has an important role in the evolution of logic (see for example Gochet et al. 1988, 1989), which has been rediscovered lately (Couturat 1901, 1903), due in particular to his search for a universal language (the *lingua characteristica universalis*) that enables the formalization of the thought and an algorithmic logical calculus (*calculus ratiocinator*), thus anticipating the project of Frege. He is also at the origin of the idea of "possible worlds", and was interested in issues in legal and deontic reasoning. Another slightly later attempt at developing a logical formalism is the one by Gottfried Ploucquet (1716–1790) (Ploucquet 2006). Leibniz (1703) is also the first to imagine the binary numeration. Moreover he proposed a machine able to perform the four arithmetic operations in 1673 (finally recognized as imperfect). Let us also add that Leibniz was a good chess player who was also interested in the scientific understanding of the game. The reader is referred to Lenzen (2016) for a recent study of Leibniz's algebra of concepts, which anticipates some aspects of modern formal concept analysis. Another German philosopher, Christian Wolff (1679–1754), should be mentioned for his treatise (1713) of logic, translated in French and English. More generally, the reader is referred to Gabbay and Woods (2004b) on the rise of modern logic after Leibniz. A particular mention should be made for Johann Christian Lange (1669–1756) who invented a tree-like diagram for solving syllogisms (1714); see Lemanski (2018).

Let us also particularly mention another philosopher as a forerunner of different AI concerns: David Hume (1711–1776), for whom the origin of our knowledge

comes from experience (Hume 1748), and ideas are not innate (as already for John Locke (1632–1704) (Locke 1690) or George Berkeley (1685–1753)). He establishes a distinction between first "impressions", and "ideas" which are weakened images, synthesized from impressions; for him, ideas are associated by different relations such as resemblance, (temporal or spatial) contiguity, or causality (a relation that he has especially analyzed). He also makes a distinction between logical truths and empirical truths which cannot be certain, but only probable, and points out that induction cannot lead to any certainty. He has also discussed analogical arguments.

Logic and Probability

The name of the Swiss mathematician Gabriel Cramer (1704–1752) is especially attached to the resolution of linear equation systems. But if his presence is relevant in this overview of the prehistory of AI, it is because of his course of logic (Cramer 1745; Martin 2006a), remained unpublished until now, that he wrote in 1745 as a preceptor in a rich family. In his introduction, he makes a distinction between the *logique naturelle* (the one used spontaneously in reasoning) and the *logique artificielle* (the one that is founded on principles and rules). The presentation of this latter topic is developed along two main parts of approximately equal importance, one dedicated to the search for truth and the understanding of "how human mind forms ideas, compare them in order to state judgements and to chain them for deductive purposes",[10] and the other devoted to the study of probabilities as measures of the likelihood of the propositions or judgements about events. Thus, in a certain way, this *Cours* could be compared in its intention to the *Laws of Thought* by George Boole who, a little more than a century later, devotes parts of equal length to what will be called later Boolean logic, and to probabilistic reasoning under uncertainty. It seems by the way that Cramer's *Cours* would be the anonymous source of the article *Probabilité* (attributed to Benjamin de Langes de Lubières (1714–1790), see (Candaux 1993)) in the *Encyclopédie* by Denis Diderot (1713–1784) and Jean le Rond D'Alembert (1717–1783) (edited from 1751 to 1772). Moreover let us indicate that the article (also anonymous) *Logique* in the *Encyclopédie* also contrasts natural logic and artificial logic and refers for this latter to the article *Syllogisme* also anonymous.

Let us also mention the Alsatian mathematician Johann-Heinrich Lambert (1728–1777) who in his *Neues Organon* (Lambert 1764) develops a probabilistic theory of syllogisms, with application to the handling of the probability of testimonies (Shafer 1978; Martin 2006b, 2011). Lambert, as the article *Probabilité* in the *Encyclopédie*, proposes a reinforcement rule of the confidence in corroborating testimonies, which may retrospectively appear as a particular case of the combination rule in Dempster–Shafer belief function theory. Besides, a landmark work in probability is the posthumous article by Reverend Thomas Bayes (1702–1761), communicated by his friend Richard Price (1723–1791), about the famous theorem of the computation of the a posteriori probability from priors and conditional probability (Bayes 1763), result

[10] In French: "comment l'esprit humain se forme des idées, les compare pour en porter des jugements et enchaîner ces jugements pour déduire les uns des autres".

found again by Pierre-Simon Laplace (1749–1827) in his works on probability and induction (Laplace 1814).

Some other names are also worth mentioning, on different issues. The grammarian César Chesneau Du Marsais (1676–1756), also a contributor to the *Encyclopédie*, studies the patterns in rhetorics in his *Traité des Tropes* (Dumarsais 1730), and has concerns that might still have some relevance in automated treatment of languages and in argumentation in natural language. Besides, the philosopher and mathematician Nicolas de Condorcet (1743–1794), probabilist, pioneer in statistics, studied the representativity of voting systems (as also his contemporary Jean-Charles de Borda (1733–1799), mathematician, physicist, and sailor De Borda 1781), and stated the famous paradox on the possible intransitivity of majoritarian relative preferences (Condorcet 1785).

Literature

The Age of Enlightenment, a century of progress towards reason and rationality, is also marked by literary works that contribute to feed our collective psyche. Jonathan Swift (1667–1745), in his novel *Gulliver's travels* (Swift 1726), develops an ironical criticism of the society of his time and intends to show the inadequation of humans with reason. More particularly, during the fourth travel, Gulliver stays with the *Houyhnhnms*, which are "*reasonable*" animals ignoring contradiction and argumentation (chap. VIII) and whose language does not include any word for expressing lies, since saying something false would be betraying the functions of language (chap. III and IV). During a previous trip to *Laputa*, Gulliver visits the *Academy of Lagado* (chap. V) where he sees a machine that generates sentences for helping to write books.

Automata

The XVIIIth century is also marked by the automata built by Jacques Vaucanson (1709–1782), such as his Tambourine Player (Vaucanson 1738), or his Digesting Duck (1744). These automata are in some way echoing the mechanical view of human (La Mettrie (Offray de) 1747) supported by the philosopher Julien Offray de La Mettrie (1709–1751). These automata impressed the minds of the contemporaries. For instance, Mrs de Genlis, born Stéphanie-Félicité Du Crest (1746–1830), in one her educative and moral tale (Du Crest, comtesse de Genlis 1797) stages two child automata, one making drawings and the other playing music. The idea of an animated toy may fuel all the fantasies, like in the novel *Pigmalion* (Boureau-Deslandes 1742) by André-François Boureau-Deslandes (1690–1757), or in the novel (Galli de Bibiena 1747) by Jean Galli de Bibiena (1709–c.1779), where the narrator is fascinated by a doll found in a store and later discovers that it is a sylph! Shortly after, in 1769, the Hungarian Johann Wolfgang van Kempelem (1734–1804), born in Slovakia, built an automaton, attracting considerable attention, a Mechanical Turk or Automaton Chess Player, able to answer questions. This "Turk" has opponents as famous as Catherine the Great, Napoleon Bonaparte, or Benjamin Franklin. Resold at the death of van Kempelem, it had a long career, and it took time before discovering how a man could be hidden in the "machine", but van Kempelem was in spite of

that the author of a genuine vocal synthetizer (in 1791)! The chess player from van Kempelem has fascinated and inspired several novels in the next centuries.

Another famous opponent to this false automaton (against whom he lost two times) is Charles Babbage (1791–1891), who later in 1837, designed the first programmable computer (with punched cards) having a memory, the *Analytical Engine*, and for which Ada Lovelace (1815–1852) (the daughter of the poet George Byron) wrote the first programmed algorithm.

5 The XIXth Century: The Rise of Modern Logic

After some isolated attempts at formalizing syllogisms at the very beginning of the XIXth century, modern logic finally appears in the middle of the century.

Pioneers of Formal Logic

The beginning of XIX^{e} century is marked by the publication of some isolated works which may retrospectively appear as important milestones between the theory of syllogisms inherited from Aristotle and modern logic. Thus, Frédéric de Castillon (1747–1814) proposed a formal calculus for solving syllogisms (De Castillon 1804, 1805). Besides, starting from the idea of set diagrams, beautifully introduced by Leonhard Euler (1707–1783) for visualizing syllogistic reasoning (in seven of his famous *Lettres à une Princesse d'Allemagne sur Divers Sujets de Physique & de Philosophie* Euler 1761, publ. 1768[11]), Joseph D. Gergonne (1771–1859), a French mathematician, mainly known as a geometer, published an article (Gergonne 1816b; Giard 1972) in 1816 where he identified the five possible relations between two sets,[12] and characterized the valid syllogisms for the first time. A modern counterpart of this work can be found in Faris (1955). Besides, Gergonne has also proposed polynomial regression, and was interested in the rule of three (Gergonne 1815, 1816a). Quite ironically, although he was a geometer, Gergonne emphasized, as early as 1813, the interest of algebraic methods in mathematics (algebra was at that time mainly restricted to operations on the reals) (Dahan-Dalmedico 1986), but this is George Boole (1815–1864) who will be the first to apply this idea to logic. In other respects, Bernard Bolzano (1781–1848), a German-speaking mathematician, logician, philosopher and theologian is also worth-mentioning here as a precursor of

[11] Although researchers nowadays speak of 'Euler diagrams', similar diagrams have already been used by many authors before (Lemanski 2017). Among others, the diagrams of Juan Luis Vives (1493–1540) (who used a 'V'-like nested representation for the three items in the syllogism in *Barbara* "Any B is a C, but any A is a B, therefore any A is a C ", in a treatise entitled *De Censura Veri*, part of his encyclopedic compendium *De Disciplinis Libri*), as well as those of Nicolaus Reimarus Ursus (Nicolaus Reimers) (1551–1600) in his *Metamorphosis Logicae* (Strasbourg,1589), of Erhard Weigel (1625–1699) in his *Philosophia Mathematica, Theologia Naturalis Solida* (1693), of Johann Christoph Sturm (1635–1703) in the *Universalia Euclidea* (1661), or still those of Leibniz, and Johann Christian Lange (1669–1756) can be mentioned as precursors of the logic diagrams used by Euler.

[12] This result was already anticipated in the line diagrams (using pairs of segments) by Abū al-Barakāt; see Hodges (2018).

predicate logic, for his original view of logic in terms of variations (where different types of propositions are defined depending on the ways changes in their truth value can occur), and for the analysis of five meanings that the words 'true' and 'truth' may have in different uses (1837).

Boole, De Morgan and Their Time

The middle of the XIXth century is marked by the publication of the founding works of Boole and Augustus De Morgan (1806–1871) on the mathematisation of reasoning (Boole 1847; De Morgan 1847). Boole develops a symbolic view of logic, and an equational theory of deduction, based on the binary algebra named from him. It is quite noticeable that both Boole and De Morgan were interested both in logic and probability in their works, which enables them to have a renewed approach of syllogisms (Boole 1854; De Morgan 1868). Indeed logic and probability have a pretty much equal place in the celebrated book by Boole (1854) *An Investigation of The Laws of Thought on which are founded the mathematical theories of logic and probabilities*. It should be also emphasized that studies on logic and the laws of thought had become a topic relatively popular at that time with the books of the archibishops Richard Whately (1787–1863) and William Thomson (1819–1890), and of the philosopher John Stuart Mill (1806–1873)[13] (Whately 1826; Thomson 1842; Stuart Mill 1843), published before the first works of Boole and De Morgan on this topic. The final version, substantially expanded (which even includes an appendix on the logic in India) of the *Outline of The Laws of Thought* (Thomson 1857) by Thomson pays an homage to De Morgan in turn. Let us also note that Stuart Mill presents new views on induction in his book among other things, and proposes five qualitative inference rules for causal reasoning. In a more amusing style, Lewis Carroll (1832–1898), the author of *Alice's Adventures in Wonderland*, under his nom de plume, actually wrote a treatise of symbolic logic (Carroll 1896; Braithwaite 1932) (where he is using an original diagrammatic representation), with many exercises and problems presented in a funny way. The subtitle of his book was indeed "*A fascinating mental recreation for the young*"! See Moretti (2014) for a discussion of 'logical diagrams' and 'logical charts' that can be found in this treatise. Besides, under his patronymic name Charles L. Dodgson, Lewis Carroll had also refined a voting method due to Condorcet (Dodgson 2001; Ratliff 2010) some twenty years before. More material on the richness of British logic in the XIXth century can be found in Gabbay and Woods (2008a). Lastly, more perhaps as a curiosity than as influential contributions, two books by the Irish engineer and mathematician Oliver Byrne (1810–1880) are worth mentioning where he respectively dealt with analogical proportions in a pre-symbolic manner (1841) and with colored diagrams and symbols for helping the understanding of proofs (1847). Besides, the German philosopher Christoph von Sigwart (1830–1904) examined English induction theories in his *Logik* (1873–1878), while Joseph Delbœuf (1831–1896), a Belgian psychologist and philosopher outlines a "*Logique*

[13] Stuart Mill is perhaps better known as an economist, and a strong advocate of utilitarism (Stuart Mill 1863), following Jérémy Bentham (1748–1832), i.e. a consequentialist approach to decision making.

Algorithmique" based on algebra (1876), somewhat departing from Boole's work, of which he had a limited knowledge coming from a teaching monograph by the Scottish philosopher Alexander Bain (1818–1903) (1870) containing some account of the novel schemes of De Morgan and Boole. Bain was a follower of Stuart Mill, and William Hamilton (1788–1856), another Scottish philosopher (not to be confused with the Irish mathematician who invented quaternions), who worked on syllogisms (1859–1860) (Pratt-Hartmann 2011).

Further Developments in Logic

As a following of Boole's and De Morgan's works, the algebra of logic was developed by Ernst Schröder in Germany (Schröder 1890), Charles Sanders Peirce (1839–1914) (Peirce 1870, 1880, 1885, 1931, 1955) and his followers Oscar Howard Mitchell (1851–1889) (Mitchell 1883) and Christine Ladd-Franklin (1847–1930) (Ladd 1883) in the United States, and in France (Couturat 1905) by Louis Couturat (1868–1914), who was also a great specialist of the logic of Leibniz. The Euler set diagrams were improved by John Venn (1834–1923) who shaded the empty parts of his diagrams rather than representing the sets in the exact configuration where they are supposed to be (Venn 1880, 1881), and by Peirce for taking into account existential statements and disjunctive information (Shin and Lemon 2008). Besides, Venn in the multiple editions of his book *The Logic of Chance* also developed probabilistic aspects of reasoning, privileging the frequentist interpretation (Venn 1866).

William Stanley Jevons (1835–1882), who wrote one of the most popular introductory text to Boolean logic in his time (Jeavons 1870), also built a logic machine in 1869, called "Logic Piano", based on a substitution principle (Jeavons 1869), which was able to draw conclusions mechanically from premisses. In a quite different perspective, automata are built during the XIXth century. Let us mention the speaking head made by Joseph Faber (1800–1850) named "Euphonia", able to articulate words, the speaking doll of Thomas Edison (1847–1931) commercialized in 1889, the "Steam-Men" of the American Zadock Dederick in 1868 and of the Canadian George Moore, which walked in 1893 at the speed of 8 km/h, and closer to us, the automaton by Leonardo Torres y Quevedo (1852–1936) which in 1914 was able to automatically play a king and rook endgame against king from any position; see Vigneron (1914) for a detailed description.

Literature

In relation with the romantic movement, and in reaction to the rationality of the previous century, the beginning of the XIXth century sees the emergence of fantasy literature with the development of gothic fiction. One of the classics is Mary Shelley's (1797–1851) famous novel *Frankenstein or the Modern Prometheus* (1818). Victor Frankenstein, throughout his investigations, succeeds in discovering the secret of life and creates a superhuman artificial man whose terrifying appearance scares him away. Left on her own, this creature learns to speak by observing humans and tries to make contact with them, but she is rejected by the fright she inspires. Suffering from isolation, she has a fierce hatred of her creator on whom she wants to take revenge. Even today this novel has achieved widespread popularity and inspired many movie

adaptations since the beginning of moving pictures. Other authors carry on known themes of previous centuries. Hence, Prosper Mérimée (1803–1870) reuses the old theme of the animated statue in *La Vénus d'Ille* (1837), Edgar Allan Poe (1809–1849) takes his inspiration from automata with *Maelzel's chess player* (1836), or Gustav Meyrink (1868–1932) revisits the myth of *Golem* with the eponymous novel (1915). The scientific and technological developments in the context of early industrialization inspire literary creation. The artificial creatures are no longer purely mechanical, thus electricity and electromagnetism play a major role in *L'Eve future* (1886) by Auguste de Villiers de L'Isle-Adam (1838–1889), the human-machines (*engine-men*) or "steam-human" in the novel *Ignis* (1883) by the less well-known writer, Didier de Chousy (1834–1895), or even the "rammer" (in French,"hie" or "dame"), professor Cantarel's mechanical "young lady" in *Locus Solus* (1914) by Raymond Roussel (1877–1933) (in the same spirit, see also Clair and Szeemann (1975)). In addition, let us mention the president-automaton who operates with three keys held by the president of the Chamber of deputies, the president of the Senate and the president of the Council respectively (1883) by Albert Robida (1848–1926).

6 The First Half of the XXth Century: From Mathematical Logic to Cybernetics

The beginning of the XXth century, regarding logic, is principally marked by the development of predicate logic, after the seminal works by Gottlob Frege (1848–1925), with the introduction of quantifiers (also (re)discovered independently by O. H. Mitchell, already cited, see (Dipert 1994)). A logical system is then thought both as a representation language, and a formal system for deduction (Geach and Black 1980; Gochet et al. 1988, 1989). This has led to a series of very important developments which have primarily concerned the foundations of mathematics like the *Principia Mathematica* (1910) by Alfred North Whitehead (1861–1947) and Bertrand Russell (1872–1970), or in 1931 the Kurt Gödel's (1906–1978) incompleteness theorems (Nagel and Newman 1958). We owe the notations (completed by Whitehead and Russell) of modern logic to the mathematician Giuseppe Peano (1858–1932). It is not the place here for presenting an history nor to even sketch a panorama of modern mathematical logic. We just cite some names, closely related to:

- the foundations of computability theory like Alonzo Church (1903–1995) and Alan Turing (1912–1954) whose research works first concern the foundations of computer science;
- the development of intuitionistic logic, which rejects the law of excluded middle, thanks to the works (1956) of Arend Heyting (1898–1980), following the constructivist approach to mathematics advocated by Luitzen Egbertus Jan Brouwer (1881–1966);
- the deduction theory in classical logic like those of Leopold Löwenheim (1878–1957), Thoralf Skolem (1887–1963), Jacques Herbrand (1908–1931), and Gerhard

Gentzen (1909–1945) for natural deduction and sequent calculus (about them consult (Largeault 1972; Herbrand 1968; Gentzen 1969)). These are the starting points of the seminal works on logical deduction of Martin Davis (born in 1928) and Hilary Putnam (1926–2016) (1960) and John Alan Robinson (1928–2016) (1965);

- the concept of truth, semantics and model theory with Alfred Tarski (1902–1983) (1956).

Philosophy and Cognitive Aspects

The above listed works aroused many philosophical echoes or counterpoints. Concerning the first half of the XXth century, they can be found particularly with Bertrand Russell (1956) (see also (Vuillemin 1971)), with Ludwig Wittgenstein (1889–1951) (1921; 1969), or even with Willard Van Orman Quine (1908–2000) (Quine 1941), or with Rudolf Carnap (1891–1970) (Carnap 1942), the latter was interested both in logic and probability,[14] like Hans Reichenbach (1891–1953). Let us also mention Carl Gustav Hempel (1905–1997), who, like the last two authors, was a significant representative of logical empiricism and who proposed a model for scientific explanation; he left his name associated with a paradox about confirmation, "the raven paradox" (Hempel 1965) (seeing a black raven can be taken as a confirmation that all ravens are black, while seeing a white swan does not confirm it, although equivalently all non black things are non ravens). Karl Popper (1902–1994) emphasized the idea that a scientific theory can be falsified by one counterexample, while it can never be proven, but only indirectly supported by the observation of consequences. Other worth citing philosophers are John Langshaw Austin (1911–1960) for his works on speech acts (Austin 1955) (but also on the language of perception), Paul Grice (1913–1988) on linguistic pragmatics and dialogue (1957), and Stephen Toulmin (1922–2009) on argumentation (1958). All these writings have indirectly later influenced various research works in artificial intelligence, even if they first concern the philosophy of mathematics, the philosophy of spirit, the epistemology, or the philosophy of language.

Gregorius Itelson (1852–1926), André Lalande (1867–1963) and Louis Couturat at the IInd international congress of philosophy in Geneva in 1904 (Collective 1904) observed that they had been independently led to propose the French term "logistique" to refer to symbolic logic in its new algebraic and mostly algorithmic developments, and decided to adopt this new term. The term "logistique" in this sense is now completely disused. However it is worth noticing that it was still in use until the sixties by authors still inspired by *The Laws of Thought* by George Boole, who proposed treaties on "operational logistics", like the psychologist Jean Piaget (1896–1980), or the physicist Augustin Sesmat (1885–1957), or even the logician and philosopher Robert Blanché (1898–1975) (Piaget 1949; Sesmat 1951; Blanché 1970). This school of thought has been focusing not only on the formal dimension

[14] Carnap (1930) also underlines that from the tautological nature of deduction in modern logic "results the impossibility of any metaphysics which could pretend to conclude from experience to transcendent".

of reasoning, but also on cognitive aspects, common sense reasoning, plausible reasoning, argumentative reasoning, and has continued with the works of Jean-Blaise Grize (1922–2013), or even Nicholas Rescher (born in 1928) (Blanché 1966, 1973; Rescher 1976; Grize 1982). On the cognitive side, let us also cite the books by the mathematician Georg Polya (1887–1885) (1945, 1954) that analyze the discovery process of the solution of a mathematical problem, and emphasize the role played, in particular, by analogical reasoning. Besides, Kenneth Craik (1914–1947), a philosopher and a psychologist, first proposed the concept of mental models (1943). Finally, let us also name the American logician Jon Barwise (1942–2000) as a pioneer for advocating the cognitive interest of logical reasoning with diagrams in formal proofs, see, e.g., Takemura (2013).

Non-classical Logics

The first half of the XXth century also saw the introduction of various non-classical logics: multiple-valued, modal or probabilistic logics. These topics, already addressed in the pioneering works of a Scottish-born, French logician, Hugh MacColl (1837–1909) (Rahman and Redmond 2007), have been significantly developed during this period. Multiple-valued logics introduce new truth values for in particular reflecting the ideas of possibly, unknown, contradictory, or not applicable, or use intermediary truth values between true and false. Among the main contributors in this period, let us cite Jan Łukasiewicz (1878–1956) (Łukasiewicz 1913, 1930), Nicolai A. Vasiliev (1880–1940), Emil L. Post (1897–1954), Dmitrii A. Bochvar (1903–1990) (Bochvar 1984), Stephen Cole Kleene (1909–1994) (Kleene 1952), and Gregore Moisil (1906–1973) (Moisil 1972).[15] Let us also mention the mathematician Karl Menger (1902–1985), who in his works on stochastic geometry introduced a family of associative aggregation operators, called "triangular norms" (1942), which had a significant impact on multi-valued and fuzzy logics especially. Besides, the systematic study of modal logics starts with the works of Clarence Irving Lewis (1883–1964), before being equipped with a semantics in terms of possible worlds and accessibility relations by Saul Kripke (born in 1940) (Kripke 1959, 1963) and Jaako Hintikka (1929–2015) (Hintikka 1962). Georg von Wright (1916–2003) investigated their modeling capacity for many topics such as deontic logic or logic of action (von Wright 1951). Stanisław Leśniewski (1886–1939) coined "mereology" in 1927 to refer to a formal theory of part-whole relationships (Leśniewski 1992), while another Polish logician, Roman Suszko (1919–1979) introduced a non-Fregean logic (1968), inspired by Wittgenstein's *Tractatus*, which is a first order modal logic where 'situations' and 'facts' are not handled as predicates.

[15] Incidentally, it is notable that the first issue of one of the very first journals in computer science (Collective 1952), a journal dedicated to both computational machinery and theoretical logic, included in its table of contents a paper on a tri-valued logic by Bolesław Sobociński (1906–1980), which has turned out to be the logic of conditional objects (see chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" in this volume).

Probability, Decision Theory, Entropy

Regarding probability and decision theory, the period from the twenties to the sixties of the past century[16] saw a great deal of important works. Let us cite the economist John Maynard Keynes (1883–1946) supporter of a non-frequentist vision of probability closer to logic (1921),[17] the engineer Richard von Mises (1883–1953) defender of the frequentist point of view, the mathematicians Andreï Markov (1856–1922) and Andreï Kolmogorov (1903–1987) for their works respectively on stochastic processes and on the formalization of probability theory, Frank P. Ramsey (1903–1930) friend and translator of Wittgenstein, for his works on the decision problem in first-order logic and on the idea of subjective probability stemming from the idea of gamble (1931), Bruno De Finetti (1906–1985) who developed (independently from the previous one) the theory of subjective probability (1937, 1974) which is at the basis of Leonard Savage's (1917–1971) decision theory (1954) grounded in an axiomatic justification of expected utility, the statistician I. J. Good (1916–2009) who worked with Turing in cryptology during the Second World War and contributed to many subjects such as causality modeling, imprecise probability, or the possibility of constructing intelligent machines (Good 1961, 1962a, b, 1965), the mathematician, physicist and economist John von Neumann (1903–1957) (von Neumann 1958) who with Oskar Morgenstern (1902–1977) modeled decision making under risk and founded game theory (von Neumann and Morgenstern 1944), and the mathematician John Forbes Nash (1928–2015) for his equilibrium theory in non-cooperative games (Nash 1951) who received the Nobel Memorial Prize together with the economist John Harsanyi (1920–2000) especially known for his analysis of games of incomplete information (1967). Another Nobel laureate in economics, Kenneth Arrow (1921–2017) established an impossibility theorem (1951) of a collective, democratic and rational choice in social choice theory, providing a broader framework to Condorcet's paradox. We should also mention the economists Maurice Allais (1911–2010), Gérard Debreu (1921–2004), Lloyd Shapley (1923–2017) and Robert Aumann (born in 1930), the three latter being primarily mathematicians, who by some of their results in decision theory or in game theory have later impacted research on AI. Finally, beyond probability, the economist George L. S. Shackle (1903–1992), influenced by Keynes, has proposed a non-additive approach to decision under uncertainty (1949, 1961) based on the notion of degree of surprise (which will turn out to be an impossibil-

[16] In the previous period, some economists such as Léon Walras (1834–1910) and Carl Menger (1840–1921) (Karl Menger's father), as well as the logician William Stanley Jevons, introduce the notion of marginal utility in value theory for reflecting the interest a particular agent takes in a good or service, while Vilfredo Pareto (1848–1923), who advocated an ordinal view of utility, characterized situations where one cannot increase an agent's well-being without decreasing another agent's one, giving rise to the notion of optimum which bears his name; besides, he makes a distinction between logic actions like the ones studied in economy and non-logical actions studied in sociology (Pareto 1961).

[17] His father was also a distinguished economist, fond of logic (Keynes 1900).

ity degree in the sense of possibility theory[18]), while the philosopher L. Jonathan Cohen (1923–2006) has advocated a theory of "Baconian probabilities" in terms of measures of inductive support, which can be seen as a counterpart of possibility theory (1970). In addition, let us cite two philosophers of probability, Henry Kyburg (1928–2007), inventor of the so-called lottery paradox, and supporter of a logical standpoint of probability stemming from the notion of reference class, and Isaac Levi (born in 1930), defender of Shackle, and pioneer of belief revision and imprecise probabilities.

Entropy is a basic notion that can be encountered in different fields. Indeed there exists a parallel between entropy in statistical thermodynamics (as established by Ludwig Boltzmann (1844–1906) and J. Willard Gibbs (1839–1903)), and information-theoretic entropy (as proposed later by Ralph Hartley (1888–1970) and by Claude Shannon (1916–2001) the founder of information theory). Edwin Thompson Jaynes (1922–1998), who extensively contributed to the foundations of probability and statistical inference, initiated the maximum entropy interpretation of thermodynamics (Jaynes 1957). A measure of relative entropy, called Kullback–Leibler divergence (1951) (introduced by Solomon Kullback (1907–1994) and Richard Leibler (1914–2003)) is a key notion for evaluating the similarity between probability distributions. Besides, the idea of structural equations, an instrumental notion in the probabilistic analysis of causality, has emerged from the work of Sewall Wright (1889–1988) (Wright 1921), a geneticist and a statistician, before being fully theorized by the Nobel laureate economist Trygve Haavelmo (1911–1999) (Haavelmo 1943).

Cybernetics

In another vein, cybernetics (Wiener 1949) has emerged in the forties and the fifties as a new transdisciplinary field of research, under the leadership of Norbert Wiener (1899–1969) who considered it as "the scientific study of control and communication in the animal and the machine", where ideas coming from mechanics, biology and electronics interact (Rosenblueth et al. 1943). This field of investigation has been influenced by works in neurology which led Warren McCulloch (1898–1969) and Walter Pitts (1923–1969) (McCulloch and Pitts 1943) to propose the first model of formal neuron[19] (able to implement monotonic logical functions). Besides, the role of neurons in learning mechanisms was highlighted by the neuro-psychologist Donald O. Hebb (1904–1985) (Hebb 1949). The perceptron (1962), invented in 1957 by Frank Rosenblatt (1928–1971) can be seen as the simplest type of formal neurons network (perceptrons had one layer). Questioning the possibility of a Boolean representation of human intelligence activities, Rosenblatt's approach was sidelined for a while after the limited capabilities of perceptrons were discovered and emphasized by Marvin Minsky (1927–2016) and Seymour Papert (1928–2016) (1969).

[18]This theory was rediscovered independently by Lotfi Zadeh (1921–2017) in his approach to the representation of linguistic information, and for its qualitative counterpart, by the philosopher David Lewis (1941–2001) in his work on counterfactuals (1973).

[19]The neurons as basic units of the nervous system were discovered by the neuro-anatomist S. Ramon y Cajal (1852–1934) in the late1880's.

In Great Britain, William Ross Ashby (1903–1972), a psychiatrist and one of the main cyberneticists (1952, 1956), constructed a system called "homéostat" in 1948, which was made of interconnected control modules, able of self-adaptation with respect to its environment, and equipped with reinforcement learning capabilities. Besides, the neuro-physiologist William Grey Walter (1910–1977) built two "tortoises" robots (named "Elsie" and "Elmer"), capable of adaptative behavior in response to light stimuli. Let us also cite Gregory Bateson (1904–1980) for his hierarchical view of learning, influenced by cybernetics (1972). In Germany, cybernetics was supported by the philosopher and logician Gotthard Günther (1900–1984) (Günther 1957), and in France by Louis Couffignal (1902–1967) who, as a specialist in machine-based computation, became highly interested in the idea of "thinking machines" (1952), after meetings with the neuro-physiologist Louis Lapicque (1866–1952) (1943). Let us note that from the beginning research works in cybernetics raised great interest and questions well beyond laboratories (Wiener 1950; De Latil 1953; Delpech 1972; Dubarle 1948), on the use of science, while others, more radically, already worried about the dangers for humanity caused by the development "with a frightening speed" of "the civilization of machines" (Bernanos 1947).

Information Theory and Computability

Apart from the boom of cybernetics, landmark works in the years preceding the official birth of AI are those of Claude Shannon (1916–2001) on the foundation of information theory (after his pioneering works on the use of algebra and logic for describing relay and switching circuits (Shannon 1938)), those of John von Neumann (1903–1957) on the architecture of computer systems and on the theory of automata (1966), and those of Alan Turing (1912–1954) on the functions that can be computed by machine. These three authors were also much involved in discussing questions related, on the one hand, to the possibility of building "thinking machines", and, on the other hand, to the comparison of the functioning of human brain with the first computers that just came out at that time and were essentially devoted to numerical computing (Shannon 1950, 1956; von Neumann 1956; Turing 1950, 1956). The year 1950 sees the publication of several papers referring to the idea of thinking machines: those just cited, by Shannon (on the basic principle of chess game programming), and by Turing (where his famous test is proposed for determining whether a machine demonstrates intelligence or not), but also a paper (1950), by the young Lotfi Zadeh, future father of fuzzy logic. Let us also have a particular mention for Konrad Zuse (1910–1995), German pioneer of the transition from mechanical calculators to modern computers, and author in 1945 of a computing program for chess game, and the English computer scientist Christopher Strachey (1916–1975), author in 1951 of a program able to play checkers (Link 2012; Strachey 1952).

Let us also emphasize that Alan Turing, apart his famous "Turing test", foresaw the importance of machine learning (especially reinforcement learning) in his 1950 paper (and in a report before, see (Turing 1948)). Alan Turing, who died two years before the first AI meetings at Dartmouth College, may be certainly regarded as the main grandfather of AI. At that time, the intelligence of a machine was mainly considered in terms of computation and memory capabilities required for its implementation,

or is influenced by cybernetics, as for example in the work on the representation of events in neuronal networks by Stephen Kleene (1956) (who contributed before to the characterization of recursive functions).

The 1950s also saw the very beginning of research on automatic translation with Yehoshua Bar-Hillel (1915–1975), linguist (1954) and mathematician, student of Carnap (1952), who organized in 1952 the first *International Conference on Machine Translation*; he also pioneered information retrieval (1963) (of which, since 1945, the engineer Vannevar Bush (1890–1974) prophesied the rise with the advent of computers 1945).

Literature and Cinema

Before coming to modern AI, let us end this section by mentioning the impact of machines and computer science on literature and cinema. At the beginning of the XXth century, while the role of machines in industry was increasing, robots started to appear in literature. For instance, the protagonist of a short story (1913) by the humanistic philosopher Miguel de Unamuno (1864–1936), visits a city, "Mecanópolis", exclusively inhabited by machines. This narrative is in line with the novel *Erewhon* (1872) by the English writer Samuel Butler (1835–1902); it is a satire of the Victorian era where the author imagines that machines could develop consciousness by a kind of Darwinian selection. Another example is provided by the play *Poupées Électriques*, (i.e., "Electric Dolls") published in French, with two puppet characters (in French, "fantoche"), where the Italian writer Filippo Tommaso Marinetti (1876–1944), founder of the Futurist movement (1909), has sketched a parallel between humans and the electric puppets, able to react, built by one of the characters of the play. The term "robot" (coming from Czech "robota", which means "heavy work") was used for the first time by Karel $\tilde{C}$apek (1890–1938) in his play *R. U. R. (Rossum's Universal Robots)* (1921). The robots have then inspired a whole trend in science fiction literature starting, in particular, with the I. Asimov's (1920–1992) collection of short stories *I, Robot* (1950). Besides, the effervescence of discussions aroused by cybernetics inspired writers such as Elsa Triolet (1896–1970) in her novel *L'âme* (1963), or Henry Certigny (1919–1995) who in *Les automates* (1954), both renewing the tradition of animated doll stories of the XVIIIth century in different manners. Marvin Minsky (1927–2016), one of the fathers of AI, has also contributed later to this literary trend (Harrison and Minsky 1992). Mentioning the influence of cybernetics and AI on more contemporary science fiction literature is outside the scope of this brief historical overview.[20]

From a completely different viewpoint, works on combinatorial reasoning and advances in computer science have offered new opportunities to literary creation. Thus, the writer Raymond Queneau (1903–1976) and François Le Lionnais (1901–1984), engineer by training, created a research group, the OULIPO ("OUvroir de Littérature POtentielle")[21] in 1960, which developed works in experimental literature

[20]The reader could consult the website http://en.wikipedia.org/wiki/Artificial_intelligence_in_fiction.

[21]http://www.oulipo.net/.

that rely on the use of syntactic and semantic constraints. It is worth noticing that from its beginning, this group included writers such as Italo Calvino (1925–1985) or Georges Perec (1936–1982), but also scientists such as the mathematician Claude Berge (1926–2002), one of the modern founders of combinatorics and graph theory. This group later inspired the birth of another group named ALAMO ("Atelier de Littérature assistée par la Mathématique et les Ordinateurs"),[22] founded in 1981 by Paul Braffort (1923–2018),[23] engineer in cybernetics and by the poet Jacques Roubaud (born in 1932), and more oriented towards computer-aided literary creation.

From the very beginning of cinema, fantasy and science fiction literatures inspired numerous movie adaptations. We limit ourselves to some emblematic movies.[24] Among the first adaptations, let us cite *Gulliver's Travels Among the Lilliputians and the Giants* (1902) by G. Méliès (1861–1938), *Frankenstein* (1910) by J. S. Dawley (1877–1949), Fritz Lang's (1890–1976) *Métropolis* (1927) adapted from the Thea von Harbou's (1888–1954) eponymous novel (1926). The transition to talking pictures generated further adaptations such as *Frankenstein* (1931) and then *Bride of Frankenstein* (1935) by J. Whale (1889–1957), parodied in M. Brooks' (born in 1926) *Young Frankenstein* (1974), *Pinocchio* (1940) by W. Disney (1901–1966), S. Kubrik's (1928–1999) *2001: A Space Odyssey* (1968), inspired from the A. C. Clarke's (1917–2008) short story *The sentinel* (1951), R. Scott's (born in 1937) *Blade Runner* (1982) adapted from the Ph. K. Dick's (1928–1982) novel *Do Androids Dream of Electric Sheep?* (1968), S. Spielberg's (born in 1946) *AI. Artificial Intelligence* (2001) inspired from the B. Aldiss' (1925–2017) short story *Supertoys Last All Summer Long - and Other Stories of Future Time* (2001).

7 The Beginnings of the AI Era

As already said in the introduction of this chapter, the birth certificate of Artificial Intelligence corresponds to a two-month meeting program with ten participants held at Dartmouth College (Hanover, New Hampshire, USA) in the summer of 1956, led by two young researchers[25] who, for different reasons, would then strongly mark

[22] http://www.alamo.free.fr/.

[23] Paul Braffort has also been the author of the first French monograph on AI (1968). We are very glad that he kindly accepted to write the foreword of the Volume 3 of this treatise.

[24] For more details, the reader may, for example, consult the website http://homepages.inf.ed.ac.uk/rbf/AIMOVIES/AImovai.htm.

[25] With the help of Claude Shannon and Nathaniel Rochester (1919–2001). The latter was the designer of the IBM701 computer and the author of the first program in assembly language, and had interests close to AI (Rochester et al. 1956). The request for support, already titled "A proposal for the Dartmouth summer research project on Artificial Intelligence" dates back from the previous summer and was jointly signed by McCarthy, Minsky, Rochester and Shannon (McCarthy et al. 2006). The six other participants were Trenchard More, Allen Newell, Arthur Samuel, Oliver Selfridge, Herbert A. Simon, and Ray Solomonoff (1926–2009). This last researcher, who was a pioneer of the concept of algorithmic probability, circulated a report (1956) the same year, which was the beginning of his future theory of universal inductive inference and one of the first approaches to probability-based

the development of the discipline: John McCarthy (1927–2011) and Marvin Minsky (1927–2016), the former advocating a purely logical view of knowledge representation (1996, 1990), the latter coming from neural nets and reinforcement learning (1954) and favoring the use of structured representations of stereotypes of situations (alias "frames" (Minsky 1975)) that may include different types of information. It was during these meetings that the expression "Artificial Intelligence" (defended by McCarthy) was used for the first time in a systematic way to designate the new field of research. However, it was far from being a consensual term, some of the researchers participating in the program seeing there a complex processing of information, and nothing more. Among them were Alan Newell (1927–1992) and Herbert Simon (1916–2001), who, as well, were going to have a significant impact on the development of AI.

First AI Programs

It was indeed in 1956 that Newell and Simon, in collaboration with John Cliff Shaw (1922–1991), presented a first computer program, the "Logic Theorist" capable of demonstrating logical theorems (such as those appearing at the beginning of the *Principia Mathematica* of Whitehead and Russell) (Newell and Simon 1956; Newell et al. 1957), before presenting a "General Problem Solver", or "GPS") (Newell et al. 1959) based on the evaluation of the difference between the situation at which the solver has arrived and the goal it has to achieve (aka means-end analysis). Another participant in the Dartmouth encounters, Oliver Selfridge (1926–2008) is a pioneer (already cited) of pattern recognition[26] and machine learning (1959) (see also his work with Minsky, see (Minsky and Selfridge 1961)). He has also been at the origin of the notions of "pattern matching" (Selfridge 1959) and "daemon" (which allows us to associate some pieces of code with the filtering process), two notions that have proven very useful for knowledge-based systems. Herbert Gelernter (1929–2015) achieved the first automated theorem prover ("GTP") in elementary geometry (1959). At the same time, Robert Lindsay developed SADSAM (which stands for "Syntactic Appraiser and Diagrammer Semantic Analyzing Machine"), a program capable of establishing and reasoning on relations between items in a speech (1963), while James Slagle (born in 1934) conceived a symbolic integration program (SAINT for "Symbolic Automatic INTegrator") (1963), and the program "Student" developed by Daniel Bobrow (1935–2017) solved elementary problems of arithmetic, expressed in natural language (1964). For more details, one can find a collection of articles representative of early works in AI until the early 1960's in Feigenbaum and Feldman (1963). In United Kingdom, Donald Michie (1923–2007), a biologist and a pioneer of artificial intelligence, developed the Machine Educable Noughts And Crosses Engine

machine learning in artificial intelligence (this latter phrase is used in his report of August 1956!). As to Trenchard More, he was preparing a thesis on the concept of natural deduction which he later defended (1962).

[26]Pattern recognition is born in the same time as AI (Dinneen 1955; Selfridge 1955; Clark and Farley 1955). Moreover Selfridge's work has in turn influenced the work of the cyberneticians Jerome Lettvin (1920–2011), Humberto Maturana (born in 1928), Warren McCulloch, and Walter Pitts (Lettvin et al. 1959).

(MENACE), one of the first programs capable of learning to play the game of Tic-Tac-Toe (1963). He was also the founder and editor-in-chief of the *Machine Intelligence* series, of which nineteen volumes were published,[27] which was especially influential in the late sixties and seventies (Collins and Michie 1967; Dale and Michie 1968; Meltzer and Michie 1968–1972; Elcock and Michie 1977; Hayes et al. 1979).

Among the different works that marked the beginnings of AI, one can still mention the program "Analogy" (1964) developed by Thomas G. Evans (born in 1934), that was capable to find out, as in a IQ test, the fourth geometrical figure among several possible choices in order to complete a series of three (which required a conceptual representation of the figures). Processing texts or dialogues in natural language also concerned AI very early, either for trying to understand something of their contents, or for generating sentences automatically. The program "ELIZA" (1966) by Joseph Weizenbaum (1923–2008), was able in 1965 to dialogue in natural language by identifying key phrases in sentences and reconstructing sentences from them by filling in ready-made structures (it succeeded for a moment in duping some human users who thought they were dealing with a human!). Yet "ELIZA" did not construct any representation of the sentences of the dialogue and therefore had no understanding of the dialogue at all. The program "SHRDLU" (1971) by Terry Winograd (born in 1946) was the first to construct such representations and to exploit them in dialogues concerning the relative positions of blocks in a simplified block world.

Programming Languages

In order to write such programs more easily, programming languages devoted to the symbolic processing of information were necessary. Specified as early as 1958 by McCarthy, and inspired from the λ-calculus of Alonzo Church, LISP (for "LISt Processing") developed in the 1960s (McCarthy et al. 1962) quickly became a reference language for AI programming. While LISP is a functional programming language, PROLOG (for PROgramming in LOGic) is, as its name suggests it, a logic programming language (it is based on the first-order predicate calculus) (Colmerauer 1978; Colmerauer and Roussel 1992). PROLOG appeared in the 1970s and became another key language for AI programming.[28] Other pioneering works such as Carl Hewitt's one about the actor model of computation (1969, 2009) also contributed to the development of AI programming. This period was also marked by a certain amount of research which based knowledge representation on logic. Let us mention the situation calculus (McCarthy and Hayes 1979), a formal framework for reasoning on dynamic worlds, the application of automated theorem proving to query answering systems (Green 1979), and the STRIPS ("STanford Research Institute Problem Solver") language for planning, and its algorithm based on the means-end analysis (as in the already mentioned "General Problem Solver") (Fikes and Nilsson 1971). Let us not forget the progress achieved during this period in automated theorem proving (1971), especially with the work of Woodrow Bledsoe (1921–1995).

---

[27] http://www.doc.ic.ac.uk/~shm/MI/mi.html.

[28] Alain Colmerauer (1941–2017), the father of the PROLOG programming language, was also the inventor of the founding principles of constraint logic programming. We are very glad that he kindly accepted to write the foreword of the Volume 2 of this treatise.

Problem Solving and Cognitive Issues

While logic plays a key role in knowledge representation, problem solving has been influenced by cognitive psychology (Newell and Simon 1972). The psychologist Roger Schank (born in 1946) is in particular at the origin of the idea of case-based reasoning (Schank and Abelson 1977). The need of control structures for solving problems in order to avoid a scattered search, or, on the contrary, to go too far into a dead end, led to the use of if-then rules and sophisticated filtering procedures (see for instance Moore and Newell 1974). Newell was also influenced by George Pólya (1887–1985) and the importance of the concept of analogy in the search for solutions (Newell 1981). Let us also mention the "Logo" programming language (created in 1967 by D. Bobrow, W. Feurzeig, S. Papert and C. Solomon), related to LISP and conceived as an interactive learning tool for children (a small turtle-robot allowed them to visualize the result of actions) (Papert 1980), a project inspired by the works of Jean Piaget. Another lasting influence on AI (and on Theoretical Computer Science) was the one of the linguist Noam Chomsky (born in 1928) in the field of formal language structures and grammars.

Checkers and Chess

As has been said, AI was interested, even before its name had been found, by the development of programs capable of playing checkers or chess. The first programs, notably those by Arthur Samuel (1901–1990) for the checkers (1959),[29] and by Alex Bernstein (1958) for chess, appeared in the early 1960s. Over the decades, such programs, such as the "MacHack" program by Richard Greenblatt (born in 1944) in the late 1960s, succeeded in beating players of increasingly higher levels. In the 1970s, research in this field (Berliner et al. 1977) is marked by the idea of endowing the computers with capacities for the implementation of sophisticated strategies, evolving dynamically (as in the work of Hans Berliner (1929–2017)).[30] However, it is first and foremost the computational power of a computer capable of exploring gigantic combinatorial spaces that finally overcome the world champion of the discipline (victory of the Deep Blue computer on Gary Kasparov in 1997).

Expert Systems

The 1970s and early 1980s were marked by the achievement of many expert systems (Smith 1984) where pieces of knowledge about a specialized field were expressed as if-then rules, and applied to any set of facts describing a situation on which the system must produce conclusions. The first ones were DENDRAL in organic chemistry (Lindsay et al. 1980), MYCIN in medicine (Buchanan and Shortliffe 1984), HEARSAY-II in speech understanding (Erman et al. 1980), PROSPECTOR in geology (Duda et al. 1976, 1981). Alongside the mainstream of AI, let us mention the

[29] Samuel's program initiated the use of tree-pruning procedures of alpha-beta type, and already had skills to learn its cost function.

[30] In relation to theorem proving and then to chess, let us also cite Jacques Pitrat (1970, 1977), who among other things highlighted the role of metacognition in problem solving and learning processes (2000). We are very glad that he kindly accepted to write the foreword of this volume.

parallel development of "fuzzy" rule-based systems, in particular for interpolation purposes, where rules have a graded applicability due to their representation based on fuzzy sets (Zadeh 1965). The proper handling of such rules are part of a theory of approximate reasoning, itself based on the possibility theory by Zadeh (1978). Fuzzy rule-based systems quickly found applications for the automatic control of many different devices, thanks to the pioneering work of E.H. Mamdani (1942–2010) (Mamdani and Assilian 1975; Dubois and Prade 2012) (fuzzy rules were representing expertise in piloting the device under consideration, in case no mathematical model was available).

Constraints, Vision and Natural Language

Among the remarkable advances of the 1970s, let us also cite heuristic search algorithms (Hart et al. 1968), and systems exploiting constraints by propagating them, as in the approach of David Waltz (1943–2012) to recognize in a picture the lines corresponding to the edges of solids and their relative positions (1975), which would later extend to many other domains where constraint representation is naturally required. It was also the beginnings of the research activities in computer vision, marked by the work (1982) of David Marr (1945–1980), in collaboration with Tomaso Poggio (born in 1947). In this work, vision is understood as an information processing process with three distinct, yet complementary levels of analysis (a computational level, an algorithmic / representational level, and an implementational / physical level). Another research area, directly related to AI, and for which some important achievements have been got during this period, is that of natural language understanding, with the works of Robert Schank (already cited) (1973), William Woods (born in 1942) (1975), Yorick Wilks (born in 1939) (1972), and the controversy about procedural semantics (Fodor 1978; Johnson-Laird 1978).

Mobile Robots and Planning

The 1970s were also the years of the first experiments with mobile robots (especially, the "Shakey" robot at SRI (Menlo Park, Ca), see (Raphael 1976)), which jointly posed problems of computer vision (Nevatia and Binford 1977), knowledge representation, and motion planning.[31] This was the period when the first theoretical works on planning appeared, such as those of Earl D. Sacerdoti (born in 1948) (1977). Ten years later, at MIT, Rodney Brooks (born in 1954), has been interested in the design of robots that are reactive to their immediate environment, but act without using a representation of the world in which they live (1989). This research is contemporaneous with the development of the study of multi-agent systems in AI, the premises of which being in Minsky's writings (1986), see also Georgeff (1983).

In the 1970's, the first academic and institutional criticisms of AI also emerged in the United States and in Great Britain (Lighthill 1973; Hendler 2008), and this has had a significant impact for a decade on the funding of AI research. AI was accused

[31] Other experiments with mobile robots of that time are the Stanford "Cart" project in the late 1960s, and a bit later the French "HILARE" project (Giralt et al. 1979) and the Carnegie Mellon University rover (Moravec 1982).

of not keeping the excessive promises made at the beginning, because of the very limited power of the computers and also of the lack of maturity of the area.[32]

This historical tour stops voluntarily at the beginning of the 1980s, considering that the reader will be able to find additional elements for the more recent history of each of the many facets of AI in the different chapters of this book. For a global picture of AI research in the 1980s, let us also mention (Barr and Feigenbaum 1981, 1982; Cohen and Feigenbaum 1986, 1990; Grimson and Patil 1987). As can be seen in this brief historical overview, AI has developed largely in the United States, before becoming a research area in Europe in the 70s (even in the 60s for the UK), and then in Asia.

8 Conclusion

There exist a number of books and documents relating various aspects of the history of AI, including the modern one, to which the interested reader is referred for further details (Anderson 1964; McCorduck 1979; Rose 1984; Pratt 1987; Kurzweil 1990; Crevier 1993; Nilsson 2010; Buchanan 2005; Buchanan et al. 2013). However, this chapter substantially departs from the above references by attempting to draw a large historical picture of the many events that led to the emergence of AI. It is clear that such an entreprise, done for the first time (as far as we know) faces unavoidably the risk of missing noticeable names. We apologize in advance for such omissions.

AI is not just a matter of technology. As any science, its concerns have roots far away in the past. Our objective was to give here a picture of AI rooted in a long tradition of research, and to show the synergies that are still at work between imagination, science and technology. It is in this will that the main originality of this chapter lies. It may be important to know the history of AI for better understanding from where it comes and where it goes, especially if we consider the present effervescence, and the questions and fears that are induced.

Acknowledgements The authors are especially indebted to Jens Lemanski who provided them with valuable references and comments, in particular on Johann Christian Lange and the history of Euler-type diagrams.

[32] In one of the answers to this report, that of Christopher Longuet-Higgins (1923–2004), one can find for the first time, the expression "cognitive science (s)" (Hünefeldt and Brunetti 2004). Longuet-Higgins was a co-founder with Richard Gregory (1923–2010) and Donald Michie (1923–2007) of the Department of Machine Intelligence and Perception of the University of Edinburgh in 1967.

References

Abraham M, Gabbay DM, Hazut G, Maruvka YE, Schild U (2010) Studies in talmudic logic. Vol. 2: the textual inference rules Klal uPrat (How the Talmud defines sets). College Publications

Abraham M, Gabbay DM, Schild UJ (2010–2013) Studies in talmudic logic. Vol. 1: Non-deductive inferences in the Talmud. Vol. 3: Talmudic deontic logic. Vol. 5: Resolution of conflicts and normative loops in the Talmud. Vol. 10: Principles of Talmudic logic. College Publications

Abraham M, Belfer I, Gabbay DM, Schild U (2011–2016) Studies in Talmudic logic. Vol. 4: Temporal logic in the Talmud. Vol. 7: Delegation in Talmudic logic. Vol. 8: Synthesis of concepts in the Talmud. Vol. 9: Analysis of concepts and states in Talmudic reasoning. Vol. 11: Platonic realism and Talmudic reasoning. Vol. 12: Fuzzy logic and quantum states in Talmudic reasoning. Vol. 13: Partition problems in Talmudic reasoning. College Publications

Akrami M (2017) From logic in Islam to islamic logic. Log Univers 11(1):61–83

Anderson AR (ed) (1964) Minds and machines. Prentice-Hall, Includes: Introduction (A.R. Anderson), Computing machines and intelligence (A.M. Turing), The mechanical concept of mind (M. Scriven), Minds, machines and Gödel (J.R. Lucas), The imitation game (K. Gunderson), Minds and machines (H. Putman), The feelings of robots (P. Ziff), Professor Ziff on robots (J.J.C. Smart), Robots Inc (N. Smart)

Anselm of Cantorbery (2001) Proslogion. Hackett Publishing Company, Indianapolis, with the replies by Gaunilo and then by Anselm. Translation and introduction by Th. Williams

Apollonius of Rhodes (1959) The Voyage of argo: the argonautica. Penguin Classics

Aquinas T (1975) Summa contra gentiles. Book one: god. University of Notre Dame Press

Aquinas T (2006) In: Leftow B, Davies B (eds) Summa theologiae, questions on god. Cambridge University Press, Cambridge

Arnauld A, Nicole P (1662) La Logique ou l'Art de Penser contenant, outre les règles communes, plusieurs observations nouvelles, propres à former le jugement. Flammarion, Champs, 1978; English translate: Logic or the art of thinking: containing, besides common rules, several new observations appropriate for forming judgement. Cambridge University Press, Cambridge, 1996

Arrow KJ (1951) Social Choice and Individual Values, 2nd edn. Yale University Press, New Haven, 1963

Ashby WR (1952) Design for a brain. Chapman & Hall, London

Ashby WR (1956) An introduction to cybernetics. Chapman & Hall, London

Asimov I (1950) I, Robot. Gnome Press, New York

Augustine of Hippo (1995) Against the academicians. The teacher. Hackett Publishing Company, Indianapolis, translated with introduction and notes by P. King

Austin JL (1955) How to do things with words. Oxford University Press, Oxford, 1962; The William James lectures, Harvard University, Cambridge, 1955

Bacon F (1605) Of the proficience and advancement of learning, divine and human. Oxford University Press, Oxford, 1974; Everyman Paperbacks, 1991

Bacon R (2009) The art and science of logic [Summulae Dialectices]. Mediaeval sources in translation, No. 47, Pontifical institute of mediaeval studies, Toronto, translated with introduction and notes by Th. S. Maloney

Bain A (1870) Logic. Part first: deduction. Part second: induction. Longmans, Green and Co., London, 784 p

Bar-Hillel Y (1954) Indexical expressions. Mind 63(251):359–379

Bar-Hillel Y (1963) Is information retrieval approaching a crisis? American documentation 14(ii):95–98

Barnes J (1983) Terms and sentences: theophrastus on hypothetical syllogisms. Proc Br Acad 69:279–326

Barr A, Feigenbaum EA (eds) (1981) The handbook of artificial intelligence, vol I. William Kaufman, Los Altos

Barr A, Feigenbaum EA (eds) (1982) The handbook of artificial intelligence, vol II. Addison-Wesley, Menlo Park

Bateson G (1972) Steps to an ecology of mind: collected essays in anthropology, psychiatry, evolution, and epistemology. Chandler Publishing Company, San Francisco
Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond 53:370–418, by the Late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a Letter to John Canton, A.M.F.R.S
Bellhouse DR (2000) De vetula: a medieval manuscript containing probability calculations. Int Stat Rev 68(2):123–136
Berliner HJ, Greenblatt R, Pitrat J, Samuel A, Slate D (1977) Computer game playing. In: Reddy R (ed) Proceedings of the 5th international joint conference on artificial intelligence, Cambridge, MA, pp 975–982
Bernanos G (1947) La France contre les Robots. Robert Laffont
Bernstein A, De V Roberts M (1958) Computer vs. chess-player. Sci Am 198:96–105
Bessot D, Lanier D, Le Goff JP, Leparmentier J, Levard M, A-M Sainson DT, Domain R (2006) L'Espérance du Hollandais ou le Premier Traité de Calcul du Hasard. Ellipses
Blanché R (1966) Structures Intellectuelles. Essai sur l'Organisation Systématique des Concepts. Librairie philosophique J. Vrin, Paris
Blanché R (1970) La Logique et son Histoire d'Aristote à Russell. Amand Colin, collection U, 2nd ed. augmented with a chapter "La Logique depuis Russell" by J. Dubucs, 1996
Blanché R (1973) Le Raisonnement. Presses Universitaires de France, Bibliothèque de Philosophie Contemporaine, Paris
Bledsoe WW (1971) Splitting and reduction heuristics in automatic theorem proving. Artif Intell 2(1):55–77
Bobrow DG (1964) A question-answering system for high school algebra word problems. In: Proceedings of the fall joint computer conference (AFIPS '64), Part I, 27–29 October 1964. ACM, pp 591–614
Bobzien S (2004) Peripatetic hypothetical syllogistic in Galen. Rhizai J Anc Philos Sci 2:57–102
Bocheński IM (1947) La Logique de Théophraste. Librairie de l'Université de Fribourg en Suisse
Bochvar DA (1984) On the consistency of a three-valued logical calculus. Topoi 3(1):3–12, translate by M. Bergmann of a 1938 article
Bolzano B (1837) Wissenschaftslehre. Versuch einer ausfürlichen und grösstentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter. J E von Seidel, Sulzbach, edited and translated by R. George, as "Theory of Science. Attempt at a detailed an in the main novel exposition of logic with constant attention to earlier author", University of California Press, California, 1972
Boole G (1847) The mathematical analysis of logic, being an essay towards a calculus of deductive reasoning. Macmillan, Cambridge
Boole G (1854) An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities. Macmillan, Cambridge, reprinted by Dover, New York, 1958
Boureau-Deslandes AF (1742) Pigmalion, ou la statue animée. Samuel Harding, London, Pigmalion, oder, Die belebte Statüe translated by J.J. Bodmer, Hamburg Martini 1748
Braffort P (1968) L'Intelligence Artificielle. P.U.F., Paris
Braithwaite RB (1932) Lewis Carroll as logician. Math Gaz 16(219):174–178
Brooks R (1989) A robot that walks; emergent behaviors from a carefully evolved network. Neural Comput 1(2):253–262
Bru MF, Bru B (2018) Le jeu de l'infini et du hasard - Vol. 1: Les probabilités dénombrables à la portée de tous ; Vol. 2: Les probabilités indénombrables à la portée de tous. Annexes et appendices. Presses Université de Franche-Comté, Besançon
Buchanan BG (2005) A (very) brief history of artificial intelligence. AI Mag 26(4):53–60
Buchanan BG, Shortliffe EH (eds) (1984) Rule-based expert systems - MYCIN experiments of the Stanford heuristic programming. Addison-Wesley, Reading
Buchanan BG, Eckroth J, Smith R (2013) A virtual archive for the history of AI. AI Mag 34(2):86–98
Bush V (1945) As we may think. Atl Mon 176(1):101–108

Busquets J (2006) Logique et Langage: Apports de la Philosophie Médiévale. Presses Universitaires de Bordeaux
Butler S (1872) Erewhon, or, over the range. Trubner & Co, London, includes three chapters, numbered 23, 24, 25, intitled "The Books of Machines", first published as articles since 1863
Byrne O (1841) The doctrine of proportion clearly developed: on a comprehensive, original, and very easy system; or, the fifth book of Euclid simplified. J. Williams, London
Byrne O (1847) In: Oechslin W (ed) The first six books of the elements of Euclid in which coloured diagrams and symbols are used instead of letters for the greater ease of learners. William Pickering, London, reprinted by Taschen Gmbh, 2013
Candaux JD (1993) Monsieur de Lubières, encyclopédiste. Recherches sur Diderot et sur l'Encyclopédie 15:71–96
Capek K (1921) Rossumovi Univerzál Roboti - R.U.R. (Rossum's Universal Robots). English translate R.U.R. and The Insect Play (with J. Capek), Oxford Paperbacks, 1963
Carnap R (1930) Die alte und die neue logik. Erkenntnis 1(1):12–26, English translate by I. Levi, The old and the new logic. In: Ayer AJ (ed) Logical positivism. Free Press, 1959, pp 133–146
Carnap R (1942) Meaning and necessity: a study in semantics and modal logic, 2nd edn. University of Chicago Press, Chicago, 1956
Carnap R, Bar-Hillel Y (1952) An outline of the theory of semantic information. MIT, Research Laboratory of Electronics, Technical report 247
Carroll L (1896) Symbolic logic. Part 1. Elementary. Macmillan and Co., London, Part 2. Advanced
Certigny H (1954) Les Automates. Gallimard, Paris
Chassay JF (2010) L'imaginaire de l'être artificiel. Presses Universitaires du Québec
Clair J, Szeemann H (eds) (1975) Junggesellenmaschinen / Les Machines Célibataires. Alfieri, Venezia, exhibition catalogue, Italian/English version Le Macchine Celibi / The Bachelor Machines, Rizzoli, New York, 236 p
Clark WA, Farley BG (1955) Generalization of pattern recognition in a self-organizing system. In: Proceedings of the Western joint computer conference, 1–3 March 1955, Institute of Radio Engineers, New York, pp 86–90
Cohen J (1968) Les Robots Humains dans le Mythe et dans la Science. Librairie philosophique Vrin
Cohen LJ (1970) The implications of induction. Methuen, London
Cohen PR, Feigenbaum EA (eds) (1986) The handbook of artificial intelligence, vol III. Addison-Wesley, London
Cohen PR, Feigenbaum EA (eds) (1990) The handbook of artificial intelligence, vol IV. Addison-Wesley, London
Collective (1904) Logique et Philosophie des Sciences. Séances de Section et Séances Générales. II^e Congrès de Philosophie - Genève. Revue de Métaphysique et de Morale T. XII:1037–1046
Collective (1952) The foundations of computing machinery (J.D. Goodell); The realization of a universal decision element (T. Lode); Axiomatization of a partial system of three-value calculus of propositions (B. Sobocinski). J Comput Syst 1(1):1–55, publ. by The Institute of Applied Logic, St. Paul MN
Collins NL, Michie D (eds) (1967) Machine intelligence 1. Oliver & Boyd, Edinburgh & London, preface by Sir Edward Collingwood
Colmerauer A (1978) Metamorphosis grammars. In: Bolc L (ed) Natural language communication with computers. LNCS, vol 63. Springer, Berlin, pp 133–189
Colmerauer A, Roussel A (1992) La naissance de Prolog. Internal report, Groupe Intelligence Artificielle, Faculté des Sciences de Luminy, Université Aix-Marseille II, France
Condorcet N (1785) Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix. Reprinted by American Mathematical Society, 1972
Conti A (2017) Paul of venice. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. Metaphysics Research Lab, Stanford University
Couffignal L (1952) Les Machines à Penser. Les Editions de Minuit, Paris, 2nd edn. revised, 1964

Couturat L (1901) La Logique de Leibniz: d'après des documents inédits. Félix Alcan, Paris, reprinted by OLMS, Hildesheim, 1969 and 1985
Couturat L (1903) Opuscules et Fragments Inédits de Leibniz. Extraits des manuscrits de la bibliothèque royale de Hanovre. Félix Alcan, Paris, 1903; reprinted by Olms, 1966
Couturat L (1905) L'Algèbre de la Logique. Gauthier-Villars, Paris
Craik KJW (1943) The nature of explanation. Cambridge University Press, Cambridge, reprinted 1967
Cramer G (1745) Cours de Logique. Bibliothèque Publique et Universitaire de Genève, MS Trembley 34, 348 pages, structured in 576 paragraphs, unpublished manuscript; partly reproduced (89 p., paragraphs 1–10, 260, 448–547), by Th. Martin, in Journl électronique d'Histoire des Probabilités et des Statistiques 2(1), 2006, 6
Crevier D (1993) The tumultuous history of the search for artificial intelligence. Basic books, HarperCollins Publishers, New York
Crossley JN (2005) Raymond Llull's contributions to computer science. Technical report (13 p), Monash University, Clayton, Australia
Curley AJ (1996) Augustine's critique of skepticism. A study of Contra Academicos, Peter Lang, New York
Dahan-Dalmedico A (1986) Un texte de philosophie mathématique de Gergonne. Revue d'Histoire des Sciences 39(2):97–126
Dale E, Michie D (eds) (1968) Machine intelligence 2. Oliver & Boyd, London
Davis M, Putnam H (1960) A computing procedure for quantification theory. J ACM 7(3):201–215
De Borda JC (1781) Mémoire sur les élections au scrutin. Mémoires de l'Académie Royale des Sciences, pp 657–664
de Bovelles C (1510) Ars oppositorum. Translated in French, L'art des opposés by P. Magnard, Vrin, Paris, 1984
De Castillon F (1804) Réflexions sur la logique. Mém de l'Acad Royale des Sciences et Belles-Lettres de Berlin, pp 29–49
De Castillon F (1805) Mémoire sur un nouvel algorithme logique. Mém de l'Acad Royale des Sciences et Belles-Lettres de Berlin, pp 3–24
De Ceriziers R (1650) Le Philosophe François. Antoine Molin, Lyon
De Chousy, comte D (1883) Ignis. Rééd. Col. Ressources 114, Slatkine, 1981, English translate Ignis, the Central Fire, Hollywood Comics, 2009
De Finetti B (1937) La prévision: ses lois logiques, ses sources subjectives. Annales de l'Institut Poincaré 7:1–68
De Finetti B (1974) Theory of probability. Wiley, New York
De Lacy PH, De Lacy EA (1941) Philodemus: on methods of inference. A study in ancient empiricism. American Philological Association, Philadelphia, with translation and commentary
De Latil P (1953) La Pensée Artificielle. Introduction à la Cybernétique. Gallimard, L'Avenir de la Science 34
De Moivre A (1718) Doctrine of chances, or a method of calculating the probability of events in play, 3rd edn. Printed by W. Pearson for the Author, London, 1756
De Morgan A (1847) Formal Logic: or, the calculus of inference, necessary and probable. Taylor & Walton, London
De Morgan A (1868) On the syllogism and other logical writings. Routledge & Kegan Paul, London, articles 1846–1868; edited with an Introduction by P. Heath, 1966
De Unamuno M (1913) Mecanópolis. In: García Blanco M (ed) Obras Completas, Vol. 2: Novelas. Escélicer, Madrid, pp 833–836, 1966
De Villiers de l'Isle-Adam A (1886) L'Eve future. Charpentier, English translate The Future Eve, Fantasy and Horror Classics, 2011
Delboeuf C (1876) Logique algorithmique: 1. Exposé de la logique déductive au moyen d'un système conventionnel de signes. Deuxième partie: Caractères généraux d'un algorithme. Troisième partie. Revue Philosophique de la France et de l'Etranger 2:225–252, 335–355, 545–595. Also published

as a book, Logique Algorithmique. Essai sur un système de signes appliqué à la logique. J. Desoer & C. Muquardt, Liège & Bruxelles, 1877
Delpech LJ (1972) La Cybernétique et ses Théoriciens. Casterman / Poche, collection Mutations. Orientations
Descartes R (1637) Discours de la Méthode pour bien conduire sa raison, et chercher la vérité dans les sciences. English version: A discourse on the method of correctly conducting one's reason and of seeking truth in the sciences, translated by I. Maclean, Oxford World's Classics, 2008
Dinneen GP (1955) Programming pattern recognition. In: Proceedings of the Western joint computer conference, 1–3 March 1955, Institute of Radio Engineers, New York, pp 94–100
Dipert RR (1994) The life and logical contributions of O.H. Mitchell, Peirce's gifted student. Trans Charles S Peirce Soc 30(3):515–542
Dodgson CL (2001) The political pamphlets and letters of Charles Lutwidge Dodgson and related pieces: a mathematical approach. In: Abeles FF (ed) The pamphlets of Lewis Carroll, vol 3. Lewis Carroll Society of North America, New York
Du Crest, comtesse de Genlis SF (1797) Alphonse et Dalinde, ou La féérie de l'Art et de la Nature: conte moral. Berthevin, Orléans
Dubarle D (1948) Vers la machine à gouverner ? Le Monde, 28 décembre See also Existe-t-il des machines à penser?, Revue des Questions Scientifiques, Ve série, t. XI, 210–230, 1950, and Scientific Humanism and Christian Thought, Blackfriars, London, 1956
Dubois D, Prade H (2012) Abe Mamdani: a pioneer of soft artificial intelligence. In: Trillas E, Bonissone PP, Magdalena L, Kacprzyk J (eds) Combining experimentation and theory - a hommage to Abe Mamdani. Springer, Berlin, pp 49–60
Dubucs J, Sandu G (eds) (2005) Les Chemins de la Logique. Pour la Science, dossier n^o 49
Duda RO, Hart PE, Nilsson NJ (1976) Subjective Bayesian methods for rule-based inference systems. In: Proceedings of the national computer conference (AFIPS Conference Proceedings, vol 45), pp 1075–1082, sRI Technical Note 124
Duda RO, Gaschnig J, Hart PE (1981) Model design in the PROSPECTOR consultant system for mineral exploration. In: Michie D (ed) Expert systems in the micro-electronic age. Edinburgh University Press, Edinburgh, pp 153–167
Dumarsais CC (1730) Traité des Tropes. Reprinted, Fayard, Paris, 1992
Dupleix S (1603) La Logique ou Art de Discourir et de Raisonner. Edition de 1607, Fayard, Paris, 1984, 370 p
Dutilh Novaes C, Read S (eds) (2016) The Cambridge companion to medieval logic. Cambridge University Press, Cambridge
Elcock EW, Michie D (eds) (1977) Machine intelligence 8. Ellis Horwood Ltd. and Wiley, New York
Erman LD, Hayes-Roth F, Lesser VR, Reddy DR (1980) The Hearsay-II speech-understanding system: integrating knowledge to resolve uncertainty. Comput Surv 12(2):213–253
Euler L (1761, publ. 1768) Lettres cii-cviii. In: Lettres à une Princesse d'Allemagne sur Divers Sujets de Physique & de Philosophie, vol 2; English version: Letters of Euler, on different subjects in natural philosophy, addressed to a German princess 2. J. & J. Harper, 1833
Evans TG (1964) A heuristic program to solve geometry-analogy problems. In: Proceedings of the A.F.I.P. spring joint computer Conference, vol 25, pp 5–16
Faris JA (1955) The Gergonne relations. J Symb Log 20(3):207–231
Feigenbaum EA, Feldman J (eds) (1963) Computers and Thought. McGraw-Hill, New York, articles by P. Armer, C. Chomsky, G.P.E. Clarkson, E.A. Feigenbaum, J. Feldman, H. Gelernter, B.F. Green Jr, J.T. Gullahorn, J.E. Gullahorn, J.R. Hansen, C.I. Hovland, E.B. Hunt, K. Laughery, R.K. Lindsay, D.W. Loveland, M. Minsky. U. Neisser, A. Newell, A.L. Samuel, O.G. Selfridge, J.C. Shaw, H.A. Simon, J.R. Slagle, F.M. Tonge, A.M. Turing, L. Uhr, C. Vossler, A.K. Wolf
Fidora A, Sierra C (eds) (2011) Ramon Llull: from the Ars Magna to artificial intelligence. Artificial Intelligence Research Institute, IIIA, CSIC, 146 p, Barcelona, Contributions by S. Barberà, M. Beuchot, E. Bonet, A. Bonner, J.M. Colomer, J.N. Crossley, A. Fidora, T. Sales, G. Wyllie

Fikes RE, Nilsson NJ (1971) STRIPS: a new approach to the application of theorem proving. Artif Intell 2:189–208
Fodor JA (1978) Tom swift and his procedural grandmother. Cognition 6:229–247
Gabbay DM, Woods J (eds) (2004a) Greek, Indian and Arabic logic. Handbook of history of logic, vol 1. Elsevier, Amsterdam
Gabbay DM, Woods J (eds) (2004b) The rise of modern logic: from Leibniz to Frege. Handbook of History of Logic, vol 3. Elsevier, Amsterdam
Gabbay DM, Woods J (eds) (2008a) British logic in the nineteenth century. Handbook of history of logic, vol 4. Elsevier, Amsterdam
Gabbay DM, Woods J (eds) (2008b) Mediaeval and renaissance logic. Handbook of history of logic, vol 2. Elsevier, Amsterdam
Galli de Bibiena J (1747) La Poupée. Desjonquères, 1987; English translate The fairy doll. Chapman and Hall, 1925 and Amorous Philandre (Store Window Doll Lives). Avon Book, 1948
Geach P, Black M (eds) (1980) Translations from the philosophical writings of Gottlob Frege, 3rd edn. Basil Blackwell (1st edn. 1952)
Gelernter H (1959) Realization of a geometry theorem proving machine. In: Proceedings of the international conference on information processing, Paris, pp 273–282
Gentzen G (1969) The collected papers of Gerhard Gentzen. Studies in logic and the foundations of mathematics. North-Holland Publising Company, Amsterdam
Georgeff M (1983) Communication and interaction in multi-agent planning. In: Genesereth MR (ed) Proceedings of the national conference on artificial intelligence, Washington, D.C., 22–26 August 1983. AAAI Press, pp 125–129
Gergonne JD (1815) Application de la méthode des moindres quarrés à l'interpolation des suites. Annales de Mathématiques Pures et Appliquées 6:242–252
Gergonne JD (1816a) Théorie de la règle de trois. Annales de Mathématiques Pures et Appliquées 7:117–122
Gergonne JD (1816b) Variétés. Essai de dialectique rationnelle. Annales de Mathématiques Pures et Appliquées 7:189–228
Giard L (1972) La "dialectique rationnelle" de Gergonne. Revue d'Histoire des Sciences 25(2):97–124
Gillon BS (ed) (2010) Logic in earliest classical India. Motilal Banarsidass Publishers Private Limited, Delhi, papers of the 12th World sanskrit conference, vol 10.2
Giralt G, Sobek RP, Chatila R (1979) A multi-level planning and navigation system for a mobile robot: a first approach to HILARE. In: Buchanan BG (ed) Proceedings of the 6th international joint conference on artificial intelligence (IJCAI'79), Tokyo, 20–23 August 1979. William Kaufmann, pp 335–337
Gochet P, Grégoire E, Gribomont P, Hulin G, Pirotte A, Roelants D, Snyers D, Thayse A, MVauclair, Wolper P (1988 and 1989) From standard logic to logic programming: introducing a logic based approach to artificial intelligence and From modal logic to deductive databases: introducing a logic based approach to artificial intelligence. Wiley, New York
Gödel K (1995) Collected works Vol. 3. Unpublished Essays and lectures. Oxford University Press, Oxford, edited by S. Feferman, W. Goldfarb, J.W. Dawson Jr, Ch. Parsons, R.M. Solovay
Gombocz WL (1990) Apuleius is better still: a correction to the square of opposition. Mnemosyne XLIII(1–2):124–131
Good IJ (1961) A causal calculus I. Br J Philos Sci 11:305–318
Good IJ (1962a) A causal calculus II. Br J Philos Sci 12:43–51
Good IJ (1962b) Subjective probability as the measure of a non-measurable set. In: Nagel E, Suppes P, Tarski A (eds) Logic, methodology, and philosophy of science. Stanford University Press, Stanford, pp 319–329
Good IJ (1965) Speculations concerning the first ultra intelligent machine. In: Alt FL, Rubinoff M (eds) Advances in computers, vol 6. Academic, London, pp 31–88

Green CC (1979) Theorem proving by resolution as a basis for question answering systems. In: Meltzer B, Michie D (eds) Machine intelligence, vol 4. Edinburgh University Press, Edinburgh, pp 183–205
Grice P (1957) Meaning. Philos Rev 66:377–388
Grimson WEL, Patil RS (eds) (1987) AI in the 1980s and beyond. A MIT survey. MIT Press, Cambridge
Grize JB (1982) De la Logique à l'Argumentation. Librairie Droz, Genève
Günther G (1957) Das Bewusstsein der Maschinen. Eine Metaphysik der Kybernetik, 1. Aufl. 1957, 2. Aufl. 1963, 3. Aufl. 2002. Agis Verlag, Krefeld, Baden, French translate La Conscience des Machines - Une Métaphysique de la Cybernétique followed by Cognition et Volition, 2008
Guo Z (2017) Pensée Chinoise et Raison Grecque. Editions Universitaires de Dijon
Haavelmo T (1943) The statistical implications of a system of simultaneous equations. Econometrica 11(1–2):1–12, reprinted in D.F. Hendry and M.S. Morgan (eds) The foundations of econometric analysis. Cambridge University Press, Cambridge, pp 477–490, 1995
Hamilton W (1859–1860) Lectures on metaphysics and logic, vol 4. William Blackwood and Sons, Edinburgh
Hansen C (Chen Han Sheng) (1983) Language and logic in ancient China. University of Michigan Press, Michigan Studies on China, Ann Arbor
Harrison H, Minsky M (1992) The turing option. Viking
Harsanyi JC (1967) Games with incomplete information played by "Bayesian" players, I–III. Part I. The basic model. Manag Sci (Theory Series) 14(3):159–182
Hart PE, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cybern 4(2):100–107
Hayes JE, Michie D, Mikulich LJ (eds) (1979) Machine intelligence 9. Ellis Horwood Ltd. and Wiley, New York
Hebb DO (1949) The organization of behaviour. Wiley, New York
Hempel CG (1965) Studies in the logic of confirmation. Aspects of scientific explanation and other essays in the philosophy of science, pp 3–46
Hendler J (2008) Avoiding another AI winter. IEEE Intell Syst 23(2):2–4
Herbrand J (1968) Écrits Logiques. Presses Universitaires de France, Logical writings edited by W.D. Goldfarb, translation of the Écrits logiques edited by J. van Heijenoort, Harvard University Press, Cambridge, 1971
Hewitt C (1969) PLANNER: a language for proving theorems in robots. In: Walker DE, Norton LM (eds) Proceedings of the 1st international joint conference on artificial intelligence, Washington, DC, May 1969, pp 295–302
Hewitt C (2009) Middle history of logic programming: resolution, planner. Edinburgh LCF, Prolog, Simula, and the Japanese fifth generation project. CoRR arXiv:0904.3036v25
Heyting A (1956) Intuitionism: an introduction. North-Holland Publishing Co., Amsterdam
Hintikka J (1962) Knowledge and belief: an introduction to the logic of the two notions. Cornell University Press, Ithaca
Hobbes of Malmesbury T (1651) Leviathan, or the matter, forme and power of a common-wealth ecclesiasticall and civil. Basil Blackwell, Oxford, 1955
Hobbes of Malmesbury T (1655) Elementa Philosophiae I. De Corpore. Vrin, Paris, Bibliothèque des Textes Philosophiques, 2000, English translate Elements of philosophy, the first section, concerning body, 1656; The English works of Thomas Hobbes of Malmesbury edited by W. Molesworth, Vol. 1: elements of philosophy, parts I–IV, The first section concerning body, John Bohn, London 1839; the quoted text is in Part First, Computation or Logic, chap. 1 of Philosophy
Hodges W (2018) Two early Arabic applications of model-theoretic consequence. Log Univ 12(1–2):37–54
Homer (1984) The iliad. Oxford Paperbacks
Hubien H (1977) Logiciens médiévaux et logique d'aujourd'hui. Revue Philosophique de Louvain 75(26):219–233

Hughes G (1982) John Buridan on self-reference: chapter eight of Buridan's Sophismata. An edition and translation with an introduction, and philosophical commentary. Cambridge University Press, London

Hume D (1748) An enquiry concerning human understanding. Oxford world's classics, 2008

Hünefeldt T, Brunetti R (2004) Artificial intelligence as "theoretical psychology": Christopher Longuet–Higgins' contribution to cognitive science. Cogn Process 5(3):137–139

Jaynes ET (1957) Information theory and statistical mechanics I & II. Phys Rev 106(4):620–630 & 108(2):171–190

Jeavons WS (1869) The substitution of similars, the true principle of reasoning, Derived from a Modification of Aristotle's Dictum. Macmillan & Co

Jeavons WS (1870) Elementary lessons in logic: deductive and inductive, with copious questions and examples, and a vocabulary of logical terms. Macmillan & Co, reprinted by Elibron Classics

Johnson-Laird PN (1978) What's wrong with Grandma's guide to procedural semantics: a reply to Jerry Fodor. Cognition pp 249–261

Jones RB (ed) (2012) The Organon: the works of Aristotle on logic. CreateSpace independent publishing platform, includes six works on logic: the categories; On interpretation; The prior analytics; The posterior analytics; The topics; The sophistical refutations

Kalinowski G (1982) La logique juridique et son histoire. Archives de Philosophie du Droit 27:275–289, republ. in Anuario Filosófico, vol 16, pp 331–350, 1983

Keynes JM (1921) A treatise on probability. Macmillan & Co., London

Keynes JN (1900) Studies and exercises in formal logic, including a generalization of logical processes in their application to complex inferences. Macmillan & Co., London

Kleene SC (1952) Introduction to metamathematics. North Holland, Amsterdam

Kleene SC (1956) Representation of events in nerve nets. In: Shannon CE, McCarthy J (eds) Automata studies. Princeton University Press, Princeton, pp 3–40. 1st version: Representation of events in nerve nets and finite automata, U.S. Air Force, Project RAND, Research Memorandum 704, 98 p, 15 December 1951

Kripke S (1959) A completeness theorem in modal logic. J Symb Log 24(1):1–14

Kripke S (1963) Semantical considerations on modal logic. Acta Philos Fenn 16:83–94

Kullback S, Leibler R (1951) On information and sufficiency. Ann Math Stat 22:79–86

Kurzweil R (1990) The age of intelligent machines. MIT, Cambridge

La Mettrie (Offray de) J (1747) Man a machine and man a plant. Hackett Publishing Company, 1994, also author of L'Homme plus que Machine (1748) (republished by Rivages, Payot, 2004) and of Les Animaux plus que Machines (1750)

Ladd C (1883) On the algebra of logic. In: Peirce CS (ed) Studies in logic by members of the Johns Hopkins University, Little, Brown, and Company, Baltimore, pp 17–71

Lambert JH (1764) Neues Organon oder Gedanken über die Erforschung und Bezeichnung des Wahren und dessen Unterscheidung vom Irrthum und Schein. Reprinted in Philosophische Schriften. Volume II, Georg Olms Verlagsbuchhandlung, Hildesheim, 1965; and by Akademie Verlag Berlin, 1990

Lambert of Auxerre (2015) Logica, or Summa Lamberti. University of Notre Dame Press, translated, with introduction and notes, by Th. S. Maloney

Langius IC (1714) Inventvm Novvm Quadrati Logici Vniversalis [...]. Gissa Hassorum: Henningius Mllerus

Lapicque L (1943) La Machine Nerveuse. Flammarion, Paris

Laplace PS (1814) Essai Philosophique sur les Probabilités. Madame Veuve Courcier, Paris, republ. by Christian Bourgois, 1986; English translate A Philosophical Essay on Probabilities, Dover Publications, New York, 1951

Largeault J (1972) Logique Mathématique. Textes. Collection U, Armand Colin, Paris, texts by J. Lukasiewicz, E. Post, E.W. Beth, Th. Skolem, L. Löwenheim, K. Gödel, L. Henkin, D. Hilbert

Leibniz GW (1703) Explication de l'arithmétique binaire, qui se sert des seuls caractères 0 & 1; avec des Remarques sur son utilité, & sur ce qu'elle donne le sens des anciennes figures Chinoises de Fohy. Compte Rendu de l'Académie des Sciences (Paris), Mémoires pp 85–89,

English translation by L. Strickland available online: Explanation of binary arithmetic, which uses only the characters 0 and 1, with some remarks on its usefulness, and on the light it throws on the ancient Chinese figures of Fuxi
Lemanski J (2017) Periods in the use of Euler-type diagrams. Acta Baltica Historiae et Philosophiae Scientiarum 5(1):50–69
Lemanski J (2018) Logic diagrams in the Weigel and Weise circles. Hist Philos Log 39(1):3–28
Lenzen W (2016) Leibniz's logic and the "cube of opposition". Log Univ 10(2):171–189
Leśniewski S (1992) Collected works. In: Surma SJ, Srzednicki JT, Barnett DI, Rickey VF (eds) Nijhoff international philosophy series, vol I, II. Kluwer, Dordrecht
Lettvin JY, Maturana HR, McCulloch WS, Pitts WH (1959) What the frog's eye tells the frog's brain. Proc IRE 47(11):1940–1951
Lewis D (1973) Counterfactuals and comparative possibility. J Philos Log 2(4):418–446
Lighthill J (1973) Artificial intelligence: a paper symposium. Science Research Council, UK, contents: Part I: Artificial intelligence: a general survey (Sir James Lighthill) Part II: Some comments on the Lighthill report and on artificial intelligence (N.S. Sutherland), Part III: Comments on the Lighthill report and the Sutherland reply, par R.M. Needham, H.C. Longuet-Higgins, et par D. Michie
Lindsay RK (1963) Inferential memory as the basis of machines which understand natural language. In: Feigenbaum EA, Feldman J (eds) Computers and thought. McGraw-Hill, New York, pp 217–233
Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J (1980) Applications of artificial intelligence for organic chemistry: the DENDRAL project. McGraw-Hill, New York
Link D (2012) Programming ENTER: Christopher Strachey's draughts program. Computer resurrection. Bull Comput Conserv Soc 60(3):23–31
Locke J (1690) An essay concerning human understanding. Penguin Classics, 1998
Londey D, Johanson C (1984) Apuleius and the square of opposition. Phronesis 29(2):165–173. Correction, Phronesis 30:209 (1985)
Londey D, Johanson C (1987) The Logic of Apuleius, including a complete Latin text and English translation of the Peri Hermeneias of Apuleius of Madaura. E.J. Brill, Leiden & New York, Philosophia Antiqua series, v. 47
Łukasiewicz J (1913) Die Logischen Grundlagen der Wahrscheinlichkeitsrechnung. In: Borkowski L (ed) Jan Łukasiewicz - Selected works. North-Holland, Amsterdam. Polish Scientific Publishers, Warsaw, 1970. Logical foundations of probability theory, English translation, pp 16–63
Łukasiewicz J (1930) Philosophical remarks on many-valued systems of propositional logic. In: Borkowski L (ed) Jan Łukasiewicz - Selected works, North-Holland, Amsterdam. Polish Scientific Publishers, Warsaw, 1970, pp 153–179
Mamdani EH, Assilian S (1975) An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man-Mach Stud 7(1):1–13
Marinetti FT (1909) Poupées Électriques, drame en trois actes, avec une Préface sur le Futurisme. E. Sansot & Cie, Paris
Mariotte E (1678) Essai de Logique, contenant les principes des sciences, et la manière de s'en servir pour faire de bons raisonnements. Reprinted by Fayard, Paris, 1992, followed by Les principes du devoir et des connaissances humaines, attributed to Roberval
Marquis P, Papini O, Prade H (2014) Some elements for a prehistory of artificial intelligence in the last four centuries. In: Schaub T, Friedrich G, O'Sullivan B (eds) Proceedigns of the 21st European conference on artificial intelligence (ECAI'14), 18–22 August, Prague. Frontiers in artificial intelligence and applications, vol 263. IOS Press, Amsterdam, pp 609–614
Marr D (1982) Vision. W.H. Freeman and Co., San Francisco, réd. MIT Press, Cambridge, 2010. Foreword: S. Ullman; Afterword: T. Poggio
Martin T (2006a) La logique probabiliste de Gabriel Cramer. Math Sci Hum 44^{e} année(4):43–60
Martin T (2006b) Logique du probable de Jacques Bernoulli à J.-H. Lambert. Journ@l électronique d'Histoire des Probabilités et des Statistiques 2(1b)
Martin T (2011) J.-H. Lambert's theory of probable syllogisms. Int J Approx Reason 52:144–152

Mavrodes GI (1963) Some puzzles concerning omnipotence. Philos Rev 72:221–223
McCarthy J (1990) In: Lifschitz V (ed) Formalizing common sense: papers by John McCarthy. Intellect Books
McCarthy J (1996) Defending AI research: a collection of essays and reviews. CSLI Publications, Stanford
McCarthy J, Hayes P (1979) Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer B, Michie D (eds) Machine intelligence, vol 4. Edinburgh University Press, Edinburgh, pp 463–502
McCarthy J, Abrahams PW, Edwards DJ, Hart TP, Levin MI (1962) LISP 1.5 programmer's manual, 2nd edn. MIT Press, Cambridge, 1985. The Computation Center And Research Laboratory of Electronics
McCarthy J, Minsky M, Rochester N, Shannon CE (2006) A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Mag 27(4):12–14
McCorduck P (1979) Machines who think. A personal inquiry into the history and prospects of artificial intelligence. W.H. Freeman and Company, San Francisco
McCulloch WS, Pitts W (1943) A logical calculus of ideas immanent in nervous activity. Bull Math Biophys 5:115–133
Meltzer B, Michie D (eds) (1968–1972) Machine intelligence 3–7. American Elsevier Publishing Company, New York
Menger K (1942) Statistical metrics. Proc Nat Acad Sci USA 28:535–537
Mérimée P (1837) La Vénus d'Ille. Emile Colin et Cie, English translated by Venus of Ille and other stories. Oxford library of French classics. Oxford University Press, Oxford, 1966
Meusnier N, Piron S (2007) Medieval probabilities: a reappraisal. In: Journ@l électronique d'Histoire des Probabilités et des Statistiques 3(1)
Meyrink G (1915) Der Golem. English translated by The Golem. European Classics Paperback, Dedalus Ltd., 1985
Michie D (1963) Experiments on the mechanization of game-learning. Part I. Characterization of the model and its parameters. Comput J 6(3):232–236
Minsky M (1954) Theory of neural-analog reinforcement systems and its application to the brain-model problem. Typewritten ms., dated 1953; PhD thesis, Princeton University
Minsky M (1975) Minsky's frame system theory. In: Proceedings of the 1975 workshop on theoretical issues in natural language processing (TINLAP '75), Association for computational linguistics, pp 104–116, the article originally appeared without any author name
Minsky M (1986) The society of mind. Simon & Schuster, Inc
Minsky M, Papert S (1969) Perceptrons: an introduction to computational geometry, 2nd edn. The MIT Press, Cambridge, revised 1972
Minsky M, Selfridge OG (1961) Learning in random nets. In: Proceedings of the 4th London symposium on information theory, Butterworth Ltd., London, pp 335–347
Mitchell OH (1883) On a new algebra of logic. In: Peirce CS (ed) Studies in logic by members of the Johns Hopkins University, Little, Brown and Company, Baltimore, pp 72–106
Moisil G (1972) La logique des concepts nuancés. In: Essais sur les Logiques Non Chrysippiennes, Editions Acad. Repub. Soc. Roum., Bucharest, pp 157–163
Moore J, Newell A (1974) How can Merlin understand? In: Gregg L (ed) Knowledge and cognition. Erlbaum, Hillsdale, pp 201–252
Moravec HP (1982) The CMU rover. In: Waltz DL (ed) Proceedings of the national conference on artificial intelligence. Pittsburgh, 18–20 August. AAAI Press, pp 377–380
More T (1962) Relations between implicational calculi. Technical report. MIT, Cambridge. PhD Dissertation, May
Moretti A (2014) Was Lewis Carroll an amazing oppositional geometer? Hist Philos Log 35(4):383–409
Nagel E, Newman JR (1958) Gödel's proof. New York University Press, New York
Nash J (1951) Non-cooperative games. Ann Math (2nd Series) 54:286–295
Nevatia R, Binford TO (1977) Description and recognition of curved objects. Artif Intell 8(1):77–98

Newell A (1981) The heuristic of George Polya and its relation to artificial intelligence. Technical report, Computer Science Department, Carnegie Mellon University, Paper 2413
Newell A, Simon HA (1956) The logic theory machine. A complex information processing system. The Rand Corporation, Santa Monica, CA, report P-868, 15 June 1956; Proceedings of the IRE Transactions on Information Theory (IT-2), September 1956, pp 61–79
Newell A, Simon HA (1972) Human problem solving, 1st edn. Prentice-Hall, Englewood Cliffs. 1st print. 920 pp; 2nd print. 784 pp
Newell A, Shaw JC, Simon HA (1957) Empirical explorations of the logic theory machine. A case study in heuristic. In: Proceedings of the Western joint computer conference, pp 218–239
Newell A, Shaw JC, Simon HA (1959) Report on a general problem-solving program. In: Proceedings of the international conference on information processing, pp 256–264
Nilsson NJ (2010) The quest for artificial intelligence: a history of ideas and achievements. University Press, Cambridge
Ovid (1998) Metamorphoses. Oxford Paperbacks
Papert S (1980) Mindstorms: children, computers, and powerful ideas. Prentice Hall/Harvester
Pareto V (1961) On logical and non-logical action. In: Parsons T, Shils E, Naegele KD, Pitts JR (eds) Theories of society. Foundations of modern sociological theory, vol II. The Free Press of Glencoe, Inc., pp 1061–1063
Parsons T (2017) The traditional square of opposition. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, Metaphysics Research Lab, Stanford University
Peirce CS (1870) Description of a notation for the logic of relatives, resulting from an amplification of the conceptions of Boole's calculus of logic. Memoirs of the American academy of arts and sciences 9:317–378, reprinted in Collected Papers, vol 3, pp 45–149
Peirce CS (1880) On the algebra of logic. Am J Math 3:15–57, reprinted in Collected Papers, vol 3, pp 154–251, 1960
Peirce CS (1885) On the algebra of logic: a contribution to the philosophy of notation. Am J Math 7(2):180–202, reprinted in Collected Papers, vol 3, pp 359–403, 1960
Peirce CS (1931) Collected papers of Charles Sanders Peirce. Harvard University Press, Cambridge. Publication 1931–1935, 1958
Peirce CS (1955) Philosophical writings. Selected and edited, with an introduction by J. Buchler, Dover Publications
Peter of Spain (2014) Summaries of logic [Tractatus]. Oxford University Press, New York, translated with introduction and notes by B.P. Copenhaver, C.G. Normore and T. Parsons
Petronius (1969) The Satyricon & the fragments. Penguin Books
Piaget J (1949) Traité de Logique. Essai de Logistique Opératoire. Armand Colin, Paris, 2nd revised ed.: Essai de logique opératoire, in collaboration with Jean-Blaise Grize, Dunod, Paris, 1972
Pitrat J (1970) Un programme de démonstration de théorèmes. Dunod, Paris
Pitrat J (1977) A chess combination program which uses plans. Artif Intell 8(3):275–321
Pitrat J (2000) Métaconnaissance : Futur de l'Intelligence Artificielle. Hermes Science Publications
Ploucquet G (2006) Logik. Georg Olms, Hildesheim, herausgegeben, übersetzt und mit einer Einleitung versehen von M. Franz
Poe EA (1836) Maelzel's chess player. Southern Literary Messenger, Richmond
Polya G (1945) How to solve it, 2nd edn. Princeton University Press, Princeton, 1957
Polya G (1954) Mathematics and plausible reasoning, 2nd edn. Vol. 1: Induction and analogy in mathematics. Vol. 2: Patterns of plausible inference. Princeton University Press, Princeton, 1968
Pratt V (1987) Thinking machines: the evolution of artificial intelligence. Basil Blackwell Ltd., Oxford and New York
Pratt-Hartmann I (2011) The Hamiltonian syllogistic. J Log Lang Inf 20(4):445–474
Quine WVO (1941) Elementary logic, 2nd edn. Harper & Row, New York, 1965
Rahman S, Redmond J (2007) Hugh MacColl. An overview of his logical work with anthology. College Publications
Ramsey FP (1931) Foundations - essays in philosophy, logic, mathematics and economics. Ed. by D.H. Mellor, republ. by Humanities Press, 1978

Raphael B (1976) The thinking computer: mind inside matter. W.H. Freeman and Co., San Francisco
Ratliff TC (2010) Lewis Carroll, voting, and the taxicab metric. Coll Math J 41:303–311
Read S (2010) Thomas Bradwardine, "Insolubilia". Dallas Medieval Texts and Translations, 10, Peeters Editions, Leuven, Latin text and English translation
Read S (2012) John Buridan's theory of consequence and his octagons of opposition. In: Béziau J, Jacquette D (eds) Around and beyond the square of opposition. Studies in universal logic. Springer, Basel, pp 93–110
Rescher N (1963) Studies in the history of Arabic logic. University of Pittsburgh Press, Pittsburgh
Rescher N (1964) The development of Arabic logic. University of Pittsburgh Press, Pittsburgh
Rescher N (1967) Temporal modalities in Arabic logic. Foundations of language, supplementary series. D. Reidel Publishing Company, Dordrecht
Rescher N (1976) Plausible reasoning. Van Gorcum, Amsterdam
Robida A (1883) Le Vingtième Siècle. Georges Decaux, Paris, illustrated by the author; English transl The twentieth century. Wesleyan University Press, Early classics of science fiction, 2004
Robinson JA (1965) A machine-oriented logic based on the resolution principle. J ACM 12(1):23–41
Rochester N, Holland JH, Haibt LH, Duda WL (1956) Tests on a cell assembly theory of the action of the brain using a large digital computer. IRE Trans Inf Theory IT-2:80–93
Rose F (1984) Into the heart of the mind: an American quest for artificial intelligence. Harper & Row
Rosenblatt F (1962) Principles of neurodynamics: perceptrons and the theory of brain mechanisms. Spartan Books
Rosenblueth A, Wiener N, Bigelow J (1943) Behavior, purpose and teleology. Philos Sci 10(1):18–24
Roussel R (1914) Locus Solus. Alphonse Lemerre, English translation Locus Solus by R.C. Cunningham, New Directions Publishing Corporation, 2017
Russell B (1956) Logic and knowledge. Essays 1901–1950. George Allen & Unwin Ltd., London & Macmillan, New York, edited by R.C. Marsh
Sacerdoti ED (1977) Structure for plans and behaviour. Elsevier, Amsterdam
Samuel A (1959) Some studies in machine learning using the game of checkers. IBM J 3(3):210–229, some studies in machine learning using the game of checkers. II. Recent progress. IBM J 11(6):601–617, 1967
Sarukkai S (ed) (2018) Handbook of logical thought in India. Springer, Berlin. Contents: Buddhist logic: sample texts (P.P. Gokhale); Convergence and divergence of Nyāya and Tattvavāda (Dvaita) theories of logic (V. Nishanka); Dependency of inference on perception and verbal testimony (P Vinay); Early Nyāya logic: rhetorical aspects (K. Lloyd); General introduction to Buddhist logic (J. Tuske); General introduction to logic in Jainism with a list of logicians and their texts (J. Soni); Introduction to Buddhist logicians and their texts (M. Chattopadhyay); Later Nyāya logic: computational aspects (A. Kulkarni); Logic in Tamil didactic literature (T. Jayaraman); Logic of Syād-Vāda (A. Clavel); Logical argument in Vidyānandins Satya-śāsana-parīkṣā (H. Trikha) Some issues in Buddhist logic (P.P. Gokhale, K. Bhattacharya) The logic of late Nyāya: a property-theoretic framework for a formal reconstruction (E. Guhe); The logic of late Nyāya: problems and issues (E. Guhe); The opponent: jain logicians reacting to Dharmakīrtis theory of inference (M.-H. Gorisse)
Savage LJ (1954) The foundations of statistics. Wiley, New York. 2nd revised edition, 1972
Schank R (1973) Identification of conceptualizations underlying natural language. In: Schank R, Colby K (eds) Computer models of thought and language. W.H. Freeman and Co., San Francisco, pp 187–247
Schank R, Abelson RP (1977) Scripts, plans, goals and understanding: an inquiry into human knowledge structures. Erlbaum
Schröder E (1890) Vorlesungen über die Algebra der Logik, 3 vols. B.G. Teubner, Leipzig. Publication 1890–1905, republ. by Chelsea, 1966; Thoemmes Press, 2000
Schumann A (2012) Studies in Talmudic logic. Vol 6: Talmudic logic. College Publications

Schumann A (ed) (2017) Studies in Talmudic logic. Vol. 14: Philosophy and history of Talmudic logic. College Publications
Selfridge O (1955) Pattern recognition and modern computers. In: Proceedings of the Western joint computer conference, 1–3 March 1955. Institute of Radio Engineers, New York, pp 91–93
Selfridge OG (1959) Pandemonium: a paradigm for learning. In: Blake DV, Uttley AM (eds) Symposium on mechanisation of thought processes, London, 24–27 November 1958, pp 511–529
Sesmat A (1951) Logique. I: Les Définitions. Les Jugements. Logique II: Les Raisonnements, la Logistique. Hermann, 2 vols. 1950–1951, Paris
Shackle GLS (1949) Expectation in economics. Cambridge University Press, Cambridge
Shackle GLS (1961) Decision, order and time in human affairs, 2nd edn. Cambridge University Press, Cambridge
Shafer G (1978) Non-additive probabilities in the work of Bernoulli and Lambert. Arch Hist Exact Sci 19(4):309–370
Shafer G (2018) Marie-France Bru and Bernard Bru on dice games and contracts. Statistical Science To appear
Shannon CE (1938) A symbolic analysis of relay and switching circuits. Trans AIEE 57(12):713–723, the master thesis of the author, with the same title, is from 1937
Shannon CE (1950) Programming a computer for playing chess. Philos Mag (7th Series) XLI(314):256–275, presented at the National Institute of Radio Engineers Convention, 9 March 1949, New York
Shannon CE (1956) A chess-playing machine. In: Newman JR (ed) The world of mathematics - a small library of the literature of mathematics from A'H-Mose the scribe to Albert Einstein (4 Vols), vol 4. Simon & Schuster, New York, pp 2124–2135. In: Part XIX: Mathematical machines: can a machine think?
Shelley MW (1818) Frankenstein: or the modern prometheus. Oxford Paperbacks, 1980
Sherwood W (1966) William of Sherwood's introduction to logic. University of Minnesota Press, Minneapolis, with translation, introduction and notes by N. Kretzmann
Shin SJ, Lemon O (2008) Diagrams. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Winter 2008 Edition)
Slagle JR (1963) A heuristic program that solves symbolic integration problems in freshman calculus. In: Feigenbaum EA, Feldman J (eds) Computers and thought, McGraw-Hill, New York, pp 191–203
Smalbrugge MA (1986) L'argumentation probabiliste d'Augustin dans le *Contra Academicos*. Revue des Études Augustiniennes XXXII:41–55
Smith RG (1984) On the development of commercial expert systems. AI Mag Fall:61–73
Sobel JH (2004) Logic and theism: arguments for and against beliefs in god. Cambridge University Press, Cambridge
Solomonoff RJ (1956) An inductive inference machine. Technical report, Technical Research Group, New York City, 61 p. http://world.std.com/~rjs/indinf56.pdf. Revised version published later as A formal theory of inductive inference, Information and Control, vol 7, pp 1–22 and 224–254, 1964
Sousa Silvestre R (2015) On the logical formalization of Anselm's ontological argument. Revista Brasileira de Filosofia da Religiã o 2(2):142–161
Spade PV (1979) Roger Swyneshed's Insolubilia: Edition and Comments. Archives d'Histoire Doctrinale et Littéraire du Moyen Âge 46:177–220
Spade PV, Read S (2018) Insolubles. In: Zalta E (ed) The Stanford encyclopedia of philosophy, Metaphysics Research Lab, Stanford University
Strachey CS (1952) Logical or non-mathematical programmes. In: Proceedings of the 1952 ACM national meeting, Toronto, 8–10 September 1952, pp 46–49
Stuart Mill J (1843) A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence and the methods of scientific investigation. John W. Parker, London. 2 vols.: XVI + 580, XII + 624 pp; republished by University of Toronto Press, Routledge & Kegan Paul, 1974

Stuart Mill J (1863) Utilitarianism. Parker, Son, and Bourn, London
Suszko R (1968) Non-fregean logic and theories. Acta Log (Annals of the University of Bucarest) 11:105–125
Swift J (1726) Gulliver's travels. Worlds Classics, Oxford Paperbacks, 1987
Takemura R (2013) Proof theory for reasoning with Euler diagrams. A logic translation and normalization. Stud Log 101(1):157–191
Tarski A (1956) Logic, semantics, metamathematics. Papers from 1923 to 1938. Oxford University Press, Oxford, translated by J.H. Woodger
Teixidor J (2003) Aristote en Syriaque. Paul le Perse, logicien du VIe siècle. CNRS Editions, Paris
Thomas of Britain (1969) Tristan. Penguin Books, includes Tristan by Gottfried von Strassburg, with the surviving fragments of the Tristan of Thomas
Thomson W (1842) Outline of the laws of thought. William Pickering, London
Thomson W (1857) An outline of the necessary laws of thought: a treatise on pure and applied logic. Sheldon and Company, New York
Toulmin SE (1958) The uses of argument, 2nd edn. Cambridge University Press, Cambridge, 2003
Triolet E (1963) L'âme. Gallimard, Paris
Turing A (1948) Intelligent machinery. Report National Physical Laboratory, London, 1948. Reprinted in: Machine intelligence, vol 5. Edinburgh University Press, Edinburgh, pp 3–23, 1969
Turing A (1950) Computing machinery and intelligence. Mind 59:433–460
Turing A (1956) Can a machine think? In: Newman JR (ed) The world of mathematics - a small library of the literature of mathematics from A'H-Mose the scribe to Albert Einstein (4 Vols), vol 4. Simon & Schuster, New York, pp 2099–2123. In: Part XIX: Mathematical machines: can a machine think?
Uckelman SL (2017) Medieval logic. In: Malpass A, Marfori MA (eds) The history of philosophical and formal logic: from Aristotle to Tarski. Bloomsbury, London, pp 71–99
Vaucanson J (1738) Le mécanisme du fluteur automate présenté à messieurs de l'Académie Royale des Sciences, avec la description d'un Canard Artificiel, mangeant, beuvant, digerant & se vuidant, épluchant ses aîles & ses plumes, imitant en diverses manières un Canard vivant, et aussi d'une autre figure également merveilleuse, jouant du Tambourin et de la Flute ... (24 p). Paris, English translated An account of the mechanism of an automaton, or image playing on the German-flute: as it was presented in a memoire, to the gentlemen of the royal academy of sciences at Paris. Together with a description of an artificial duck, Gale ECCO, Print Editions, 2010
Venn J (1866) The logic of chance. Macmillan, London and Cambridge, revised, 1888; reprinted by Dover, New York, 2006
Venn J (1880) On the diagrammatic and mechanical representation of propositions and reasonings. Lond Edinb Dublin Philos Mag J Sci 10(58):1–18
Venn J (1881) Symb Log. Macmillan, London
Vigneron H (1914) Les automates. La Nature, Revue des Sciences et de leurs Applications aux Arts et à l'Industrie Quarante deuxième année (2142):56–61, 13 juin
von Neumann J (1956) The general and logical theory of automata. In: Newman JR (ed) The world of mathematics - a small library of the literature of mathematics from A'H-Mose the scribe to Albert Einstein (4 Vols), vol 4. Simon & Schuster, New York, pp 2070–2098. In: Part XIX: Mathematical machines: can a machine think?
von Neumann J (1958) The computer and the brain. Yale University Press, New Haven
von Neumann J (1966) Theory of self-reproducing automata. University of Illinois Press, Urbana, edited and completed by A.W. Burks
von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton
von Sigwart C (1873–1878) Logik. Laupp, Tübingen, English translation by H. Dendy: Logic. Volume 1. The judgment, concept, and inference. Volume 2. Logical methods, Swan Sonnenschein & Co., London, 1895
von Wright G (1951) An essay in modal logic. North-Holland Publishing Co., Amsterdam

Vuillemin J (1971) La Logique et le Monde Sensible. Flammarion, Paris
Wade Savage C (1967) The paradox of the stone. Philos Rev 76(1):74–79
Waltz D (1975) Understanding line drawings of scenes with shadows. In: Winston PH (ed) The psychology of computer vision. McGraw-Hill, New York, pp 19–91
Weizenbaum J (1966) Eliza - a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45
Whately R (1826) Elements of logic, comprising the substance of the article in the encyclopaedia metropolitana. J. Mawman, London
Whitehead AN, Russell B (1910) Principia Mathematica, vol 3. Cambridge University Press, Cambridge. Publication 1910–1913; 2 éd. 1925–1927
Wiener N (1949) Cybernetics or control and communication in the animal and the machine. Hermann/Wiley, Paris/New York
Wiener N (1950) The human use of human beings. Cybernetics and society. Houghton Mifflin, Boston
Wilks Y (1972) Grammar, meaning and the machine analysis of language. Routledge, London
Winograd T (1971) Procedures as a representation for data in a computer program for understanding natural language. MIT AI Technical report 235
Wittgenstein L (1921) Tractatus Logico-Philosophicus. Annalen der Naturphilosophie, introduction by B. Russell
Wittgenstein L (1969) On Certainty. Basil Blackwell; Harper & Row Publisher, 1972
Wolff Ch (1713) Vernünfftige Gedancken von den Kräfften des menschlichen Verstandes und ihrem richtigen Gebrauche in Erkäntniβ der Wahrheit. French translation (by J. Deschamps, Berlin, 1736): Réflexion sur les forces de l'entendement humain et sur leur légitime usage dans la connaissance de la vérité, itself translated in English: Logic, or Rational Thoughts on the Powers of the Human Understanding, with their Use and Application in the Knowledge and Search of Truth (Gale Ecco, Print Editions 2010)
Woods WA (1975) What's in a link: foundations for semantic networks. In: Bobrow D, Collins A (eds) Representation and understanding: studies in cognitive science. Academic, New York, pp 35–82
Wright S (1921) Correlation and causation. J Agric Res 20:557–585
Zadeh LA (1950) Thinking machines. A new field in electrical engineering. Columbia Eng Q 3:12–13, 30–31
Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353
Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1(1):3–28

Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning

Andreas Herzig and Philippe Besnard

Abstract The aim of the present chapter is to overview three important tools for knowledge representation that are strongly interrelated. All three can be traced back to a fundamental limitation of classical logic: its connectives are truth-functional, which does not allow to reason about some concepts such as modalities and "if-then" relationships between propositions. To witness, most of the students in an introductory course on logic have a hard time to accept that the implication "if A then B" should be identified with "A is false or B is true". Indeed, such an identification leads to validities that are rather counter-intuitive, such as "B implies A implies B" or "A implies B, or B implies A". In introductory courses it is often omitted that the above interpretation of the so-called material implication was subject of much concern among scholars in the past. Their work led to the development of several families of formalisms that will be presented in this chapter: modal logics, conditional logics, and nonmonotonic formalisms. The next three sections detail the definitions of each of these: the modal logics K and S5, the conditional logics due to Stalnaker and Lewis, and the preferential and rational nonmonotonic reasoning formalisms. We then study the relationship between conditional logics and dynamic epistemic logics. The latter are a family of modal logics that got popular recently. We show that they can be viewed as particular logics of indicative conditionals: they are in the Stalnaker family and violate all of Lewis's principles.

1 Introduction

The program of the logical approach to AI is to develop methods for the representation of knowledge by means of logical formulas in order to enable the inference of conclusions from these formulas. Scholars first focussed on classical logic. The language of that logic provides logical operators whose most important are negation

A. Herzig (✉) · P. Besnard
IRIT-CNRS, Université Paul Sabatier, Toulouse, France
e-mail: herzig@irit.fr

P. Besnard
e-mail: besnard@irit.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_2

($\neg$), conjunction ($\wedge$), disjunction ($\vee$), material implication ($\rightarrow$) and equivalence ($\leftrightarrow$). These operators are all *truth-functional*. For example, the truth value of the implication $A \rightarrow B$ is a function of the truth value of A and the truth value of B, given that $A \rightarrow B$ is true if and only if A is false or B is true.

However, it became clear rather quickly that truth-functional operators do not allow us to work with some concepts that are important in knowledge representation. Here are three examples of such concepts:

- necessity and possibility;
- an agent's knowledge and belief[1];
- conditionals "if-then".

For example, when A is false there are two possibilities: either A is necessarily false, or A is not necessarily false, and therefore possibly true. The same distinction applies when A is true. This shows that the truth value of "necessarily A" cannot be a function of the truth value of A.

Similarly, when some agent believes that A is true then A can be either true or false.

As to conditionals, one may criticise the truth-functionality of the material implication itself: the formula $\neg A \rightarrow (A \rightarrow B)$ being valid in classical logic, falsehood of A is enough for the implicational link between A and B. This is however not how things work when we use the "if-then" construction in natural language. An example that can be found in several variations in the literature is obtained by reading A as "the moon is made of green cheese" and B as "Earth is flat": it sounds odd to say that A implies B, but this is nevertheless true in classical logic given that A is (believed to be) false. What we would like to have is an operator "$\Rightarrow$" such that falsehood of A is not enough to establish an implicational link between A and B. Formally, we are looking for a semantics where $\neg A \rightarrow (A \Rightarrow B)$ fails to be valid.[2]

The problem of truth-functionality of the classical operators was studied by philosophers well before the beginnings of AI. They proposed extensions of classical propositional logic by non truth-functional concepts:

- modal logics in order to reason about necessity and possibility (Lewis and Langford 1959);
- epistemic logics to reason about knowledge and belief (Hintikka 1962),

[1]We note that knowledge and belief are part of an agent's mental attitudes. Other such attitudes exist and cannot be represented by means of truth-functional operators either. These attitudes are presented in detail in chapter ",Formalization of Cognitive-Agent Systems, Trust, and Emotions" of this volume.

[2]Note that one might as well wish to avoid validity of the nested conditional formula $\neg A \Rightarrow (A \Rightarrow B)$. Such a project does not require a material implication, which amounts to studying $\Rightarrow$ as a 'full-fledged' implication operator that is an alternative to $\rightarrow$. This can e.g. be done in a logical language with operators $\Rightarrow$, $\neg$, $\wedge$ and $\vee$. This leads us to so-called substructural logics such as intuitionistic logic or linear logic (Troelstra 1992). However, most researchers in AI tend a less radical position and study extensions of classical logic by a further logical operator $\Rightarrow$. We consequently restrict our presentation to such approaches.

- conditional logics to reason about implications other than material implication (Stalnaker 1968; Lewis 1979).

Taking advantage of the invention of a simple and intuitive possible worlds semantics by the end of the 1950s by Saul Kripke (1963), the so-called non-classical logics found numerous applications both in philosophy and in AI.

Let us have a closer look at the properties of the material implication $\rightarrow$. Beyond truth-functionality it has other properties that are considered undesirable by many:

1. Monotony: if $A \rightarrow B$ then $(A \wedge A') \rightarrow B$, for every A';
2. Contraposition: if $A \rightarrow B$ then $\neg B \rightarrow \neg A$;
3. Transitivity: if $A \rightarrow B$ and $B \rightarrow C$ then $A \rightarrow C$;
4. Simplification of disjunctive antecedents: if $(A \vee A') \rightarrow B$ then $A \rightarrow B$ and $A' \rightarrow B$.

It was shown that these properties are related (Nute 1980, 1984). A classical counter-example against monotony is the proposition "if I pour sugar in my coffee (A) then I like my coffee (B)" which does not license the conclusion "if I pour sugar in my coffee (A) and I pour diesel in my coffee (A') then I like my coffee (B)".[3]

In AI, scholars focussed on the monotony property, notwithstanding the fact that the four undesirable properties of material implication are related, as we have said above.[4] They aimed at a conditional differing from material implication not only by the absence of truth-functionality, but also by the absence of monotony.

Let us inspect monotony and contraposition in more detail. Each of them can be formulated in two different ways, viz. as axioms:

$$(A \rightarrow B) \rightarrow ((A \wedge A') \rightarrow B) \text{ and } (A \rightarrow B) \rightarrow (\neg B \rightarrow \neg A)$$

and as inference rules:

"if $\models A \rightarrow B$ then $\models (A \wedge A') \rightarrow B$" and "if $\models A \rightarrow B$ then $\models \neg B \rightarrow \neg A$".

The requirements of absence of monotony and contraposition can therefore be formulated in two different ways. The first leads to the study of logics of the operator $\Rightarrow$ that fail to validate the formulas

$$(A \Rightarrow B) \rightarrow ((A \wedge A') \Rightarrow B) \text{ and } (A \Rightarrow B) \rightarrow (\neg B \Rightarrow \neg A).$$

[3]The example is Goodman's (1947), who proposes the requirement that A' should be *cotenable* with A for such an inference. His paper is dedicated to the quest of a definition of such a conditional; however, having discussed several unsatisfactory proposals he ends up defining cotenability in terms of the conditional, thus resulting in a circular definition.

[4]Precisely, for Nute's weak conditional logic W the following holds: "Any $\Box$-normal extension of the conditional logic W which is closed under one, is closed under all" (Nute 1980). It is however not the case that the principles are always equivalent: for example, in System C of Sect. 4.1, monotony and transitivity are equivalent, but monotony does not imply contraposition.

In the second perspective, we may reformulate the two properties using the deduction theorem as follows:

$$\text{“if } A \models B \text{ then } A \wedge A' \models B\text{” and “if } A \models B \text{ then } \neg B \models \neg A\text{”}$$

Then the object of study is no longer an operator of the object language, but rather a relation of logical consequence that is not in the object language but rather in the *metalanguage*. We are then interested in logical consequence relations $\mid\!\!\sim$ that have none of the following properties:

$$\text{“if } A \mid\!\!\sim B \text{ then } A \wedge A' \mid\!\!\sim B\text{” and “if } A \mid\!\!\sim B \text{ then } \neg B \mid\!\!\sim \neg A\text{”}$$

While the conditional operator $\Rightarrow$ was mainly studied by philosophers, the symbol of nonmonotonic consequence $\mid\!\!\sim$ was introduced by researchers in AI. The main difference is that just as the operator of necessity $\Box$ and the epistemic operator K (“the agent knows”), $\Rightarrow$ is an object language operator, while $\mid\!\!\sim$ is a metalanguage relation. Just as the symbol of logical consequence $\models$, it is therefore not part of the logical language. The operator $\Rightarrow$ is a weakening of $\rightarrow$, while the operator $\mid\!\!\sim$ is a weakening of $\models$. The term “weakening” has to be understood as “weakening of the logical properties”. Somewhat in contrast with that, in set-theoretic terms the relation $\mid\!\!\sim$ is a superset of the relation $\models$. Indeed, the nonmonotonic deduction is supposed to “go beyond” monotonic deduction: from the same hypotheses B, $\mid\!\!\sim$ should allow us to deduce more than $\models$. For each nonmonotonic relation $\mid\!\!\sim$, we expect that $B \models C$ implies $B \mid\!\!\sim C$. This postulate is called *supra-classicality*.[5]

The property of nonmonotony that is shared by $\Rightarrow$ and $\mid\!\!\sim$ takes two different forms: for $\Rightarrow$, it corresponds to the non-validity (falsifiability) of the axiom schema

$$(A \Rightarrow B) \rightarrow ((A \wedge A') \Rightarrow B).$$

For $\mid\!\!\sim$, it corresponds to the fact that there are A, A' and B such that $A \mid\!\!\sim B$ and $A \wedge A' \not\mid\!\!\sim B$. The rejection of monotony of the relation $\mid\!\!\sim$ is illustrated by resorting to similar counter-examples. The classical example in AI is the proposition “if Tweety is a bird (A) then Tweety flies (B)” which should not allow us to conclude that “if Tweety is a bird (A) and Tweety is a penguin (A') then Tweety flies (B)”.

In the next three sections we introduce the three families of formalisms that we have mentioned above: modal logics (Sect. 2), epistemic logics (Sect. 3), and conditional logics (Sect. 4). For each family we give two important formalisms. In the last Sect. 5 we revisit conditional logics in the light of a family of logics called dynamic epistemic logics (DEL), which are modal logics that became popular about 20 years ago. Just as the conditional operator, their dynamic operator is a binary

[5]The terms ‘postulate’ and ‘axiom’ both designate formal properties that are desired to hold. As customary we call such a property ‘axiom’ when it is formulated in the object language and ‘postulate’ when it is formulated in the metalanguage (see e.g. the AGM revision postulates Alchourrón et al. 1985). Inference rules are therefore particular postulates having one or more object language formulas as premisses and a single object language formula as conclusion.

modal operator (relating two formulas). In its most basic form it is written $[A!]B$ and is read "after the public announcement of A, B is the case". It can be viewed as a subjective, epistemic version of a conditional operator that is evaluated w.r.t. an agent's beliefs. This becomes clear when one reads $[A!]B$ as "if the agent learns that A then B will be the case". We study the logical properties of this operator from the conditional logic perspective: somewhat surprisingly, it will turn out that while the principles of the basic conditional logic CK are valid, almost all the principles beyond those of CK are invalid. As we will see, most of them fail when A has the form of particular epistemic formulas, the so-called "Moore sentences". The latter are formulas of the form $A \wedge \neg \mathsf{K} A$ ("A is true and the agent does not know this").

2 Two Basic Modal Logics

We now present the basic modal logic K and its extension S5. The latter has been chosen by many as the logic of knowledge.

Formulas are built from a countable set of propositional variables Prp and the operators $\neg$ and $\wedge$ of propositional logic, plus the modal operator $\Box$. Formally, the modal language is defined by the following grammar:

$$A ::= p \mid \neg A \mid A \wedge A \mid \Box A$$

where p is a propositional variable. The formula $\Box A$ is read "A is necessary". We use A, B, C,…, to denote formulas.

Here are some examples of formulas: $A = \Box p \rightarrow p$, read "if p is necessary then p is true" and $B = \Box p \rightarrow \Box\Box p$, read "if p is necessary then p is necessarily necessary". We will see that neither of these formulas is valid in the modal logic K, while both are valid in modal logic S5. Finally, the formula $C = p \rightarrow \Box p$ ("if p is true then p is necessarily true") will be invalid in both K and S5.

The operators $\top$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined as usual by the following abbreviations: $\top$ is $p \vee \neg p$, for some arbitrary propositional variable p; $A \vee B$ is $\neg(\neg A \wedge \neg B)$, $A \rightarrow B$ is $\neg A \vee B$, and $A \leftrightarrow B$ is $(A \rightarrow B) \wedge (B \rightarrow A)$. Finally, the formula $\Diamond A$ abbreviates $\neg\Box\neg A$. It can be read "A is possible".

2.1 The Modal Logic K

We now briefly present the semantics and the axiomatisation of the basic normal modal logic K (whose name honours its inventor Saul Kripke).

The *models* of K are triples of the form $M = \langle W, R, V \rangle$ where W is a non-empty set ("the possible worlds"), $R \subseteq W \times W$ is a binary relation on W ("the accessibility relation"), and $V : \mathsf{Prp} \longrightarrow 2^W$ is a valuation associating to each propositional variable p its extension $V(p) \subseteq W$: the set of possible worlds where p is true.

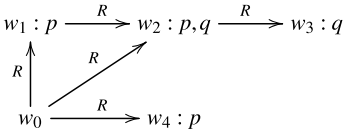


Fig. 1 An example Kripke model M

Figure 1 provides an example model M. Its set of possible worlds is $W = \{w_0, \ldots, w_4\}$, the accessibility relation is

$$R = \{(w_0, w_1), (w_0, w_2), (w_0, w_4), (w_1, w_2), (w_2, w_3)\}$$

and the valuation V is such that $V(p) = \{w_1, w_2, w_4\}$ and $V(q) = \{w_2, w_3\}$.

The satisfaction relation $\Vdash$ determines whether a formula is true in a world of a model. $M, w \Vdash A$ reads "in M, A is true at w" and is defined recursively as follows:

$$\begin{array}{ll} M, w \Vdash p & \text{iff } w \in V(p), \text{ for } p \in \mathsf{Prp} \\ M, w \Vdash \neg A & \text{iff } M, w \nVdash A \\ M, w \Vdash A \wedge B & \text{iff } M, w \Vdash A \text{ and } M, w \Vdash B \\ M, w \Vdash \Box A & \text{iff } M, v \Vdash A \text{ for every } v \text{ such that } (w, v) \in R \end{array}$$

In the model M of Fig. 1 we have for example $M, w_0 \Vdash \neg p \wedge \Box p \wedge \neg\Box\neg p$, $M, w_0 \Vdash \neg\Box q \wedge \neg\Box\neg q$, $M, w_0 \Vdash \neg\Box(p \wedge q)$, $M, w_0 \Vdash \Box(\Box q \rightarrow p)$.

A formula A is *valid* in K if and only if $M, w \Vdash A$ for every world w of every model M of K. The formula A is *satisfiable* in K if its negation $\neg A$ is not valid in K. The model M of Fig. 1 illustrates that the formula $\Box A \rightarrow A$ is not valid in K: indeed, $M, w_0 \Vdash \neg p \wedge \Box p$; and as—by the properties of material implication of classical logic—the formula $\neg p \wedge \Box p$ is equivalent to $\neg(\Box p \rightarrow p)$ we have that $M, w_0 \nVdash \Box p \rightarrow p$. This also illustrates that the formula $\neg p \wedge \Box p$ is satisfiable in K.

Here is the *axiomatisation* of the set of formulas that are valid in K (Chellas 1980):

(Class) every axiom schema of propositional logic
(M) $\Box(A_1 \wedge A_2) \rightarrow (\Box A_1 \wedge \Box A_2)$
(C) $(\Box A_1 \wedge \Box A_2) \rightarrow \Box(A_1 \wedge A_2)$
(N) $\Box\top$
(R.MP) $\dfrac{A \quad A \rightarrow B}{B}$
(R.E) $\dfrac{A_1 \leftrightarrow A_2}{\Box A_1 \leftrightarrow \Box A_2}$

(M) is called the axiom of monotony and its converse direction (C) is called the axiom for the conjunction. (N) is called the axiom of necessity. (R.MP) and (R.E) are respectively the rules of modus ponens and equivalence.

We observe that the inference rule of monotony

$$\text{(R.M)}\ \frac{A \rightarrow B}{\Box A \rightarrow \Box B}$$

is derivable from (M) by the rule (R.E).

A formula is *provable* modal logic K if it is derivable from instances of the axioms (Class), (M), (C) and (N), by the inference rules (R.MP) and (R.E).

Our axiomatisation is sound: every provable formula is valid. It is also complete: every valid formula is provable.

2.2 The Modal Logic S5

The models of S5 are a sub-class of the class of models of K: the class of models where the accessibility relation is an equivalence relation. The formulas that are valid in that class of models can be characterised by adding three further axioms to the axiomatisation of K[6]:

(T) $\Box A \rightarrow A$
(4) $\Box A \rightarrow \Box\Box A$
(5) $\neg\Box A \rightarrow \Box\neg\Box A$

The logic S5 is considered by many philosophers as *the* logic of necessity. It is also considered in AI as *the* logic of knowledge. Let us however observe that its *omniscience* properties can be criticised as being too strong: while a 'real' agent typically does not know all the logical consequences of her knowledge, the rule of monotony (R.M) stipulates exactly that. A realistic agent also does not know everything she knows, and a fortiori she does not know everything she does not know. These two principles are called positive and negative introspection, and are expressed by the axioms (4) and (5).

In place of $\Box$, the modal operator of knowledge is often noted K (from "know"). Up to now we have only considered the case of a single agent; it is possible to index the operator K by the name of an agent and to write e.g. $\mathsf{K}_1 p \wedge \neg\mathsf{K}_2 p$ to express that agent 1 knows that p and that agent 2 does not know that p. In an epistemic interpretation the axiom schemas (4) and (5) express what is called positive introspection ("I know what I know") and negative introspection ("I know what I don't know"). The schema (T) says that knowledge is true, which distinguishes it from beliefs (which in the case of a wrong belief is false).

The concept of knowledge that is captured here is binary: either the agent knows that A, or she does not know it. There are some approaches in the literature equipping the modal operator with degrees in order to express more fine-grained distinctions. For example, the logic proposed by Noël Laverny and Jérôme Lang (2005) possesses

[6] Actually axiom (4) is superfluous: it can be derived from (T) and (5).

operators of belief $\mathsf{K}^{\geq k}$, and $\mathsf{K}^{\geq k} A$ reads "A is true for the agent with a degree of at least k".[7]

Numerous other interpretations of the operator $\Box$ exist. For example, the formula $\Box A$ can be read "the agent intends that A" or "it is obligatory that A". We refer readers to the presentation of these modalities in chapter "Qualitative Reasoning" of this volume (Qualitative reasoning about time and space), chapter "Norms and Deontic Logic" of this volume (Norms and deontic logic), chapter "Reasoning about Action and Change" of this volume (Reasoning about action and change), and chapter "Formalization of Cognitive-Agent Systems, Trust, and Emotions" of this volume (Formalization of cognitive-agent systems, trust and emotions).

3 Two Logics of Conditionals

Our presentation of conditional logics is essentially syntactical: we focus on the reasoning principles and only give the basics of the semantics. The formulas of the language of conditional logics are built from a countable set of propositional variables together with the operators $\neg$ and $\wedge$ of propositional logic, plus the conditional operator $\Rightarrow$. Precisely, the language is defined by the following grammar:

$$A ::= p \mid \neg A \mid A \wedge A \mid A \Rightarrow A$$

where p is a propositional variable. The formula $A \Rightarrow C$ is read "if A then C". Throughout the chapter, we are going to use A for the antecedent and C for the consequent of a conditional.

We economise parentheses by considering that $\Rightarrow$ binds weaker than $\neg$ and stronger than the other operators. Therefore $\neg A \Rightarrow C \wedge B$ is $((\neg A) \Rightarrow C) \wedge B$ and $A \Rightarrow C \rightarrow B$ is $(A \Rightarrow C) \rightarrow B$.

3.1 The Normal Conditional Logic CK and Its Extensions

The semantics of normal conditional logics is due to Stalnaker (1968) and is based on *selection functions*. The basic logic is called CK, "C" standing for "conditional" and "K" standing for "Kripke".

A *model* of CK is a triple of the form $\langle W, f, V\rangle$ where W is a set of possible worlds (just as in the logic K), $f : (W \times 2^W) \longrightarrow 2^W$ is a mapping—called selection function—associating every 'world/set of worlds' couple with a set of worlds: intuitively, $f(w, U)$ is the set of those worlds of U that are most similar to w. (We note that this intuition should not be taken too literally: the basic logic CK allows for models where the set $f(w, U)$ is not contained in U, as well as for models where

[7] We have adapted the original notation.

$f(w, U)$ contains worlds v where no propositional variable has the same truth value at w and v. Just as in logic K, the function $V : \mathsf{Prp} \longrightarrow 2^W$ is a valuation.

The satisfaction relation $\Vdash$ links a model, a world of that model and a formula. The definition is recursive. The cases of propositional variables, negation and conjunction are as for K, while the case of the conditional operator is:

$$M, w \Vdash A \Rightarrow C \text{ iff } M, v \Vdash C \text{ for every } v \in f(w, ||A||_M)$$

where $||A||_M$ is the set of A-worlds of M, defined as: $||A||_M = \{v \mid M, v \Vdash A\}$. Thus, $f(w, ||A||_M)$ provides the set of A-worlds that are most similar to w (with the above proviso about the notion of similarity in logic K).

Just as for the logic K, a formula A is valid in CK if and only if $M, w \Vdash A$ for every world w of every model M.

Here is the *axiomatisation* of the set of valid formulas in the basic conditional logic CK (Chellas 1975, 1980).

(Class) every axiom schema of propositional logic

(C.M) $A \Rightarrow (C_1 \wedge C_2) \rightarrow (A \Rightarrow C_1 \wedge A \Rightarrow C_2)$

(C.C) $(A \Rightarrow C_1 \wedge A \Rightarrow C_2) \rightarrow A \Rightarrow (C_1 \wedge C_2)$

(C.N) $A \Rightarrow \top$

(R.MP) $\dfrac{A \quad A \rightarrow C}{C}$

(RC.EA) $\dfrac{A_1 \leftrightarrow A_2}{A_1 \Rightarrow C \leftrightarrow A_2 \Rightarrow C}$

(RC.EC) $\dfrac{C_1 \leftrightarrow C_2}{A \Rightarrow C_1 \leftrightarrow A \Rightarrow C_2}$

(C.M) is called the axiom of monotony (for the consequent of the conditional) and the symmetric (C.C) is called the axiom for conjunction; (C.N) is the axiom of necessity. The reader may observe the symmetry of these axioms with the axioms (C), (M), and (N) for modal logic K of Sect. 2.1. (R.MP), (RC.EA) and (RC.EC) are respectively the rules of modus ponens, of equivalence in the antecedent and of equivalence in the consequent.

Just as for modal logics, a formula is *provable in CK* if it is derivable from axiom instances by the inference rules.

The axiomatisation above is sound and complete: every theorem of CK is valid in the models of CK, and every formula that is valid in the models of CK is a theorem of CK.

Similarly to the logic K, the inference rule

(RC.M) $\dfrac{C_1 \rightarrow C_2}{A \Rightarrow C_1 \rightarrow A \Rightarrow C_2}$

is derivable from (C.M) by the rule (RC.EC).

The logic CK is a rather weak logic: even a reasonably-looking principle such as

(C.ID) $A \Rightarrow A$

cannot be proved in CK. One can however guarantee the validity of this schema by modifying the semantics of CK: it suffices to evaluate the validity of formulas only in those models where $f(w, U) \subseteq U$ for every w and U.

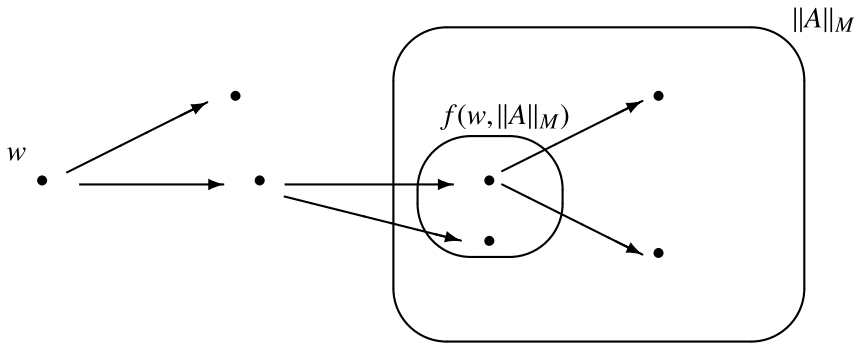


Fig. 2 The similarity relation determines a selection function

The above observation about the derivability of (RC.M) allows us to prove that the extension of the logic CK by the axiom (C.ID) satisfies the supra-classicality postulate. The latter here takes the form of the inference rule

$$\frac{A \rightarrow C}{A \Rightarrow C}$$

that can be derived by deducing in a first step $A \Rightarrow A \rightarrow A \Rightarrow C$ from the hypothesis $A \rightarrow C$; then, from (C.ID) and $A \Rightarrow A \rightarrow A \Rightarrow C$ one deduces $A \Rightarrow C$ by the rule of modus ponens (R.MP).

Several principles were discussed in the literature. We are going to overview them in the next section.

3.2 The Logic of Lewis–Burgess CL and Its Extensions

David Lewis (1973) proposed to replace Stalnaker's selection functions by a more sophisticated construction that he called a *system of spheres*: to each possible world there is associated a set of nested spheres. Burgess generalised sphere systems to partial preorders associated to worlds: such preorders also allow to compare which of two worlds is most similar to a given world. These preorders can be viewed as orders of plausibility or orders of comparative possibility.[8] In the case of a total order we retrieve qualitative possibility orderings (Fariñas del Cerro and Herzig 1991). Given a partial preorder, we can build a selection function, as illustrated in Fig. 2. The converse fails to hold.

There is no established name for the basic conditional logic of the sphere semantics. We here call it CL, where "L" is in honour of David Lewis.

[8]The link with theories of uncertainty is deepened in chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" of this volume.

A *model of CL* is a triple of the form $M = \langle W, \{\leq_w : w \in W\}, V\rangle$, where W is a non-empty set of possible worlds and V is a valuation as before. Every $\leq_w$ is a binary relation on W.[9]

Let $S_w = \{u \mid \exists v\ u \leq_w v\}$ for every $w \in W$. The elements of S_w can be seen as worlds that are accessible from w. M has to satisfy the following conditions:

- for every w, the restriction of $\leq_w$ to S_w is a partial preorder on S_w (so $\leq_w \cap (S_w \times S_w)$ is a reflexive and transitive relation).
- for every $w \in W$, $\leq_w$ satisfies the *"limit assumption"*: for every formula A and worlds $w, v \in W$, if $v \in ||A||_M \cap S_w$ then there is a $u \in ||A||_M$ such that $u \leq_w v$, and for every $u' \in ||A||_M$, $u' \leq_w u$ implies $u \leq_w u'$.

Then $f(w, U)$ can be defined as the set of elements of U that are both in S_w and minimal w.r.t. $\leq_w$:

$$f(w, U) = \min_{\leq_w}(U \cap S_w)$$

The limit assumption guarantees that $f(w, ||A||_M) = \min_{\leq_w}(||A||_M \cap S_w)$ is non-empty as soon as $||A||_M \cap S_w$ is so.

Given the above definition of the selection function, the satisfaction relation can be defined just as for CK.

We have seen that from the relations $\leq_w$ one can always build a selection function. Therefore the set of valid formulas of the logic CL contains those of the logic CK. Thus, an axiomatisation of CL can be obtained by adding to the axiomatisation of CK the following axiom schemas (see Herzig 1998).

(ID) $A \Rightarrow A$
(CA) $(A_1 \Rightarrow C \wedge A_2 \Rightarrow C) \rightarrow (A_1 \vee A_2) \Rightarrow C$
(CSO) $(A_1 \Rightarrow A_2 \wedge A_2 \Rightarrow A_1) \rightarrow (A_1 \Rightarrow C \leftrightarrow A_2 \Rightarrow C)$
(MOD$_0$) $A \Rightarrow \bot \rightarrow (A \wedge A') \Rightarrow \bot$
(MOD) $\neg A \Rightarrow A \rightarrow A' \Rightarrow A$
(ASC) $(A \Rightarrow A' \wedge A \Rightarrow C) \rightarrow (A \wedge A') \Rightarrow C$
(RT) $(A \Rightarrow A' \wedge (A \wedge A') \Rightarrow C) \rightarrow A \Rightarrow C$

(ID) is the axiom of identity. (CA) means "conjunction of antecedents". (ASC) is sometimes called cautious monotony and (RT) is restricted transitivity. (ASC) and (RT) are symmetric; the cumulativity axiom

(CUM) $A \Rightarrow A' \rightarrow (A \Rightarrow C \leftrightarrow (A \wedge A') \Rightarrow C)$

combines them in a single axiom.

Remember that Lewis' sphere systems are nested: they are total preorders. The class of all these CL-models can be axiomatised by adding to the axiomatisation of CL the following axiom schema:

(CV) $(A \Rightarrow C \wedge \neg(A \Rightarrow \neg A')) \rightarrow (A \wedge A') \Rightarrow C$

[9] As noted in Makinson (1993), the condition of transitivity that was initially imposed by Burgess can be abandoned. One might as well restrict the $\leq_w$ to strict preorders.

Observe that this axiom is stronger than CL's axiom (ASC): instead of $A \Rightarrow A'$ we have the weaker $\neg(A \Rightarrow \neg A')$.[10] Let us interpret the axioms (ASC) and (CA) in the light of Goodman's cotenability: according to (ASC), everything that follows from the antecedent (w.r.t. $\Rightarrow$) is cotenable with it; according to (CV), everything that is consistent with the antecedent (w.r.t. $\Rightarrow$) is cotenable with it.

Let us end this section by noting that conditional logics are close to logics of update via what is called the *Ramsey Test* (Herzig 1998). If $B \diamond A$ denotes the update of B by A then $(B \diamond A) \rightarrow C$ is valid if and only if $B \rightarrow (A \Rightarrow C)$ is valid. Exploiting that correspondence one can almost systematically translate each axiom, in both senses (Ryan and Schobbens 1997).

4 From Default Logic to Two Classes of Nonmonotonic Formalisms

Right from the start of AI there was an agreement in the community that commonsense reasoning requires default reasoning and that the latter is by nature *nonmonotonic*: the fact that the premise A allows to infer C does not guarantee that the premise $A \wedge A'$ allows to infer C.

Default logic (Reiter 1980) was one of the first nonmonotonic formalisms and is certainly the most popular one. The idea underlying default logic can be related to the expression "unless proven otherwise", meaning that one holds the conclusion true unless it causes a contradiction with what is known. A standard example associated to default logic is the piece of information according to which "birds normally fly" (where "fly" is understood as "able to fly"). Indeed, this piece of information can be represented according to the schema "a bird flies, unless proven otherwise". In default logic, this is formally translated by a default rule

$$\frac{bird(x) : flies(x)}{flies(x)}$$

Intuitively the rule tells us: if x is a bird and it is not contradictory to infer that it flies then infer that it flies.

We obviously have to clarify some points, first and foremost what "contradictory" refers to. But let us begin by the basic definitions.

A *default rule* (or, for short, a *default*) is defined as an expression

$$\frac{A : B_1, \ldots, B_n}{C}$$

where $A, B_1, \ldots, B_n$ and C are formulas of first-order predicate logic.

[10] Indeed, the first implies the second in presence of the axiom (MOD_0).

A *default theory* is a couple (W, D) where W is a set of formulas of first-order predicate logic and D is a set of defaults without free variables.[11] Intuitively, W expresses what is certain and D expresses laws allowing for exceptions. Here is an illustration.

$$W = \left\{ \begin{array}{c} cat(Sylvester), \\ bird(Tweety), \\ bird(Tyty), \\ ostrich(Tyty), \\ \forall x\ ostrich(x) \rightarrow \neg flies(x) \end{array} \right\} \quad \text{and} \quad D = \left\{ \frac{bird(x) : flies(x)}{flies(x)} \right\}$$

This default theory allows, among others, to conclude $flies(Tweety)$ and $\neg flies(Tyty)$ (but neither $flies(Tyty)$ nor $flies(Sylvester)$).

Formally, the consequences of a default theory (i.e., the conclusions that one can deduce from it) are grouped into extensions that are defined as follows:

A set of formulas E is an *extension* of a default theory (W, D) if and only if $E = \bigcup_{i=0}^{\infty} E_i$ where

$$E_0 = W$$
$$E_{i+1} = Th(E_i) \cup \left\{ C \mid \frac{A : B_1, \ldots, B_n}{C} \in D \text{ such that } A \in E_i \text{ and } \neg B_1 \notin E, \ldots, \neg B_n \notin E \right\}$$

where Th is the consequence operator of classical logic.

Beware: there is no typo, the tests of non-contradiction $\neg B_m \notin E$ are indeed performed w.r.t. E and not w.r.t. E_i. So the computation of extensions is not constructive because it appeals to the result of the computation. Actually extensions are defined as solutions of a fixed-point equation; however, the above characterisation is much more popular.

Let us come back to our example. The computation starts by $E_0 = W$. Then, at the level of E_1, we check whether we can "apply" the default

$$\frac{bird(Tweety) : flies(Tweety)}{flies(Tweety)}$$

because $bird(Tweety)$ is in E_0 and $\neg flies(Tweety)$ is "felt" as being obtainable neither now *nor later in the sequence* $E_2, E_3, \ldots$. The application of the default thus introduces $flies(Tweety)$ into E_1. On the contrary, $flies(Tyty)$ is not introduced because $\neg flies(Tyty)$ belongs to E (indeed, $\neg flies(Tyty)$ is a classical consequence of W, and therefore of E_0, which makes $\neg flies(Tyty)$ belong to E_1; however, E_1 is included in E by construction). In other words, the default

$$\frac{bird(Tyty) : flies(Tyty)}{flies(Tyty)}$$

[11] Actually defaults with free variables are considered to be abbreviations to be replaced by their closed instances.

fails to apply. And the default

$$\frac{bird(Sylvester) : flies(Sylvester)}{flies(Sylvester)}$$

does not apply either because $bird(Sylvester)$ is neither a classical consequence of E_0, nor of any other E_i). To sum it up, our default theory has an extension that contains W together with $flies(titi)$ and their classical consequences.

Default logic determines a kind of nonmonotonic inference because the supplementary formulas of W may block the application of a default. In our example, if $ostrich(Tweety)$ is added to W then $flies(Tweety)$ can no longer be inferred.

From the point of view of knowledge representation, default logic has some particular features (Besnard 1989). On the one hand, a default theory may have zero, one, or several extensions. These extensions intuitively stand for alternative collections of conclusions. Clearly, the existence of cases where there is no conclusion is a serious problem, and several approaches tried to delimit classes of default theories where either the existence of an extension is guaranteed (e.g. Etherington 1987), or where the definition of extensions is modified (e.g. Delgrande et al. 1994). On the other hand, some reasoning schemas are not preserved. One example is contraposition: "If A then, unless proven otherwise, C" allows—except if there is a contradiction—to conclude C when A is established, but does not necessarily allow to infer $\neg A$ when $\neg C$ is established. Finally, defaults are not expressions of the language: it is impossible to deduce a default, to negate a default,…(see Doherty and Łukaszewicz 1992).

The 1980s saw a series of propositions of alternative definitions, in particular circumscription (McCarthy 1980; McDermott and Doyle 1980; McCarthy 1986, 1990) and autoepistemic logic (Moore 1985; Konolige 1995). Several contributions proved the equivalence of fragments of default logic and other nonmonotonic formalisms (e.g. Marek and Truszczynski 1989). The book Léa Sombé (1994) contains an overview and comparisons of the literature existing at that time.

It is only by the end of the 1980s that Gabbay, Lehmann and others proposed to complete these concrete consequence relations by a study of their general properties (Gabbay 1985; Bell 1990; Kraus et al. 1990; Stalnaker 1992; Lehmann and Magidor 1992; Arló Costa and Shapiro 1992; Crocco and Lamarre 1992; Makinson 1994; Gärdenfors and Makinson 1994; Crocco et al. 1995; Levi 1996). These scholars opted for a research avenue that differs from conditional logics: as we have said in the introduction, they opted for a nonmonotonic consequence relation $\mid\!\approx$ that is part of the metalanguage, and is therefore different in nature from the operator conditional $\Rightarrow$ that is part of the object language. However, the postulates for $\mid\!\approx$ where much inspired by the axioms for conditionals that were studied more than 10 years ago. The 1980s and 1990s saw an extensive debate about the desiderata for relations $\mid\!\approx$. It turned out that default logic violates almost all of these desiderata. This then motivated the elaboration of new, concrete nonmonotonic inference mechanisms such as variants of default logic (see Brewka 1991 for systems satisfying cumulativity).

We here note that there is a link between the postulates for nonmonotonic consequence relations and the AGM postulates for revision operators. The latter allow to revise a belief base KB by a new piece of information A. The result of that revision is noted $KB * A$.[12] For a fixed belief base KB, $A \mid\!\approx C$ can be identified with $KB * A \models C$.

In the rest of this section we present two systems of nonmonotonic consequence: preferential formalisms and rational formalisms.

4.1 Preferential Formalisms

Here are the postulates for *cumulative* inference relations: the so-called system C. We stick to the nomenclature of conditional logics in order to highlight the tight links between the two families of formalisms.

(P-RC.EA) if $A \models A'$ and $A' \models A$ then ($A \mid\!\approx C$ iff $A' \mid\!\approx C$)
(P-RC.M) if $C \models C'$ and $A \mid\!\approx C$ then $A \mid\!\approx C'$
(P-ID) $A \mid\!\approx A$
(P-ASC) if $A \mid\!\approx A'$ and $A \mid\!\approx C$ then $A \wedge A' \mid\!\approx C$
(P-RT) if $A \mid\!\approx A'$ and $A \wedge A' \mid\!\approx C$ then $A \mid\!\approx C$

The following names for the above postulates can be found in the literature: 'left logical equivalence' for (P-RC.EA); 'right weakening' for (P-RC.M); 'reflexivity' for (P-ID); 'cautious monotony' for (P-ASC); 'cautious cut' for (P-RT). Just as for conditionals, the cumulativity postulate combines cautious monotony and cautious cut:

(P-CUM) if $A \mid\!\approx A'$ then ($A \mid\!\approx C$ iff $A \wedge A' \mid\!\approx C$)

Preferential inference relations moreover satisfy the following postulate:

(P-CA) if $A_1 \mid\!\approx C$ and $A_2 \mid\!\approx C$ then $A_1 \vee A_2 \mid\!\approx C$

This postulate is called 'or rule' in the literature, and the formalism is called system P.

The semantics of system P is in terms of partial preorders and correspond with that of the conditional logic CL of Sect. 3.2 (Kraus et al. 1990).

4.2 Rational Formalisms

A further postulate corresponding to the axiom (CV) of conditional logics has been studied. It is not necessarily satisfied by preferential relations.

[12] We refer to chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume for an exposition of AGM theory.

(CV) if $A \mid\!\!\sim C$ and $A \not\mid\!\!\sim \neg A'$ then $A \wedge A' \mid\!\!\sim C$

The dedicated term in the literature for that postulate is 'rational monotony'. Preferential inference relations moreover satisfying postulate (CV) are called rational inference relations. Such relations are the 'strongest' nonmonotonic systems, in the sense that they satisfy 'as much as possible' of the properties of the classical consequence relation $\models$: intuitively, if we add any other interesting property of the classical inference relation to the list of postulates of system P plus (CV) then the resulting set of postulates only allows for monotonic inference relations.

The semantics of rational formalisms is in terms of total preorders (Kraus et al. 1990) and corresponds to the extension of the conditional logic CL by the axiom (CV) that we have presented in the end of Sect. 3.2.

Several experimental verifications of the psychological plausibility of the postulates have been undertaken, see e.g. Neves et al. (2002), Benferhat et al. (2004). They have by and large confirmed the psychological validity of these inference rules.

5 Conditional Logics in the Light of Dynamic Epistemic Logics

Dynamic epistemic logics were introduced almost 30 years ago (Plaza 1989) and were intensely investigated since about 20 years ago. (Gerbrandy and Groeneveld 1997; Gerbrandy 1999; van Ditmarsch 2000; van Benthem 2006). We here consider the simplest dynamic epistemic logic: Public Announcement Logic (PAL). For our purposes it will be enough to consider the case of a single agent.

The language of PAL is defined by the following grammar:

$$A ::= p \mid \neg A \mid A \wedge A \mid \mathsf{K}A \mid [A!]A$$

The formula $\mathsf{K}A$ reads "the agent knows that A", and the formula $[A!]C$ reads "if A is publicly announced then B will be the case afterwards".

Given their reading, formulas of the form $[A!]C$ are therefore particular conditionals; given the presence of the operator K one may call them epistemic conditionals. In the present section we are going to explore that perspective. We start by showing that the fundamental requirements for conditional operators are fulfilled: absence of monotony and of contraposition; we are then going to examine the status of the axioms for conditionals that we have seen in Sect. 3.

5.1 Dynamic Epistemic Logics: Public Announcement Logic

The models of PAL are nothing but the models of (single-agent) S5 that we have seen in Sect. 2. The epistemic operator K is interpreted in the same way as the modal

operator $\Box$, while the dynamic operator $[A!]$ is interpreted by a restriction of the model to worlds where A is true.

$$M, w \Vdash \mathsf{K}A \text{ iff } M, v \Vdash A \text{ for every } v \text{ such that } (w, v) \in R$$
$$M, w \Vdash [A!]C \text{ iff } M, w \nVdash A \text{ or } M^A, w \Vdash C$$

The model $M^A = \langle W^A, R^A, V^A \rangle$ is the restriction of M to the set of worlds of M where A is true and is defined as follows.

$$\begin{aligned} W^A &= ||A||_M \\ R^A &= R \cap (||A||_M \times ||A||_M) \\ V^A(p) &= V(p) \cap ||A||_M \end{aligned}$$

We are not going to give the axiomatisation of PAL here because it does not serve our purposes and refer the reader to the literature we have cited above. We just mention that the dynamic operator $[A!]$ is a normal modal operator and that the following inference rule (which resembles the rule for conditional logics (RC.EA)) is admissible:

(RA.EA) $$\frac{A_1 \leftrightarrow A_2}{[A_1!]C \leftrightarrow [A_2!]C}$$

Observe that the schema $[A!]A$ is invalid. To see this it suffices to replace A by the so-called Moore sentence $p \wedge \neg\mathsf{K}p$: the formula $[[p \wedge \neg\mathsf{K}p!]](p \wedge \neg\mathsf{K}p)$ is not valid.

5.2 *Public Announcement Logic as a Conditional Logic*

Which of the axioms for conditionals of Sect. 3 are satisfied by the dynamic operators $[A!]$ of PAL?

Let us start by observing that the principles that are rejected by conditional logics are not validated by PAL either. Consider first monotony $[A!]C \rightarrow [A \wedge A'!]C$ and replace A and C by $\neg\mathsf{K}p$ and A' by p: then the formula $[\neg\mathsf{K}p!]\neg\mathsf{K}p$ is valid, while $[\neg\mathsf{K}p \wedge p!]\neg\mathsf{K}p$ is not. As to contraposition $[A!]C \rightarrow [\neg C!]\neg A$, replace A by the propositional variable p and C by $\mathsf{K}p$: then the formula $[p!]\mathsf{K}p$ is valid in PAL, while $[\neg\mathsf{K}p!]\neg p$ is not.

The logic PAL is therefore a serious candidate for a conditional logic according to the criteria proposed by Donald Nute (1984). As it allows us to reason about knowledge thanks to the epistemic operator K, PAL can be considered to be an interesting basis for a logic of so-called *epistemic conditionals* (Lindström and Rabinowicz 1995; Arló Costa 1995).

Our next observation is that all the principles for the basic conditional logic CK are valid. This is the case because, on the one hand, the $[A!]$ are normal modal operators;

and on the other hand, identifying $A \Rightarrow C$ with $[A!]C$, the inference rule (RA.EA) of PAL is nothing but the inference rule (RC.EA) of CK.

What about the other principles such as $A \Rightarrow A$?

Theorem 1 *The formula schema (MOD$_0$):* $[A!]\bot \rightarrow [A \wedge A'!]\bot$ *is valid in PAL.*

Proof Consider an arbitrary model M and world w of M. Then $M, w \Vdash [A!]\bot$ if and only if $M, w \nVdash A$. Hence $M, w \nVdash A \wedge A'$, and therefore $M, w \Vdash [A \wedge A'!]\bot$.

What about the other principles? We have already observed that $[A!]A$ is not valid in PAL. But the situation is more dramatic than that: as it turns out, all the axioms that are proper to CL other than (MOD$_0$) are invalid in PAL.

Theorem 2 *The following formula schemas are invalid in PAL.*

1. $[A!]A$
2. $([A_1!]C \wedge [A_2!]C) \rightarrow [A_1 \vee A_2!]C$
3. $([A_1!]A_2 \wedge [A_2!]A_1) \rightarrow ([A_1!]C \leftrightarrow [A_2!]C)$
4. $\neg[A!]A \rightarrow [A'!]A$
5. $([A!]A' \wedge [A!]C) \rightarrow [A \wedge A'!]C$
6. $([A!]A' \wedge [A \wedge A'!]C) \rightarrow [A!]C$

Proof Almost all non-validities can be established by means of a Moore sentence.

1. As said above, it suffices to replace A by the Moore sentence $p \wedge \neg \mathsf{K} p$ to see that the schema (ID): $[A!]A$ is invalid.
2. In the schema (CA), replace A_1 by p, A_2 by $\neg p$ and C by $\mathsf{K} p \vee \mathsf{K} \neg p$. Then on the left side, both $[A_1!]C = [p!](\mathsf{K} p \vee \mathsf{K} \neg p)$ and $[A_2!]C = [\neg p!](\mathsf{K} p \vee \mathsf{K} \neg p)$ are valid in PAL, while on the right side, $[A_1 \vee A_2!]C = [p \vee \neg p!](\mathsf{K} p \vee \mathsf{K} \neg p)$ is equivalent to $[\top!](\mathsf{K} p \vee \mathsf{K} \neg p)$ (by the inference rule (RA.EA)), which is not valid.
3. In the schema (CSO), replace A_1 by p, A_2 by q (for some p, q such that $p \neq q$) and C by $\mathsf{K} p$. Then $[A_1!]A_2 \wedge [A_2!]A_1 = [p!]q \wedge [q!]p$ is equivalent to $p \leftrightarrow q$, and the latter formula does not imply $[A_1!]C \leftrightarrow [A_2!]C = [p!]\mathsf{K} p \leftrightarrow [q!]\mathsf{K} p$.
4. In the schema (MOD), replace A by $p \wedge \neg \mathsf{K} p$ and A' by $\top$. Then the formula $\neg[A!]A = \neg[p \wedge \neg \mathsf{K} p!](p \wedge \neg \mathsf{K} p)$ is valid in PAL. However, $[A'!]A = [\top!](p \wedge \neg \mathsf{K} p)$ is not.
5. In the last but one schema (ASC), replace A by $\neg \mathsf{K} p$, A' by p and C by $\neg \mathsf{K} p$. Take a model M and a world w of M such that $M, w \Vdash p \wedge \neg \mathsf{K} p$. Then $M, w \Vdash [A!]A' = [\neg \mathsf{K} p!]p$ and $M, w \Vdash [A!]C = [\neg \mathsf{K} p!]\neg \mathsf{K} p$, but $M, w \nVdash [A \wedge A'!]C = [\neg \mathsf{K} p \wedge p!]\neg \mathsf{K} p$.
6. In the last schema of restricted transitivity (RT), replace A by $p \wedge \neg \mathsf{K} p$, A' by $\mathsf{K} p$ and C by $\bot$. Then the two conjuncts on the left are valid:

$$
\begin{aligned}
[A!]A' &= [p \wedge \neg \mathsf{K} p!]\mathsf{K} p \\
[A \wedge A'!]C &= [p \wedge \neg \mathsf{K} p \wedge \mathsf{K} p!]\bot \\
&\leftrightarrow [\bot!]\bot \qquad \text{(by (RA.EA))}
\end{aligned}
$$

However, $[A!]C \;=\; [p \wedge \neg \mathsf{K} p!]\bot$ is not valid.

To sum it up, among the principles for conditional logics beyond CK that were proposed and much discussed in the literature, only (MOD_0) is valid in PAL. The counter-examples for (ID), (MOD), (ASC) and (RT) make use of the famous Moore sentences. It seems to us that our negative result might shed a new light on the debate about reasoning principles associated to "if…then…" constructions.

Let us note that apart from Moore sentences—where 'successful' formula schemas (such as $[A!]A$) and self-defeating schemas (such as $[A!]\neg A$) were studied—, the PAL literature focussed on valid formula *instances* and not on valid axiom schemas as it is customary in logic. The only article undertaking a schematic study is Holliday et al. (2011).

5.3 Discussion

One may object to our analysis that PAL provides a very special kind of conditionals because announcements have to be truthful. Thus, if A is false then A cannot be announced: the formula $\neg A \rightarrow (A \Rightarrow \bot)$ is valid. Therefore, what cannot be analysed in PAL are counterfactual conditionals: sentences of the form "if A then C" whose antecedent A is false. However, our analysis applies to open conditionals: conditionals where it is unknown whether the antecedent is true or not. Beyond that, we observe that one can as well reason about announcements that are not necessarily truthful: it suffices to adopt a variant of the semantics of PAL that is due to Jelle Gerbrandy (1999) and that was studied further by Barteld Kooi (2007). These authors redefine the truth condition unconditionally as:

$$M, w \Vdash [A!]C \text{ iff } M^A, w \Vdash C$$

where the restriction $M^A = \langle W^A, R^A, V^A \rangle$ of M to the set of worlds of M where the announcement A is true is defined as follows: $W^A = W$, $V^A = V$ and

$$R^A = R \cap (W \times ||A||_M)$$

In the restricted model the worlds where the announcement is false are therefore no longer eliminated from the model. Thus $\neg A \rightarrow [A!]\bot$ is no longer valid. This said, it has to be noted that the announcement still has to be compatible with the agent's beliefs (otherwise the agent's beliefs would become inconsistent). This variant perhaps better corresponds to an open conditional: the agent entertaining it among her beliefs ignores whether the antecedent is true or not. Other variants of the truth condition are studied in Balbiani et al. (2012).

6 Conclusion

In this chapter we have seen three fundamental concepts in knowledge representation: the dual modalities "necessary" ($\Box$) and "possible" ($\Diamond$) and two concepts of the "if…then…" kind: conditionals ($\Rightarrow$) and nonmonotonic inference relations ($\mid\!\!\sim$). By turning our attention towards dynamic epistemic logics and in particular PAL we have obtained a new view of conditional logics (and, mutatis mutandis, of nonmonotonic inference relations): somewhat surprisingly, almost all the axioms that were introduced and debated as reasonable principles for conditionals turned out to be untenable in the epistemic framework of PAL.

Overall, the field of conditional and nonmonotonic reasoning has not changed a lot during the last 20 years. The biannual Nonmonotonic Reasoning workshop series (NMR) is mainly affiliated with the Knowledge Representation and Reasoning conference (KR) (Kern-Isberner and Wassermann 2016). One may also note the more recent annual International Workshop on Defeasible and Ampliative Reasoning (DARe) series (Booth et al. 2017). A recent regain of interest can also be observed in philosophy (Pfeifer 2014; Bradley and Stefánsson 2017; Alenda et al. 2016; Girlando et al. 2016; Cross 2016; Douven 2016; Koutras and Rantsoudis 2017).

A domain of AI that gets more and more interested in nonmonotonic reasoning is description logics (presented in chapter "Reasoning with Ontologies" of this volume); indeed, many of the papers at the NMR and DARe workshops concern that topic. In these logics, a knowledge base is a couple $KB = \langle T, A\rangle$ where T is an ontology or terminology ('the TBox') and A is a set of facts or assertions ('the ABox'). The TBox is made up of *concept inclusions* of the form $C \sqsubseteq D$, such as $Student \sqsubseteq \neg Prof$ expressing that students and professors are disjoint. Suppose we would like to allow exceptions: a few students (such as PhD students acting as teaching assistants) are also professors. Several authors proposed to extend description logics by nonmonotonic reasoning mechanisms, starting with Reiter's default logic (Baader and Hollunder 1995) or with nonmonotonic modal extensions (Donini et al. 2002). The resulting formalisms were however criticised as being difficult to understand and having bad computational properties: they are often undecidable, which conflicts with the 'presupposition' underlying research in description logics that their *raison d'être* is to provide decidable formalisms. A new way of representing such ontologies was recently proposed by Giordano et al. (2013). Their extension of the basic description logic ALC contains an operator of typicality $\mathbf{T}$, allowing us to write $\mathbf{T}(Student) \sqsubseteq \neg Prof$: typical students are non-professors. Their logic has a semantics in terms of preferred models that generalises that for preferential formalisms of Sect. 4.1, and they show that it is characterised by the same postulates. They also show that their logic is decidable and that the problem of deciding the satisfiability of a knowledge base is EXPTIME complete. Given that that problem is already EXPTIME hard for the underlying monotonic logic ALC and given the criticisms of nonmonotonic extensions of ALC that we have reported this has to be considered as an interesting result.

Beyond description logics, nonmonotonic inference relations were studied in the domain of *hybrid knowledge bases*. These are formalisms merging knowledge bases

in the TBox+ABox form (just as description logics) with rules (just as logic programming) (Donini et al. 2002). The integration of the non-classical semantics of rules into description logics comes with new, interesting problems.

Finally, it is often considered that one of the most interesting forms of nonmonotonic reasoning is *answer set programming* (ASP) (Lifschitz 2008). The latter is a rather recent branch of logic programming providing a logically solid answer to the long-standing problem of the semantics of negation by failure (see chapter "Logic Programming" of volume 2 for an overview of logic programming). The associated inference relation typically allows to infer $\neg p$ from an empty program, for any atomic formula p. The underlying order thus gives priority to negative information. The bi-annual Logic Programming and Nonmonotonic Reasoning conference series (LPNMR) is dedicated to that topic (Balduccini and Janhunen 2017).

Acknowledgements Thanks are due to Ricardo Caferra for his careful reading of the first version of the French version of this chapter and to Henri Prade and Christos Rantsoudis for the same job on the present English version. Thanks are also due to Hans van Ditmarsch for his comments.

References

Alchourrón C, Gärdenfors P, Makinson D (1985) On the logic of theory change: partial meet contraction and revision functions. J Symb Log 50:510–530

Alenda R, Olivetti N, Pozzato GL (2016) Nested sequent calculi for normal conditional logics. J Log Comput 26(1):7–50. https://doi.org/10.1093/logcom/ext034

Arló Costa H (1995) Epistemic conditionals, snakes and stars. In: Crocco G, Fariñas del Cerro L, Herzig A (eds) Conditionals: from philosophy to computer science. Studies in logic and computation, vol 5. Oxford University, Oxford, pp 193–239

Arló Costa H, Shapiro S (1992) Maps between nonmonotonic and conditional logics. In: Nebel B, Rich C, Swartout W (eds) Proceedings of the 4th international conference on knowledge representation and reasoning (KR'92). Morgan Kaufmann, San Francisco, pp 553–564

Baader F, Hollunder B (1995) Embedding defaults into terminological knowledge representation formalisms. J Autom Reason 14(1):149–180

Balbiani P, van Ditmarsch H, Herzig A, de Lima T (2012) Some truths are best left unsaid. In: Ghilardi S, Moss L (eds) Advances in modal logic (AiML), pp 1–15. Copenhagen, 22/08/2012–25/08/2012, College Publications. http://www.collegepublications.co.uk, www.irit.fr/~Andreas.Herzig/P/Aiml12.html

Balduccini M, Janhunen T (eds) (2017) Logic programming and nonmonotonic reasoning-14th international conference, LPNMR 2017, Espoo, Finland, 3-6 July 2017, proceedings. Lecture notes in computer science, vol 10377. Springer, Berlin. https://doi.org/10.1007/978-3-319-61660-5

Bell J (1990) The logic of nonmonotonicity. Artif Intell J 41:365–374

Benferhat S, Bonnefon JF, Neves RDS (2004) An experimental analysis of possibilistic default reasoning. In: Dubois D, Welty CA, Williams MA (eds) KR. AAAI Press, pp 130–140

Besnard P (1989) An introduction to default logic. Springer, Berlin

Booth R, Casini G, Varzinczak IJ (eds) (2017) Proceedings of the 4th international workshop on defeasible and ampliative reasoning (DARe-17) co-located with the 14th international conference on logic programming and nonmonotonic reasoning (LPNMR 2017), Espoo, Finland, 3 July 2017. CEUR workshop proceedings, vol 1872, CEUR-WS.org. http://ceur-ws.org/Vol-1872

Bradley R, Stefánsson HO (2017) Counterfactual desirability. Br J Philos Sci 68(2):485–533. https://doi.org/10.1093/bjps/axv023
Brewka G (1991) Cumulative default logic: in defense of nonmonotonic inference rules. Artif Intell J 50:183–205
Fariñas del Cerro L, Herzig A (1991) A modal analysis of possibility theory. In: Proceedings of the European conference on symbolic and quantitative approaches to uncertainty (ECSQAU'91). LNCS, vol 548. Springer, Berlin, pp 58–62. http://www.irit.fr/PERSONNEL/LILaC/Herzig/P/Ecsquau91.pdf (short version; long version published in FAIR'91)
Chellas BF (1975) Basic conditional logics. J Philos Log 4:133–153
Chellas BF (1980) Modal logic: an introduction. Cambridge University, Cambridge
Crocco G, Lamarre P (1992) On the connection between conditional logics and nonmonotonic logics. In: Nebel B, Rich C, Swartout W (eds) Proceedings of the 4th international conference on knowledge representation and reasoning (KR'92). Morgan Kaufmann, San Francisco, pp 565–571
Crocco G, Fariñas del Cerro L, Herzig A (1995) Conditionals: from philosophy to computer science. Studies in logic and computation. Oxford University, USA
Cross CB (2016) Embedded counterfactuals and possible worlds semantics. Philos Stud 173(3):665–673
Delgrande J, Schaub T, Jackson W (1994) Alternative approaches to default logic. Artif Intell J 70(1–2):167–237
Doherty P, Łukaszewicz W (1992) Defaults as first-class citizens. In: 22nd international symposium on multiple-valued logic (SMVL'92). IEEE computer society, Sendai, Japan, pp 146–154
Donini FM, Nardi D, Rosati R (2002) Description logics of minimal knowledge and negation as failure. ACM Trans Comput Log 3(2):177–225
Douven I (2016) Experimental approaches to the study of conditionals. A companion to experimental philosophy. Wiley, New Jersey, pp 545–554
Etherington D (1987) Formalizing nonmonotonic reasoning systems. Artif Intell J 31(1):41–48
Gabbay DM (1985) Theoretical foundations for non-monotonic reasoning in expert systems. In: Apt KR (ed) Logics and models of concurrent systems. Springer, Berlin, pp 439–457
Gärdenfors P, Makinson D (1994) Nonmonotonic inference based on expectation ordering. Artif Intell J 65:197–245
Gerbrandy J (1999) Bisimulations on planet Kripke. PhD thesis, University of Amsterdam
Gerbrandy J, Groeneveld W (1997) Reasoning about information change. J Log Lang Inf 6(2)
Giordano L, Gliozzi V, Olivetti N, Pozzato GL (2013) A non-monotonic description logic for reasoning about typicality. Artif Intell 195:165–202. https://doi.org/10.1016/j.artint.2012.10.004, http://www.sciencedirect.com/science/article/pii/S0004370212001269
Girlando M, Lellmann B, Olivetti N, Pozzato GL (2016) Standard sequent calculi for Lewis' logics of counterfactuals. In: Michael L, Kakas AC (eds) Logics in artificial intelligence - 15th European conference, JELIA 2016, Larnaca, Cyprus, 9–11 November 2016, proceedings. Lecture notes in computer science, vol 10021, pp 272–287. https://doi.org/10.1007/978-3-319-48758-8_18
Goodman N (1947) The problem of counterfactual conditionals. J Philos 44:113–128
Herzig A (1998) Logics for belief base updating. In: Dubois D, Gabbay D, Prade H, Smets P (eds) Handbook of defeasible reasoning and uncertainty management, vol 3 - Belief change. Kluwer, Dordrecht, pp 189–231
Hintikka J (1962) Knowledge and belief. Cornell University Press, Ithaca
Holliday WH, Hoshi T, Icard III TF (2011) Schematic validity in dynamic epistemic logic: decidability. In: Proceedings of the third international conference on logic, rationality, and interaction. Springer, Berlin, LORI'11, pp 87–96. http://dl.acm.org/citation.cfm?id=2050423.2050429
Kern-Isberner G, Wassermann R (2016) Proceedings of the 16th international workshop on nonmonotonic reasoning (NMR 2016) Cape Town, South Africa, 22–24 April 2016. Technische Universität, Department of Computer Science, Fakultät für Informatik (fi)
Konolige K (1995) Autoepistemic logic. In: Gabbay DM, Hogger CJ, Robinson JA (eds) Handbook of logic in artificial intelligence and logic programming, vol 3 (Nonmonotonic reasoning and uncertain reasoning). Oxford Science Publications, pp 217–295

Kooi B (2007) Expressivity and completeness for public update logic via reduction axioms. J Appl Non Class Log 17(2):231–253
Koutras CD, Rantsoudis C (2017) In all but finitely many possible worlds: model-theoretic investigations on overwhelming majority default conditionals. J Log Lang Inf 26(2):109–141. https://doi.org/10.1007/s10849-017-9251-5
Kraus S, Lehmann D, Magidor M (1990) Nonmonotonic reasoning, preferential models and cumulative logics. Artif Intell J 44:167–207
Kripke S (1963) Semantical analysis of modal logic. Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 9:67–96
Laverny N, Lang J (2005) From knowledge-based programs to graded belief-based programs, part II: off-line reasoning. In: Proceedings of the 9th international joint conference on artificial intelligence (IJCAI'05). Edinburgh, Gallus, pp 497–502
Sombé L (1994) Revision and updating in knowledge bases. Int J Intell Syst 9(1):1–180 (Besnard P, Cholvy L, Cordier MO, Dubois D, Fariñas del Cerro L, Froidevaux C, Lévy F, Moinard Y, Prade H, Schwind C, Siegel P)
Lehmann D, Magidor M (1992) What does a conditional knowledge base entail? Artif Intell J 55:1–60
Levi I (1996) For the sake of the argument: Ramsey test conditionals, inductive inference, and nonmonotonic reasoning. Cambridge University, Cambridge
Lewis C, Langford CH (1959) Symbolic logic. Dover reprint, original Published (1932)
Lewis D (1973) Counterfactuals. Basil Blackwell, Oxford
Lewis D (1979) Counterfactual dependence and time's arrow. Nous 13:455–476
Lifschitz V (2008) What is answer set programming? In: Fox D, Gomes CP (eds) AAAI. AAAI, pp 1594–1597
Lindström S, Rabinowicz W (1995) The Ramsey test revisited. In: Crocco G, Fariñas del Cerro L, Herzig A (eds) Conditionals: from philosophy to computer science. Studies in logic and computation, vol 5. Oxford University, Oxford, pp 147–192
Makinson D (1993) Five faces of minimality. Stud Log 52:339–379
Makinson D (1994) General patterns in nonmonotonic reasoning. In: Gabbay DM, Hogger CJ, Robinson JA (eds) Handbook of logic in artificial intelligence and logic programming, vol 3. Oxford University, Oxford, pp 35–110
Marek V, Truszczynski M (1989) Relating autoepistemic and default logics. In: Brachman RJ, Levesque HJ, Reiter R (eds) 1st international conference on principles of knowledge representation and reasoning (KR'89). Morgan Kaufmann, Toronto, Canada, pp 276–288
McCarthy J (1980) Circumscription, a form of nonmonotonic reasoning. Artif Intell J 13:27–39
McCarthy J (1986) Applications of circumscription to formalizing common-sense knowledge. Artif Intell J 28(1):1038–1044
McCarthy J (1990) Formalizing common sense: papers. Ablex, Norwood
McDermott D, Doyle J (1980) Non-monotonic logic I. Artif Intell J 13:41–72. Special issue in non-monotonic reasoning
Moore RC (1985) Semantical considerations on nonmonotonic logic. Artif Intell 25(1)
Neves RDS, Bonnefon JF, Raufaste E (2002) An empirical test of patterns for nonmonotonic inference. Ann Math Artif Intell 34(1–3):107–130
Nute D (1980) Topics in conditional logic. D. Reidel, Dordrecht
Nute D (1984) Conditional logic. In: Gabbay DM, Günthner F (eds) Handbook of philosophical logic, vol 2. D. Reidel, Dordrecht, pp 387–439
Pfeifer N (2014) Reasoning about uncertain conditionals. Stud Log 102(4):849–866
Plaza JA (1989) Logics of public communications. In: Emrich ML, Pfeifer MZ, Hadzikadic M, Ras ZW (eds) Proceedings of the 4th international symposium on methodologies for intelligent systems, pp 201–216. Reprinted in Synthese 158:2, pp 165–179 (2007)
Reiter R (1980) A logic for default reasoning. Artif Intell J 13:81–132. Special issue in non-monotonic reasoning

Ryan M, Schobbens PY (1997) Intertranslating counterfactuals and updates. J Log Lang Inf 6(2):123–146 (Preliminary version in: W. Wahlster (ed.) Proceedings of the ECAI'96)

Stalnaker R (1968) A theory of conditionals. In: studies in logical theory, American philosophical quarterly (Monograph Series, No. 2), Blackwell, Oxford, pp 98–112 (reprinted in Sosa E (ed) Causation and conditionals. Oxford University, Oxford 1975; reprinted in Harper WL, Stalnaker R, Pearce G (eds) Ifs. Reidel, Dordrecht 1981; reprinted in Harper WL, Skyrms B (eds) Causation in decision, belief change and statistics, vol 2. Reidel, Dordrecht 1988, pp 105–134; reprinted in Jackson F (ed) Conditionals. Oxford University, Oxford Readings in Philosophy 1991)

Stalnaker R (1992) What is a nonmonotonic consequence relation? In: (Informal) Working notes of the 4th international workshop on nonmonotonic reasoning. Plymouth, Vermont

Troelstra AS (1992) Lectures on linear logic. Lecture notes, vol 29, CSLI (Center for the Study of Language and Information), Stanford

van Benthem J (2006) One is a lonely number: on the logic of communication. In: Chatzidakis Z, Koepke P, Pohlers W (eds) Logic colloquium'02. ASL and Peters AK, Wellesley MA, pp 96–129, Technical report PP-2002-27, ILLC Amsterdam (2002)

van Ditmarsch HP (2000) Knowledge games. PhD thesis, Groningen University, ILLC dissertation series, Grafimedia

Representations of Uncertainty in Artificial Intelligence: Probability and Possibility

Thierry Denœux, Didier Dubois and Henri Prade

Abstract Due to its major focus on knowledge representation and reasoning, artificial intelligence was bound to deal with various frameworks for the handling of uncertainty: probability theory, but more recent approaches as well: possibility theory, evidence theory, and imprecise probabilities. The aim of this chapter is to provide an introductive survey that lays bare specific features of two basic frameworks for representing uncertainty: probability theory and possibility theory, while highlighting the main issues that the task of representing uncertainty is faced with. This purpose also provides the opportunity to position related topics, such as rough sets and fuzzy sets, respectively motivated by the need to account for the granularity of representations as induced by the choice of a language, and the gradual nature of natural language predicates. Moreover, this overview includes concise presentations of yet other theoretical representation frameworks such as formal concept analysis, conditional events and ranking functions, and also possibilistic logic, in connection with the uncertainty frameworks addressed here. The next chapter in this volume is devoted to more complex frameworks: belief functions and imprecise probabilities.

1 Introduction

The question of including, hence modeling, uncertainty in scientific matters is not specific to the field of artificial intelligence. Historically, this concern already appears in the XVIIth century, with pioneering works of Huyghens, Pascal, chevalier de Méré, and Jacques Bernoulli. There existed at that time a major distinction between

T. Denœux (✉)
Université de Technologie de Compiègne, CNRS,
Heudiasyc (UMR 7253), Paris, France
e-mail: tdenoeux@utc.fr

D. Dubois · H. Prade
IRIT, CNRS and Université Paul Sabatier, Toulouse, France
e-mail: dubois@irit.fr

H. Prade
e-mail: prade@irit.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_3

the objective notion of *chance* in connection with the study of games (of chance), and the subjective notion of *probability* in connection with the issue of unreliable testimonies at courts of law. With J. Bernoulli, chances are related to frequencies of events and are naturally additive, while subjective probabilities are not supposed to be so. This view is still present in the middle of the XVIIIth century with the works of Lambert. He proposed a combination rule which turns out to be a special case of Dempster's rule of combination; see (Shafer 1978), (Martin 2006), and chapter "Representations of Uncertainty in AI: Beyond Probability and Possibility" of this volume. However, with the rapid development of physics and actuarial sciences later on, the interest for the non-additive side of probability eventually waned and the issue was forgotten for almost two centuries, while the additive view became prominent, with the works of Laplace, whether the focus was on frequentist probability or not. Noticeably, in the middle of the XXth century, in economics, not only the main approach to decision under (frequentist) risk after (von Neumann and Morgenstern 1944), but also the mainstream theory of decision under (subjective) uncertainty relied on additive probability.

It is the emergence of computer sciences that brought issues related to knowledge representation and reasoning in the presence of imprecision, uncertainty, and conflicting information to the front. This went on till the 1980's almost independently of probability theory and the issue of decision-making. Instead, artificial intelligence first put the emphasis on logical and qualitative formalisms, as well as the modeling of linguistic information (in trends of research such as fuzzy set theory).

Indeed, the available information to be stored in a computer is often unreliable, as is human knowledge, so that reasoning is based on rules that may lead to uncertain conclusions even starting from sure premises. The need to handle uncertainty arose in fact with the emergence of the first expert systems at the beginning of the 1970's. One of the first and best known expert rule-based system, namely MYCIN (Shortliffe 1976; Buchanam and Shortliffe eds.1984), already proposed an ad hoc, entirely original, technique for uncertainty propagation based on degrees of belief and disbelief. This method will not be described here for lack of space, and because it is now totally outdated, especially due to its improper handling of exceptions in if-then rules. But the uncertainty propagation technique of MYCIN pioneered the new, more rigorous frameworks for uncertainty modeling that would appear soon after. On this point, see (Dubois and Prade 1989), and (Lucas and van der Gaag 1991) as well.

This chapter is structured in four sections. In Sect. 2, basic notions useful for describing the imperfection of information are defined and discussed. Section 3 deals with probability theory, focusing on the possible meanings of probability and the difficulty to handle plain incomplete information with probability distributions, as well as the connections between conditioning and logic. Section 4 deals with set functions extending the modalities of possibility and necessity, distinguishing between qualitative and quantitative approaches, and describing connections with reasoning tolerant to exceptions, formal concept analysis, probability and statistics. Section 5 explains the links between uncertain reasoning and Aristotelian logic, generalizing the square of opposition.

2 Imprecision, Contradiction, Uncertainty, Gradualness, and Granularity

Before presenting various representation frameworks (see (Halpern 2003), (Dubois and Prade 2009), (Liu 2001), (Parsons 2001) for interesting focused overviews), it is useful to somewhat clarify the terminology. We call *information item* any collection of symbols or signs produced by observing natural or artificial phenomena, or by human cognitive activity, whose purpose is communication. Several distinctions are in order. First, one must separate so-called *objective* information items, coming from sensor measurements or direct observations of the world, from *subjective* ones, expressed by individuals and possibly generated without using direct observations of the outside world. Information items may be couched in numerical formats, especially objective ones (sensor measurements, counting processes), or in qualitative or symbolic formats (especially subjective ones, in natural language for instance). However the dichotomy objective numerical vs. subjective qualitative is not so clearcut. A subjective information item can be numerical, and objective observations can be qualitative (like a color perceived by a symbolic sensor, for instance). Numerical information can take various forms: integers, real numbers, intervals, real-valued functions, etc. Symbolic information is often structured and encoded in logical or graphical representations. There are also hybrid representation formats, like Bayesian networks (Pearl 1988). Finally, another important distinction should be made between *singular* and *generic* information. Singular information refers to particular facts and results from an observation or a testimony. Generic information pertains to a *class* of situations and expresses knowledge about it: it can be a law of physics, a statistical model stemming from a representative sample of observations, or yet commonsense statements such as "birds fly" (in this latter case the underlying class of situations is not precise: is it here a zoological definition, or the birds of any epoch, or of any place, etc.?).

2.1 Imprecise Information

To represent the epistemic state of an agent, one must beforehand possess a language for representing the states of the world under interest, according to the agent, that is, model relevant aspects by means of suitable attributes. Let v be a vector of attribute variables[1] relevant for the agent, and let S be its domain (possibly not described in extension). S is then the set of (precise descriptions) of the set of possible states of affairs. A subset A of S is viewed as an event, or as a proposition that asserts $v \in A$.

An information item $v \in A$ possessed by an agent is said to be *imprecise* if it is not sufficient to enable the agent to answer a question of interest about v. Imprecision

[1] In fact, in this chapter, v denotes an ill-known entity that may be for instance a random variable in a probabilistic setting, or rather an imprecisely known entity but which does not vary strictly speaking.

corresponds to the idea of *incomplete* or even missing information. The question to which the agent tries to answer is of the form what is the value of v, or more generally does v satisfy a certain property B, given that $v \in A$ is known? The notion of imprecision is not absolute. When concerned with the age of a person, the term *minor* is precise if the referential set is $S = \{minor, major\}$ and the question is: has this person the right of vote? In contrast if the question is to determine the age of this person and $S = \{0, 1, \ldots, 150\}$ (in years), the term *minor* is very imprecise.

The standard format of an imprecise information item is $v \in A$ where A is a *subset* of S *containing more that one element*. An important remark is that elements of A, seen as possible values of v are mutually exclusive (since the entity v possesses only one value). So, an imprecise information item takes the form of a disjunction of mutually exclusive values. For instance, to say that John is between 20 and 22 years old, that is, $v = age(John) \in \{20, 21, 22\}$ means to assume that $v = 20$ or $v = 21$ or $v = 22$. An extreme form of imprecise information is *total ignorance*: the value of v is completely unknown. In classical logic, imprecision explicitly takes the form of a disjunction (stating that $A \vee B$ is true is less precise than stating that A is true). The set A representing an information item is called an *epistemic set*.

Two imprecise information items can be compared in terms of informational content: an information item $v \in A_1$ is said to be *more specific* than another information item $v \in A_2$ if and only if A_1 is a proper subset of A_2.

The disjunctive view of sets used to represent imprecision contrasts with the more usual conjunctive view of a set as a collection of items forming a certain complex entity. It then represents a precise information item. For instance, consider the set of languages that John can speak, say $v = Lang(John)$. This variable is set-valued and stating that $Lang(John) = \{\text{English, French}\}$ is a precise information item, as it means that John can speak English and French only. In contrast, the variable $v' = NL(John)$ representing the native language of John is single-valued and the statement $NL(John) \in \{\text{English, French}\}$ is imprecise. The domain of v' is the set of all spoken languages while the domain of v is its power set. In the latter case, an imprecise information item pertaining to a set-valued variable is represented by a (disjunctive) set of (conjunctive) subsets.

2.2 Contradictory Information

An information item is said to be contradictory if it is of the form $v \in A$, where $A = \emptyset$. Under this form there is not much we can do with such an information item. In mathematics, the presence of a contradiction ruins any form of reasoning, and it is only used to prove claims by refutation (a claim is true because assuming its falsity leads to a contradiction). In artificial intelligence, contradiction often stems from the conflict between several information items, e.g., $v \in A$ and $v \in B$ where $A \cap B = \emptyset$. It is thus a natural situation that is to be expected each time there are several sources, and more generally if collected information items are numerous. Another cause of conflicting information is the presence of exceptions in generic

information items such as rules, which may lead to simultaneously infer opposite conclusions. There are several approaches in the literature that aim at coping with contradictory information, and that are studied in this book:

- information fusion techniques that aim at restoring consistency, by deleting unreliable information items, taking into account the sources that deliver them, and analyzing the structure of the conflict between them. See chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" in this volume.
- argumentation methods that discuss the pros and the cons of deriving a proposition $v \in A$ using a graph-theoretic representation of an attack relation between conflicting arguments. See chapter "Argumentation and Inconsistency-tolerant Reasoning" in this volume.
- paraconsistent logics that try to prevent the infection of the contradiction affecting some variables or some subgroups of information items to other ones, by for instance changing the inference relation, thus avoiding the explosive nature of standard inference from inconsistent bases in classical logic. See chapter "Argumentation and Inconsistency-tolerant Reasoning" in this volume.
- nonmonotonic reasoning formalisms that try to cope with exceptions in rules by giving priority to conclusions of the most specific ones. See chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume.

2.3 Uncertain Information

An information item is said to be uncertain for an agent if the latter does not know whether it is true or false. If an elementary information item of the form of a proposition $v \in A$, where A contains a set of non-impossible values for v, is tainted with uncertainty, a token of uncertainty is attached to it. This token is a qualifier situated at the meta-level with respect to the information item. It can be numerical or symbolic: compare statements expressing uncertainty such as *The task will take at least one hour, with probability 0.7*, and *It is not fully sure that John comes to the meeting*. Uncertainty has two main origins: the lack of information, or the presence of conflicting information. A special case of the latter is aleatory uncertainty, where due to the variability of an observed phenomenon, it is difficult to predict the next event, hence the information item $v \in A$ that may describe it.

The most usual representation of uncertainty consists in assigning to each proposition $v \in A$ or event $A \subseteq S$, a number $g(A)$ in the unit interval. This number expresses the agent's confidence in the truth of the proposition $v \in A$. Note that this proposition is ultimately only true or false, but the agent may currently ignore what its actual truth-value is. Natural conditions are required for the set function g:

$$g(\emptyset) = 0; \quad g(S) = 1; \quad \text{if } A \subseteq B \text{ then } g(A) \leq g(B). \tag{1}$$

Indeed the contradictory proposition $v \in \emptyset$ is impossible, and the tautological proposition $v \in S$ is certainly true. Moreover, if A is more specific than B (and thus implies B), a rational agent cannot trust $v \in A$ more than $v \in B$. When S is infinite, one must add suitable continuity properties with respect to monotonic sequences of subsets. Such a function g is often called a *capacity* (Choquet 1953), or *fuzzy measure* (Sugeno 1977), or yet plausibility function (Halpern 2001) (not to be confused with the dual to belief functions, defined in the next chapter in this volume). An important consequence of (1) is in the form of two inequalities:

$$g(A \cap B) \leq \min(g(A), g(B)); \quad g(A \cup B) \geq \max(g(A), g(B)). \tag{2}$$

These inequalities suggest to consider extreme confidence measures g such that one of these inequalities is an equality, and more generally, when A and B are mutually exclusive, assume that $g(A \cup B)$ only depends on $g(A)$ and $g(B)$ (Dubois and Prade 1982), i.e.,

$$\text{if } A \cap B = \emptyset \text{ then } g(A \cup B) = g(A) \oplus g(B). \tag{3}$$

for some binary operation $\oplus$ on $[0, 1]$.

The *conjugate* set function, defined by $\overline{g}(A) = 1 - g(\overline{A})$, then satisfies the dual property $\overline{g}(A \cap B) = \overline{g}(A) \perp \overline{g}(B)$ if $A \cup B = S$ where $a \perp b = 1 - (1 - a) \oplus (1 - b)$ (Dubois and Prade 1982). The set functions g and $\overline{g}$ are said to be *decomposable*. Compatibility constraints with the Boolean algebra of events suggests considering operations $\oplus$ and $\perp$ that are associative, which leads to choose $\perp$ and $\oplus$ among triangular *norms* and *co-norms* (Klement et al. 2000) (they get their name from their role in the expression of the triangular inequality in stochastic geometry (Schweizer and Sklar 1963)). The main possible choices for $a \perp b$ (resp. $a \oplus b$) are the operators minimum $\min(a, b)$, product $(a \times b)$, and truncated addition $\max(0, a + b - 1)$ (resp. maximum $\max(a, b)$, probabilistic sum $a + b - a \times b$, and bounded sum $\min(1, a + b)$). Probability measures are recovered by defining $a \oplus b = \min(1, a + b)$ (equivalently $a \perp b = \max(0, a + b - 1)$), and *possibility* measures and *necessity* respectively for $a \oplus b = \max(a, b)$ and for $a \perp b = \min(a, b)$. The use of more complex operators (like ordinal sums of the above ones) may make sense (Dubois et al. 2000b).

2.4 Graduality and Fuzzy Sets

Representing a proposition in the form of a statement that can only be true or false (or an event that occurs or not) is but a convenient convention. It is not always an ideal one. Some information items are not easily amenable to respecting this convention. This is especially the case for statements involving *gradual* properties, like in the proposition *John is young*, that may sometimes be neither completely true nor completely false: it is clearly more true if John is 20 than if he is 30, even if in the latter case, John is still young to some extent. Predicates like *young* can be

modified by linguistic hedges. It makes sense to say *very young, not so young*, etc. Such linguistic hedges cannot be applied to Boolean predicates, like *single*. In other words, the proposition *John is young* is not Boolean, which denotes the presence of an ordering between age values to which it refers. This type of information can be taken into account by means of *fuzzy sets* (Zadeh 1965). A fuzzy set F is a mapping from S to a totally ordered set L often chosen to be the unit interval [0, 1]. The value $F(s)$ is the membership degree of the element s in F. It evaluates the compatibility between the situation s and the predicate F.

Fuzzy sets are useful to deal with information items in natural language referring to a clear numerical attribute. Zadeh (1975) introduced the notion of *linguistic variable* with values in a linearly ordered linguistic term set. Each of these terms represents a subset of the numerical domain of the attribute, and these subsets correspond to a partition of this domain. For instance, the set of terms $T = \{young, adult, old\}$ forms the domain of the linguistic variable *age(John)* and partitions the domain of this attribute. Nevertheless it is not surprising to admit that the transitions between the ranges covered by the linguistic terms are gradual rather than abrupt. And in this situation, it sounds counterintuitive to set precise thresholds separating these continuous ranges. Namely, it sounds absurd to define the set $F = young \in T$ by a precise threshold $s_\star$ such that $F(s) = 0$ if $s > s_\star$ and $F(s) = 1$ otherwise, beyond which an individual suddenly ceases to be young. The membership function of the fuzzy set valued in the scale [0, 1], representing here the gradual property *young*, is but a direct reflection of the continuous domain of the attribute (here the age). This also leads to the idea of a fuzzy partition made of non-empty fuzzy subsets $F_1, \cdots, F_n$, often defined by the constraint $\forall s$, $\Sigma_{i=1,n} F_i(s) = 1$ (Ruspini 1970).

If we admit that some sets are fuzzy and membership to them is a matter of degree, one issue is to extend the set-theoretical operations of union, intersection and complementation to fuzzy sets. This can be done in a natural way, letting

$$(F \cup G)(s) = F(s) \oplus G(s); \quad (F \cap G)(s) = F(s) \perp G(s); \quad \overline{F}(s) = 1 - F(s),$$

where $\oplus$ and $\perp$ are triangular co-norms and norms already encountered in the previous subsection. The choice $\oplus = \max$ and $\perp = \min$ is the most common. With such connectives, the De Morgan property between $\cup$ and $\cap$ are preserved, as well as their idempotence and their mutual distributivity. However, the excluded middle ($A \cup \overline{A} = S$) and contradiction laws ($A \cap \overline{A} = \emptyset$) fail. Choosing $\oplus = \min(1, \cdot + \cdot)$ and $\perp = \max(0, \cdot + \cdot - 1)$ re-installs these two laws, at the cost of losing idempotence and mutual distributivity of $\cup$ and $\cap$. As to fuzzy set inclusion, it is oftentimes defined by the condition $F \subseteq G \Leftrightarrow \forall s$, $F(s) \leq G(s)$. A more drastic notion of inclusion requires the inclusion of the support of F (elements s such that $F(s) > 0$) in the core of G (elements s such that $G(s) = 1$). In agreement with the spirit of fuzzy sets, inclusion can also be a matter of degree. There are various forms of inclusion indices, of the form $d(F \subseteq G) = \min_s F(s) \rightarrow G(s)$, where $\rightarrow$ is a many-valued implication connective.

Fuzzy sets led to a theory of approximate reasoning and the reader is referred to a section dedicated to interpolation in chapter "Case-Based Reasoning, Analogical

Reasoning, and Interpolation" of this volume. Besides, since the mid-1990's, there has been a considerable development of formal fuzzy logics, understood as syntactic logical systems whose semantics is in terms of fuzzy sets. These works, triggered by the book by Hájek (1998), considerably improved the state of the art in many-valued logics developed in the first half of the XXth century (see (Dubois et al. 2007)) for a detailed survey of both approximate reasoning and formal fuzzy logic.)

2.5 *Degree of Truth Versus Degree of Certainty: A Dangerous Confusion*

It is very crucial to see the difference between the degree of adequacy between a state of affairs and an information item (often called *degree of truth*) and a degree of certainty (confidence). Already, in natural language, sentences like *John is very young* and *John is probably young* do not mean the same. The first sentence expresses the fact that the degree of membership of *age(John)* (e.g., *age(John)* = 22) to the fuzzy set of young ages is for sure high. The degree of membership $F(s)$ evaluates the degree of adequacy between a state of affairs s_0, e.g., $s_0 = 22$, and the fuzzy category $F =$ *young*. According to the second sentence, it is not ruled out that John is not young at all.

Degrees of truth and degrees of certainty correspond to distinct notions that occur in distinct situations with unrelated semantic contents. Moreover they are driven by mathematical frameworks that should not be confused despite their superficial resemblances as to the involved connectives. Indeed, as seen earlier in this text, truth degrees are usually assumed to be compositional with respect to all connectives like conjunction, disjunction, and negation (respectively corresponding to intersection, union, and complementation of fuzzy sets). However, full-fledged compositionality is impossible for degrees of certainty. This is because the Boolean algebra of standard events is not compatible with the structure of the unit interval, nor any finite totally ordered set with more than 2 elements (Dubois and Prade 2001): they are not Boolean algebras. For instance, probability is compositional only for negation ($Prob(\overline{A}) = 1 - Prob(A)$), and as we shall see later on, possibility (resp. necessity) measures are compositional only for disjonction (resp. conjonction). For instance one can be sure that $v \in A \cup B$ is true (especially if $B = \overline{A}$!), without being sure at all that any of $v \in A$, and $v \in B$ is true.

A typical situation where certainty and truth tend to be confused is when using a three-valued logic to capture partial information, changing Boolean interpretations of a language into three-valued ones. The usual truth set $\{0, 1\}$ is turned into, say, $\{0, 1/2, 1\}$, with the idea that $1/2$ stands for *unknown* as in Kleene logic (Kleene 1952). Now the problem is that under the proposed calculus by Kleene with conjunction, disjunction and negation expressed by operations min, max and $1 - (\cdot)$, respectively, the excluded middle law is lost. This is a paradox here as, since a proposition $v \in A$ can only be true or false, the composite proposition $v \in A$ *or*

$v \notin A$ is always valid while, in the three-valued setting, it will have truth value 1/2 if $v \in A$ is set to *unknown*. The way out of the paradox consists in noticing that the negation of *unknown* is *known*, actually known to be true or known to be false. So the three alleged truth-values $\{0, 1/2, 1\}$ are degrees of certainty, and actually stand for the three non-empty subsets of $\{0, 1\}$, 1/2 standing for the hesitation between true and false, namely $\{0, 1\}$. And then it becomes clear the statement *either* $v \in A$ *is known to be true* or $v \in A$ *is known to be false* is not a tautology.

The Kleene approach to ignorance has been extended by Belnap (1977a; 1977b) to include contradictory information stemming from conflicting sources, adding a fourth truth value expressing contradiction. The 4-valued truth set forms a bilattice structure and is isomorphic to the four subsets of $\{0, 1\}$ (now including $\emptyset$), equipped with two partial orderings: the truth-ordering (where the two new truth-values are incomparable and lie between true and false) and the information ordering (that coincides with inclusion of subsets in $\{0, 1\}$). These "epistemic truth-values" are attached to atomic propositions, and truth-tables in agreement with the bilattice structure enable the epistemic status of complex propositions to be computed. The same kind of analysis as above applies regarding the use of compositional truth values in this logic (e.g., *true* in the sense of Belnap means approved by some source and disapproved by none, an epistemic stance). See Dubois (2012) for a discussion. Besides, Ginsberg (1990) used Belnap bilattices to propose a unified semantic view for various forms of non-monotonic inferences (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume and the Sect. 4.2.1 in this chapter).

2.6 Granularity and Rough Sets

In the preceding sections, we did not question the assumptions that underlie the definition of the set S of states of the world. It should not be taken for granted, as it presupposes the definition of a language. The logical approach to Artificial Intelligence often starts from a set of statements expressed in a propositional language, to which it may assign degrees of confidence. Then the set S is the set of states or interpretations generated by these propositions (mathematically, the subsets of S are the smallest Boolean algebra supporting these propositions). This view has important consequences for the representation and the updating of bodies of information items. For instance, a new information item may lead to a refinement of S: this is called a change of granularity of the representation.

The simplest case of change of granularity is when the basic propositions are taken as atomic ones, or more generally when describing objects by attributes. Let Ω be a set of objects described by attributes $v_1, v_2, \ldots, v_k$ with respective domains $D_1, D_2, \ldots, D_k$. Then S is the Cartesian product $D_1 \times D_2 \times \cdots \times D_k$. Each element of S can be refined into several ones if a $(k+1)$th attribute is added. Clearly, nothing prevents distinct objects from having the same description in terms of such attributes. Then they are indiscernible by means of this language of description.

Consider a subset Θ of objects in Ω. It follows from the above remark, that in general Θ cannot be described precisely using such an attribute-based language. Indeed let R be an equivalence relation on Ω clustering objects having the same description: $\omega_1 R \omega_2$ if and only if $v_i(\omega_1) = v_i(\omega_2), \forall i = 1, \ldots, k$. Let $[\omega]_R$ be the equivalence class of object ω. We only have such equivalence classes to describe the set Θ, so only approximate descriptions of it can be used. The only thing we can do is to build upper and lower approximations Θ^* and Θ_* defined as follows:

$$\Theta^* = \{\omega \in \Omega : [\omega]_R \cap \Theta \neq \emptyset\}; \quad \Theta_* = \{\omega \in \Omega : [\omega]_R \subseteq \Theta\} \tag{4}$$

The pair (Θ^*, Θ_*) is called a *rough set* (Pawlak 1991; Pawlak and Skowron 2007). Only subsets of objects such as Θ^* and Θ_* can be accurately described by means of combinations of attribute values of $v_1, v_2, \ldots, v_k$.

There are various examples of situations where rough sets implicitly appear, for instance histograms or digital images correspond to the same notions of indiscernability and granularity, where equivalence classes correspond, respectively, to the bins of the histograms and to pixels.

The concept of rough set is thus related to the ones of indiscernibility and granularity, while the concept of fuzzy set is related to gradualness. It it is possible to build concepts where these two dimensions are at work, when the set to be approximated or the equivalence relation become fuzzy (Dubois and Prade 1992). Rough sets are also useful in machine learning to extract rules from incomplete data (Grzymala-Busse 1988; Hong et al. 2002), as well as fuzzy decision rules (Greco et al. 2006) (see chapter "Designing Algorithms for Machine Learning and Data Mining" in Volume 2).

3 Uncertainty: The Probabilistic Framework

Probability theory is the oldest uncertainty theory and, as such, the best developed mathematically. Probability theory can be envisaged as a chapter of mathematics. In that case, we consider a probability space, made of a set Ω (called a sample space) and an application v from Ω to S (called random variable), where oftentimes S is taken as the real line. In the simplest case S is a finite set which determines via v a finite partition of Ω. If $\mathscr{B}$ is the Boolean algebra generated by this partition, a probability space is actually the triple $(\Omega, \mathscr{B}, \mathscr{P})$, and P is a probability measure, i.e., an application from $\mathscr{B}$ to [0, 1] such that:

$$P(\emptyset) = 0; \quad P(\Omega) = 1; \tag{5}$$

$$\text{if } A \cap B = \emptyset \text{ then } P(A \cup B) = P(A) + P(B). \tag{6}$$

Elements of $\mathscr{B}$ are called measurable subsets of Ω. The probability distribution induced by v on S is then characterized by a set of weights $p_1, p_2, \ldots, p_{card(S)}$,

defined by $p_i = P(v^{-1}(s_i))$, and such that

$$\sum_{i=1}^{card(S)} p_i = 1.$$

Probabilities of events can be extended to fuzzy events by considering the expectation of their membership functions (Zadeh 1968), which indeed generalizes the usual expression $P(A) = \sum_{s_i \in A} p_i$ of a classical event.

Beyond the apparent[2] unity of the mathematical model of probability, there are strikingly different views of what probability means (Fine 1983). The aim of this section is to discuss some of these views, emphasizing some limitations of the representation of uncertainty by means of a unique probability distribution. This section is completed by a glance at De Finetti's conditional events and their three-valued logic, and at a very specific kind of probability distribution (so-called *big-stepped*) that play a noticeable role in the representation of default rules.

3.1 Frequentists Versus Subjectivists

If probability theory is considered as a tool for knowledge representation, one must explain what probability means, what is it supposed to represent. There are at least three understandings of probability, that have been proposed since its inception.

The simplest one is combinatorial. The set Ω is finite and p_i is proportional to the number of elements in $v^{-1}(s_i)$. Then a probability degree is just a matter of counting, for each event, the proportion of favorable cases over the number of possible ones. The well-foundedness of this approach relies on considerations about symmetry (a principle of indifference or insufficient reason, after Laplace), or the assumption that the phenomenon we deal with is genuinely random (like coin flipping, fair die tossing, etc.), and follows a uniform distribution.

The most common interpretation is frequentist. It is assumed that we accumulate observations (a finite n-element subset $\Omega(n)$ of the sample space Ω). Then frequencies of observing $v = s_i$ defined by:

$$f_i = \frac{card(v^{-1}(s_i) \cap \Omega(n))}{n}$$

can be obtained. When S is infinite, we can build a histogram associated to the random variable v by considering frequencies of elements of a finite partition of S (possibly adjusting a continuous distribution to it).

[2] Apparent, because the mathematical settings proposed by Kolmogorov and De Finetti (1974) are different, especially for the notion of conditioning, even if the Kolmogorov setting seems to be overwhelmingly adopted by mathematicians.

As the number of observations increases, $\Omega(n)$ becomes a representative sampling of Ω, and it is assumed that such frequencies f_i converge to probability values defined as limits, by $p_i = \lim_{n\to\infty} f_i$. To use this definition of probability, one must clearly have a sufficient number of observations available (ideally an infinite number) for the phenomenon under study. Under this view, the probability of a non-repeatable event makes no sense. Moreover, the frequentist probability distribution is a mathematical model of a physical phenomenon, hence objective, even if it can be part of the knowledge of an agent.

Under the third, subjectivist, view, the degree of probability $P(A)$ is interpreted as a degree of belief of an agent in the truth of the information item $v \in A$. Hence it should apply to any event, be it repeatable or not. What plays the role of frequencies for making subjective probability operational for non-repeatable events is the amount of money one should pay for a gamble on the occurrence or the non occurrence of event A. More precisely the degree of probability $P(A)$ for an agent is equated to the fair price this agent is willing to pay to a bookmaker for a lottery ticket with a 1 euro reward in case the event occurs. The price is fair in the sense that the agent would also agree to sell it at this price to the bookmaker, should the latter decide to buy it. Clearly the more the agent believes in A the greater (i.e., the closer to 1 euro) the price (s)he is likely to offer. This approach then relies on a rationality principle, called *coherence*, saying that the agent is not willing to lose money for sure. It ensures that degrees of belief (betting prices) behave in an additive way like probabilities. To see it, suppose the agent buys two lottery tickets, the first one to bet on A, the second one to bet on its complement $\overline{A}$. The agent is sure to have one winning ticket, which means a profit of $1 - P(A) - P(\overline{A})$ euros in relative value. Prices such that $P(A) + P(\overline{A}) - 1 > 0$ are not rational as it means a sure loss for the agent. However, prices such that $P(A) + P(\overline{A}) - 1 < 0$ are unfair and will lead the bookmaker to buy the tickets at those prices instead of selling them, to avoid sure loss on the bookmaker side. So the only choices left for the agent is to propose prices such that $P(A) + P(\overline{A}) = 1$. The same reasoning can be carried our for three mutually exclusive events, A, B, $\overline{A \cup B}$, leading to the constraint $P(A) + P(B) + P(\overline{A \cup B}) = 1$, which, since $P(\overline{A \cup B}) = 1 - P(A \cup B)$, leads to $P(A \cup B) = P(A) + P(B)$. Note that the probability degrees so-defined are personal, and may change across agents, contrary to frequentist probabilities.

Apparently, the subjectivist approach looks like a mere copy of the calculus of frequentist probabilities. In fact as shown by De Finetti (1974) and his followers (Coletti and Scozzafava 2002), things are not so simple. First, in the subjectivist approach there is no such thing as a sample space. The reason is that a subjective probability is either assigned to a unique event (after betting one checks whether this event did occur or not), or to a single realization of a repeatable one (e.g., flipping this coin now). Next, on infinite spaces, only finite additivity (in contrast with σ-additivity for the frequentist approach) can be justified by the above betting paradigm. Finally, the initial data does not consist of statistics, but a collection of bets (prices c_i) on on the truth of propositions A_i in an arbitrary set thereof $\{A_j : j = 1, \ldots, m\}$, along with a number of logical constraints between those propositions. The state space S is then constructed based on these propositions and these constraints. It is assumed, by

virtue of the coherence principle, that the agent assigns prices c_j to propositions A_j in agreement with the probability calculus, so that there is a probability distribution that satisfies $P(A_j) = c_j$, $j = 1, \ldots, m$. While the frequentist approach leads to assuming a unique probability distribution representing the random phenomenon (obtained via an estimation process from statistical data), this is not the case in the subjectivist setting, if the bets bear on arbitrary events. Indeed there may be several probability measures such that $c_j = P(A_j), \forall j = 1, \ldots, m$. Any of those probability functions is coherent but the available information may not allow us to select a single one. It may also occur that no such probability exists (then the bets are not coherent). To compute the probability degree $P(A)$ of some arbitrary event A based on a collection of pairs $\{(A_j, c_j) : j = 1, \ldots, m\}$, one must solve linear programming problems whose decision variables are probabilities p_i attached to singletons of S of the form: maximise (or minimise) $\sum_{s_i \in A} p_i$ under constraints $c_j = \sum_{s_k \in A_j} p_k, \forall j = 1, \ldots, m$.

It is then clear that the subjectivist approach to probability is an extension of the logical approach to artificial intelligence based on propositional logic and classical inference. The latter is recovered by assigning probability $c_j = 1$ to A_j, $j = 1, \ldots, m$, which enforces $P(A) = 1$ to all logical consequences A of $\{A_j : j = 1, \ldots, m\}$.

There are other formal differences between frequentist and subjectivist probabilities when it comes to conditioning.

3.2 Conditional Probabilities

By considering S as the state space, it is implicitly assumed that S represents an exhaustive set of possible worlds. To emphasize this point of view we may as well write the probability $P(A)$ as $P(A \mid S)$. If further on the agent receives new information that comes down to restraining the state space, probabilities will be defined based on a different context, i.e., a non-empty subset $C \neq \emptyset \subset S$ and the probability $P(A)$ becomes $P(A \mid C)$ in this new context. Changing $P(A)$ into $P(A \mid C)$ essentially consists in a renormalization step for probabilities of states inside C, setting other probabilities to 0:

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} \tag{7}$$

We can indeed check that $P(A) = P(A \mid S)$. This definition is easy to justify in the frequentist setting, since indeed $P(A \mid C)$ is but the limit of a relative frequency.

Justifying this definition in the subjectivist case is somewhat less straightforward. The probability $P(A \mid C)$ is then assigned to the occurrence of a conditional event denoted by $A \mid C$.[3] The quantity $P(A \mid C)$ is again equated to the fair price of a lottery ticket for the conditional bet on $A \mid C$. The difference with a standard bet is that if the

[3] We come back to the logic of conditional events at the end of this section.

opposite of C occurs, the bet is called off and the amount of money paid for the ticket is given back to the agent (De Finetti 1974). The conditional event $A \mid C$ represents the occurrence of event A in the hypothetical context where C would be true. In this operational set-up it can be shown that the identity $P(A \cap C) = P(A \mid C) \cdot P(C)$ holds.

This definition of conditional probability contrasts with the one of Kolmogorov based on a quotient, which presupposes $P(C) \neq 0$, and proves to be too restrictive in the subjectivist setting. Indeed, in the latter setting, conditional probabilities are directly collected, so that conditional probability is the primitive concept in the subjectivist setting of De Finetti, and no longer derived from the unconditional probability function. The conditional probability satisfying $P(A \cap C) = P(A \mid C) \cdot P(C)$ still makes sense if $P(C) = 0$; see (Coletti and Scozzafava 2002).

Under the subjectivist view, a body of knowledge consists of a set of conditional probability assignments $\{P(A_i \mid C_j) = c_{ij}, i = 1, \ldots, m;\ j = 1, \ldots, n\}$. Such conditional events correspond to various hypothetical contexts whose probability is allowed to be 0. The questions of interest are then (i) to derive a probability distribution in agreement with those constraints (actually a sequence of probability measures on disjoint parts of S (Coletti and Scozzafava 2002); (ii) to find induced optimal bounds on some conditional probability $P(A|C)$. For instance, one may consider the probabilistic syllogism already studied by Boole and De Morgan. Namely suppose the quantities $P(B|A)$, $P(C|B)$ are precisely known, what can be inferred about $P(C|A)$? It turns out that if $P(C|B) < 1$, we can only conclude that $P(C|A) \in [0, 1]$. However when the values of $P(A|B)$ and $P(B|C)$ are known as well, we can compute non-trivial bounds on $P(C|A)$. These bounds can be found in (Dubois et al. 1993). For example, it can be shown that

$$P(C|A) \geq P(B|A) \cdot \max\left(0, 1 - \frac{1 - P(C|B)}{P(A|B)}\right)$$

and that this lower bound is tight.

Yet another mathematical attempt to justify probability theory as the only reasonable belief measure is the one of Cox (1946). To do so he relied on the Boolean structure of the set of events and a number of postulates, considered compelling. Let $g(A|B) \in [0, 1]$ be a conditional belief degree, A, B being events in a Boolean algebra, with $B \neq \emptyset$. Cox assumed:

(i) $g(A \cap C|B) = F(g(A|C \cap B), g(C|B))$ (if $C \cap B \neq \emptyset$);
(ii) $g(\overline{A}|B) = n(g(A|B))$, $B \neq \emptyset$, where $\overline{A}$ is the complement of A;
(iii) function F is supposed to be twice differentiable, with a continuous second derivative, while function n is twice differentiable.

On such a basis, Cox claimed $g(A|B)$ is necessarily isomorphic to a conditional probability measure.

This result is important to recall here because it has been repeated *ad nauseam* in the literature of artificial intelligence to justify probability and Bayes rule as the only reasonable approach to represent and process numerical belief degrees (Horvitz

et al. 1986; Cheeseman 1988; Jaynes 2003). However some reservations must be made. First, the original proof by Cox turned out to be faulty – see Paris (1994) for another version of this proof based on a weaker condition (iii): it is enough that F be strictly monotonically increasing in each place. Moreover, Halpern (1999a, b) has shown that the result does not hold in finite spaces, and we need an additional technical condition to get it in the infinite setting. Independently of these technical issues, it should be noticed that postulate (i) sounds natural only if one takes Bayes conditioning for granted; the second postulate requires self-duality, i.e., it rules out all other approaches to uncertainty considered in the rest of this chapter and in the next one; it forbids the representation of uncertainty due to partial ignorance as seen later on. Noticing that $P(A|B)$ can be expressed in terms of $P(A \cap B)$ and $P(\overline{A} \cap B)$, an alternative option would be to start with assuming $g(A|B)$ to be a function of $g(A \cap B)$ and $g(\overline{A} \cap B)$), adding the postulate $g((A|B)|C) = g(A|B \cap C)$, if $B \cap C \neq \emptyset$, but dropping (iii). This could lead to a general study of conditional belief as outlined in Dubois et al. (2010). The above comments seriously weaken the alleged universality of Cox results.

3.3 Bayes Rule: Revision Versus Prediction

Assuming that a single probability measure is available, the additivity property of probability theory implies two noticeable results for conditional probabilities, that are instrumental in practice:

- The theorem of total probability: If $\{C_1, \ldots, C_k\}$ forms a partition of S, then

$$P(A) = \sum_{i=1}^{k} P(A \mid C_i) P(C_i).$$

- Bayes theorem

$$P(C_j \mid A) = \frac{P(A \mid C_j) P(C_j)}{\sum_{i=1}^{k} P(A \mid C_i) P(C_i)}.$$

The first result makes it possible to derive the probability of an event in a general context S given the probabilities of this event in various subcontexts $C_1, \ldots, C_k$, provided they form a partition of the set of possible states, and if probabilities of these subcontexts are available. Bayes theorem is useful to solve classification problems: suppose k classes of objects forming a partition of S. If the probability that objects in each class C_j satisfy property A is known, as well as prior probabilities of classes C_j, then if a new object is presented that is known to possess property A, it is easy to compute the probability $P(C_j \mid A)$ that this object belongs to class C_j. Diagnosis problems are of the same kind, replacing "class" by "disease" and "observed property" by "symptom". The use of conditional probabilities in Bayesian networks

first proposed by Pearl (1988) is extensively discussed in chapter "Belief Graphical Models for Uncertainty Representation and Reasoning" of Volume 2 of this treatise.

Most of the time, the information encoded in a probability distribution refers to some population. It represents *generic* information, with a frequentist meaning. One can use this information to infer beliefs about a particular situation, in which one has made partial, but unambiguous observations. This task is referred to as *prediction*. If $P(A \mid C)$ is the (frequentist) probability of event A in context C, one measures the agent's confidence $g(A \mid C)$ in proposition A, when only information C is known, by the quantity $P(A \mid C)$, assuming the current situation is typical of context C. The agent's belief about proposition A is updated from $g(A) = P(A)$ to $g(A \mid C) = P(A \mid C)$ after observing that C is true in the current situation, and nothing else. Conditioning is thus used to update the agent's contingent beliefs about the current situation by exploiting generic information. For instance, probability measure P represents medical knowledge (often compiled as a Bayesian network). Contingent information C represents test results for a given patient. Conditional probability $P(A \mid C)$ is then the probability that disease A is present for patients with test results C; this value also measures the (contingent) probability that the particular patient under consideration has disease A. We can remark that, under inference of this kind, the probability measure P does not change. One only applies generic knowledge to the reference class C, a process called *focalization*.

In the context of subjective probability à la De Finetti, to say that a probability distribution P is known means to know $P(A \mid C)$ for all events in all contexts. The agent only chooses the conditional probability of the event of interest in the context that is in agreement with the information on the current situation.

These views of conditioning differ from a revision process leading to a change of probability measure. Indeed some authors justify conditional probability in terms of belief revision (Gärdenfors 1988). The quantity $P(A \mid C)$ is then viewed as the *new* probability of A when the agent learns that C occurred. A basic principle of belief revision is minimal change: the agent revises its beliefs minimally while absorbing the new information item, interpreted as the constraint $P(C) = 1$. Under this view, the nature of the original probability function, and of the input information is the same, as is the posterior probability. In this revision scenario (see chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume), the probability function can be generic (e.g., frequentist, population-based) or singular (a subjective probability) and the input information is of the same kind as P (we learn that C has actual probability 1). The revision problem is then defined as follows: find a new probability P' as close as possible to P such that $P'(C) = 1$, which obeys minimal change (Domotor 1985). Using a suitable measure of relative information (e.g., Kullback-Leibler relative entropy) it can be shown that $P'(A) = P(A \mid C), \forall A$.

This revision scenario contrasts with the one of making predictions based on generic knowledge (in the form of a probability measure P describing the behavior of a population) and singular information items describing a situation of interest, even if the same tool, conditional probability, is used. As will be seen later on in this chapter and in the next one, the two tasks (revision vs. prediction) will no longer be solved by the same form of conditioning in more general uncertainty theories.

3.4 Probability Distributions and Partial Ignorance

The so-called *Bayesian* approach to subjective probability theory postulates the unicity of the probability distribution as a preamble to any kind of uncertainty modeling (see, for instance, Lindley (1982)), which could read as follows: any state of knowledge is representable by a single probability distribution. Note that indeed, if, following the fair bet procedure of De Finetti, the agent decides to directly assign subjective probabilities via buying prices to all *singletons* in S, the coherence principle forces this agent to define a unique probability distribution in this case. However, it is not clear that the limited perception of the human mind makes the agent capable of providing real numbers with infinite precision in the unit interval as prices. The measurement of subjective probability should address this issue in some way. If one objects that perhaps the available knowledge of the agent hampers the assignment of precise prices, the Bayesian approach sometimes resorts to selection principles such that the Laplace Principle of Insufficient Reason that exploits symmetries of the problem, or the maximal entropy principle (Jaynes 1979; Paris 1994). Resorting to the latter in the subjectivist setting is questionable because it would select the uniformly distributed probability whenever it is compatible with the imprecise probabilistic information, even if imprecise probabilities suggest another trend.

Applying the Bayesian credo as recalled above forces the agent to use a single probability measure as the universal tool for representing uncertainty whatever its source. This stance leads to serious difficulties already pointed fourty years ago (Shafer 1976). For one, it means we give up making any difference between uncertainty due to incomplete information or ignorance, and uncertainty due to a purely random process, the next outcome of which cannot be predicted. Take the example of die tossing. The uniform probability assignment corresponds to the assumption that the die is fair. But if the agent assigns equal prices to bets assigned to all facets, how can we interpret it? Is it because the agent is sure that the die is fair and its outcomes are driven by pure randomness (because, say, they could test it hundreds of times prior to placing the bets)? Or is it because the agent who is given this die, has just no idea whether the die is fair or not, so has no reason to put more money on one facet than on another one? Clearly the epistemic state of the agent is not the same in the first situation and in the second one. But the uniformly distributed probability function is mute about this issue.

Besides, the choice of a set of mutually exclusive outcomes depends on the chosen language, e.g., the one used by the information source, and several languages or points of view can co-exist in the same problem. As there are several possible representations of the state space, the probability assignment by an agent will be language-dependent, especially in the case of ignorance: a uniform probability on one state space may not correspond to a uniform one on another encoding of the same state space for the same problem, while in case of ignorance this is the only representation left to the betting agent. Shafer (1976) gives the following example. Consider the question of the existence of extra-terrestrial life, about which the agent has no idea. If the variable v refers to the claim that life exists outside our planet ($v = l$), or not ($v = \neg l$),

then the agent proposes $P_1(l) = P_1(\neg l) = \frac{1}{2}$ on $S_1 = \{l, \neg l\}$. However it makes sense to distinguish between animal life (al), and vegetal life only (vl), which leads to the state space $S_2 = \{al, vl, \neg l\}$. The ignorant agent is then bound to propose $P_2(al) = P_2(vl) = P_2(\neg l) = \frac{1}{3}$. As l is the disjunction of al and vl, the distributions P_1 and P_2 are not compatible with each other, while they are supposed to represent ignorance. A more casual example comes from noticing that expressing ignorance by means of a uniform distribution for $v \in [a, b]$, a positive interval, is not compatible with a uniform distribution on $v' = \log v \in [\log(a), \log(b)]$, while the agent has the same ignorance on v and v'.

Finally, it is not easy to characterize a single probability distribution by assigning lottery prices to propositions that do not pertain to singletons of the state space. Probability theory and classical logic, understood as knowledge representation frameworks, do not get along very conveniently. A maximal set of propositions to each of which the same lower bound of probability strictly less than 1 is assigned is generally not deductively closed. Worse, the conditioning symbol in probability theory is not a standard Boolean connective. The values $Prob(A|B)$ and $Prob(B \rightarrow A) = Prob(\overline{B} \cup A)$ can be quite different from each other, and will coincide only if they are equal to 1 (Kyburg, Jr. and Teng 2012). A natural concise description of a probability distribution on the set of interpretations of a language is easily achieved by a Bayesian network, not by a weighted set of propositional formulas.

Besides, in first-order logic, we should not confuse an uncertain universal conjecture (Gaifman and Snir 1982) (for instance, $Prob(\forall x, P(x) \rightarrow Q(x)) = \alpha$) with a universally valid probabilistic statement (for instance, $\forall x, Prob(P(x) \rightarrow Q(x)) = \alpha$, or $\forall x, Prob(Q(x)|P(x)) = \alpha$). Extensions of Bayesian networks to first-order logical languages are proposed by Milch and Russell (2007). Finally we give a number of references to works that tried to reconcile probabilistic and logical representations (propositional, first-order, modal) in various ways: (Halpern 1990; Bacchus 1991; Nilsson 1993; Abadi and Halpern 1994; Marchioni and Godo 2004; Jaeger 2001; Halpern and Pucella 2002, 2006; Jaeger 2006). See chapter "Languages for Probabilistic Modeling Over Structured Domains" in Volume 2 for a detailed account of probabilistic relational languages.

The above limitations of expressive power for single probability distributions have motivated the emergence of other approaches to uncertainty representations. Some of them give up the numerical setting of degrees of belief and use ordinal or qualitative structures considered as underlying the former subjectivist approaches. For instance, see (Renooij and van der Gaag 1999; Parsons 2001; Bolt et al. 2005; Renooij and van der Gaag 2008) for works that try to provide a qualitative counterpart of Bayesian nets. Another option is to tolerate incomplete information in the probabilistic approaches, which leads to different mathematical models of various level of generality. They are reviewed in the rest of this chapter and in the next chapter in this volume. Possibility theory is the simplest approach of all, and is found in both qualitative and quantitative settings (Dubois and Prade 1998).

3.5 Conditional Events and Big-Stepped Probabilities

Instead of considering a conditional probability function $P(\cdot \mid C)$ as a standard probability distribution on C, De Finetti (1936) was the first scholar to consider the set of conditional probabilities $\{P(A \mid C) : A \subseteq S, C \neq \emptyset\}$ as a probability assignment to *three-events, or conditional events* $A \mid C$. A conditional event can be informally understood as a conditional statement or an if-then rule: if all the currently available information is described by C, then conclude that A holds.

A three-event is so called because it partitions the state space S into three disjoint sets of states s:

- Either $s \in A \cap C$; s is called an example of the rule "if C then A". The three-event is considered as true at state s, which is denoted by $t(A \mid C) = 1$;
- or $s \in \overline{A} \cap C$; s is called a counter-example of the rule "if C then A". The three-event is considered as false at state s, which is denoted by $t(A \mid C) = 0$;
- or $s \in \overline{C}$; then the rule "if C then A" is said not to apply to s. In this case the three-event takes a third truth-value at s, which is denoted by $t(A \mid C) = I$ where I stands for inapplicable.

A three-event $A \mid C$ can thus be interpreted as a pair $(A \cap C,\ \overline{A} \cap C)$ of disjoint sets of examples and counter-examples. A qualitative counterpart of Bayes rule holds, noticing that as the set-valued solutions of the equation $A \cap C = X \cap C$ are all sets $\{X : A \cap C \subseteq X \subseteq A \cup \overline{C}\}$, which is another possible representation of $A \mid C$ (as an interval in the Boolean algebra of subsets of S). This definition of conditional events as pairs of subsets suggests a natural consequence relation between conditional events defined as follows (Dubois and Prade 1994):

$$B \mid A \vDash D \mid C \Leftrightarrow A \cap B \vDash C \cap D \text{ and } C \cap \overline{D} \vDash A \cap \overline{B}$$

which reads: all examples of $B \mid A$ are examples of $D \mid C$ and all counter-examples of $D \mid C$ are counter-examples of $B \mid A$. Note that only the second condition coincides with the deductive inference between material conditional counterparts of the three-events. Material conditionals highlight counter-examples of rules, not examples. When ordering the truth-values as $0 < I < 1$, this semantic inference relation also reads $B \mid A \vDash D \mid C \Leftrightarrow t(B \mid A) \leq t(D \mid C)$.

Representing if-the rules by conditional events avoids some paradoxes of material implications, such as the confirmation paradox: in the material implication representation, the rule *if* C *then* A is the same as its contrapositive version *if* $\overline{A}$ *then* $\overline{C}$. If we use material implication, we are bound to say that an observation confirms a rule if it makes this material implication true. So, both $s_1 \in A \cap C$ and $s_2 \in \overline{A} \cap \overline{C}$ confirm the rule. But this is questionable: suppose the rule means all ravens are black. Then meeting a white swan would confirm that all ravens are black (Hempel 1945). This anomaly does not occur with conditional events as $A \mid C$ is not equivalent to $\overline{C} \mid \overline{A}$: they have the same counterexamples (e.g., white ravens) since they have the same material conditional representations, but they do not have the same examples:

s_2 is an example of $\overline{C} \mid \overline{A}$ (e.g., a white swan), but this three-event does not apply to $s_1 \in A \cap C$ (Benferhat et al. 2008).

It is worth noticing that a conditional probability $P(A \mid C)$ is indeed the probability of a conditional event $A \mid C$ since $P(A \mid C)$ is entirely determined by the pair $(P(A \cap C), P(\overline{A} \cap C))$. Moreover, if all probabilities of singletons are positive, and $B \mid A \models D \mid C$, it is clear that $P(B \mid A) \leq P(D \mid C)$.

A three-valued logic for conjunctions of conditional events was developed by Dubois and Prade (1994). A three-valued extension of standard conjunction is used where the third truth-value I is a semi-group identity. This three-valued logic truth-table for conjunction and the above inference rule offer an alternative simple semantics for the non-monotonic inference system P (Kraus et al. 1990) that captures exception-tolerant reasoning, where conditional events $B \mid A$ model generic rules of the form: *generally if A then B* (see also the section on non-monotonic inference in chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume and Sect. 4.2.1 of the present chapter). Non-monotonicity manifests itself by the fact that the inference $B \mid A \models B \mid A \cap C$ does not hold (the latter has less examples than the former), so that, like in probability theory, conditional events $B \mid A$ and $\overline{B} \mid (A \cap C)$ can coexist in the same rule base without ruling out any possible world (contrary to material conditionals in propositional logic that would enforce $A \cap C = \emptyset$). Under this logic, to infer a plausible conclusion F from a state of knowledge described by the epistemic set E, and a conditional base (a set of conditional events) $\mathscr{C}$ that encodes generic information, means to infer the conditional event $F \mid E$ from a conditional event obtained as a suitable conjunction of a subset of conditional events in $\mathscr{C}$ (Benferhat et al. 1997).

Note also that under this inference scheme, the conditional event $A \cap B \mid C$ follows from $\mathscr{C} = \{A \mid C, B \mid C\}$, so that the set of plausible conclusions obtained from C will be deductively closed. But as pointed our earlier, $P(A \cap B \mid C) > 1 - \theta$ does not follow from $P(A \mid C) \geq 1 - \theta$ and $P(B \mid C) > 1 - \theta$, however small θ may be. In particular, if we minimally define A as an accepted belief whenever $P(A \mid C) > P(\overline{A} \mid C)$ (in other words $P(A \mid C) > 1/2$), we see that contrary to what happens with conditional events, a set of probabilistically accepted beliefs will not be closed in the sense of classical deduction. To ensure compatibility between symbolic inference between conditional events and accepted beliefs in the above sense, we can restrict the set of probability distributions to a subset for which deductive closure will be respected. This kind of probability measure is called *big-stepped probability* and is defined as follows by the condition:

$$\forall i < n-1,\ p_i > \sum_{j=i+1,..,n} p_j \text{ where } p_i = P(s_i) \text{ and } p_1 > .. > p_{n-1} \geq p_n > 0.$$

For an example of big-stepped probability distribution when $n = 5$, consider $p_1 = 0.6$, $p_2 = 0.3$, $p_3 = 0.06$, $p_4 = 0.03$, $p_5 = 0.01$. This type of exponential-like (or super-decreasing) probability distributions are at odds with uniform distributions. They offer a full-fledged probabilistic semantics to the logic of conditional events and Kraus, Lehmann and Magidor (1990)'s system P for coping with exceptions in rule-based systems (Benferhat et al. 1999b; Snow 1999).

4 Possibility Theory

Basic building blocks of possibility theory go back to a seminal paper by Zadeh (1978) and further works by Dubois and Prade (1988) quite independently of the works of an English economist, Shackle (1961) who had outlined a similar theory some thirty years before (in terms of so-called *degrees of surprize* to be equated to degrees of impossibility). Actually, Zadeh and Shackle did not have the same intuitions in mind. Zadeh viewed his possibility distributions as representing flexible constraints representing pieces of fuzzy information in natural language (viz. "what is the possibility that John is more than 30 years old assuming he is *young*"?). In contrast Shackle tried to offer a representation of how the human mind handles uncertainty that is supposedly more faithful than probability theory. After the publication of Zadeh's paper, it soon became patent that possibility distributions were not necessarily generated from the representation of gradual properties in natural language (like *young*), but that they allowed to formalize a gradual notion of epistemic states by extending the disjunctive view of sets to fuzzy sets, whereby degrees of possibility, understood as plausibility, can be assigned to interpretations induced by any propositional language.

Possibility measures are maximum-decomposable for disjunction. There have companion set-functions called necessity measures, obtained by duality, that are minimum-decomposable for conjunction. They can be completed by two other set-functions that use the same basic setting. This general framework is first recalled in the following subsections. Then the distinction between qualitative and quantitative possibility theories is recalled. Qualitative possibility theory is best couched in possibilistic logic, which is briefly outlined. This section is completed by an exposition of the relationships between qualitative possibility theory and non-monotonic reasoning, and the modeling of default rules. We end the section by a possibility-theory rendering of formal concept analysis, which was originally developed in a very different perspective.

4.1 General Setting

Consider a mapping π_v from S to a totally ordered scale L, with top denoted by 1 and bottom by 0. It can be the unit interval as suggested by Zadeh, or generally any finite chain such as $L = \{0, 0.1, 0.2, \ldots, 0.9, 1\}$, or a totally ordered set of symbolic grades. The possibility scale can be the unit interval as suggested by Zadeh, or generally any finite chain, or even the set of non-negative integers. For convenience, it is often assumed that the scale L is equipped with an order-reversing map denoted by $\lambda \in L \mapsto 1 - \lambda$. More generally L can be a complete lattice with a top and a bottom element, denoted by 1 or 0 respectively. The larger $\pi_v(s)$, the more possible, i.e., plausible the value s for the variable v, that supposedly pertains to some attribute (like the age of John in Sect. 2.4). The agent information about v is captured by π_v

called a *possibility distribution*. Formally, the mapping π is the membership function of a fuzzy set (Zadeh 1978), where membership grades are interpreted in terms of plausibility. If the possibility distribution stems from gradual linguistic properties, plausibility is measured in terms of distance to fully plausible situations, not in terms of, e.g., frequency. Function π represents the state of knowledge of an agent (about the actual state of affairs), also called an *epistemic state* distinguishing what is plausible from what is less plausible, what is the normal course of things from what is not, what is surprising from what is expected. It represents a flexible restriction on what is the actual state with the following conventions (similar to probability, but opposite to Shackle's potential surprise scale)[4]:

- $\pi(s) = 0$ means that state s is rejected as impossible;
- $\pi(s) = 1$ means that state s is totally possible (= plausible).

If the universe S is exhaustive, at least one of the elements of S should be the actual world, so that $\exists s, \pi(s) = 1$ (normalised possibility distribution). This condition expresses the consistency of the epistemic state described by π. Distinct values may simultaneously have a degree of possibility equal to 1. In the Boolean case, π is just the characteristic function of a subset $E \subseteq S$ of mutually exclusive states, ruling out all those states considered as impossible. Possibility theory is thus a (fuzzy) set-based representation of incomplete information. There are two extreme cases of imprecise information

- *Complete ignorance*: without information, only tautologies can be asserted. It is of the form $v \in S$, corresponding to the possibility distribution $\pi_v^?(s) = 1, \forall s \in S$.
- *Complete knowledge*: it is of the form $v = s_0$ for some value $s_0 \in S$, corresponding to the possibility distribution $\pi_v^{s_0}(s) = 1$ if $s = s_0$ and 0 otherwise. Note that it is the value 0 that brings information in π_v.

Possibility theory is driven by the *principle of minimal specificity*. It states that *any hypothesis not known to be impossible cannot be ruled out*. It is a minimal commitment, cautious information principle. Basically, we must always try to maximize possibility degrees, taking constraints into account. Measures of possibilistic specificity have been proposed in a way similar to probabilistic entropy (Higashi and Klir 1982).

4.1.1 The Two Basic Set-Functions

Plausibility and certainty evaluations, induced by the information represented by a distribution π_v, pertaining to the truth of proposition $v \in A$ can then be defined. We speak of degrees of possibility and necessity of event A:

$$\Pi(A) = \max_{s \in A} \pi_v(s); \quad N(A) = 1 - \Pi(\overline{A}) = \min_{s \notin A} 1 - \pi_v(s) \tag{8}$$

[4] If $L = \mathbb{N}$, the conventions are opposite: 0 means possible and ∞ means impossible.

By convention $\Pi(\emptyset) = 0$ and then $N(S) = 1$. $\Pi(S) = 1$ (hence $N(\emptyset) = 0$) follows if π_v is normalized. The symbol $1 - (\cdot)$ should not suggest these degrees are numerical. It is just the order-reversing map on L.

When distribution π_v takes value on the binary scale $\{0, 1\}$, i.e., there is a subset $E \subseteq S$ such that $\pi_v(s) = 1 \Leftrightarrow s \in E$, it is easy to see that $\Pi(A) = 1$ if and only if the proposition $v \in A$ is not inconsistent with the information item $v \in E$, i.e., if $A \cap E \neq \emptyset$. Likewise, $N(A) = 1$ if and only if proposition $v \in A$ is implied by the information item $v \in E$ (since $E \subseteq A$). $\Pi(A) = 0$ means that it is impossible that the assertion $v \in A$ is true if $v \in E$ is true. $N(A) = 1$ expresses that the assertion $v \in A$ is certainly true if $v \in E$ is true.

Functions N and Π are tightly linked by the duality property $N(A) = 1 - \Pi(\overline{A})$. This feature highlights a major difference between possibility and necessity measures and probability measures that are self dual in the sense that $P(A) = 1 - P(\overline{A})$.

The evaluation of uncertainty in the style of possibility theory is at work in classical and modal logics. If K is a set of propositional formulas in some language, suppose that E is the set of its models. Consider a formula p which is the syntactic form of the proposition $v \in A$, then $N(A) = 1$ if and only if K implies p, and $\Pi(A) = 0$ if and only if $K \cup \{p\}$ is logically inconsistent. Of course, the presence of p inside K encodes $N(A) = 1$, while the presence of its negation $\neg p$ in K encodes $\Pi(A) = 0$. In contrast, in the propositional language of K, one cannot encode $N(A) = 0$ nor $\Pi(A) = 1$, e.g., that $v \in A$ is unknown. To do this inside the language, one must use the formalism of modal logic (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume), that prefixes propositions by modalities of possibility ($\Diamond$) and necessity ($\Box$): in a modal base K^{mod}, $\Diamond p \in K^{mod}$ directly encodes $\Pi(A) = 1$, and $\Box p \in K^{mod}$ encodes $N(A) = 1$ (the latter merely encoded by $p \in K$ in propositional logic). The duality relation between Π and N is very well known in modal logic, where it reads $\Diamond p = \neg\Box\neg p$. A simple modal logic (a very elementary fragment of the *KD* logic), called MEL (for minimal epistemic logic), has been defined by Banerjee et Dubois (2014) with a semantics in terms of non-empty subsets of interpretations ($\{0, 1\}$-valued possibility distributions). A similar idea was first suggested by Mongin (1994). The satisfaction of $\Box p$ by an epistemic set E means that $E \subseteq A$, if p encodes $v \in A$.

In the possibilistic setting one distinguishes three extreme epistemic attitudes pertaining to an information item $v \in A$:

- the certainty that $v \in A$ is true: $N(A) = 1$, hence $\Pi(A) = 1$;
- the certainty that $v \in A$ is false: $\Pi(A) = 0$, hence $N(A) = 0$:
- ignorance pertaining to $v \in A$: $\Pi(A) = 1$, and $N(A) = 0$.

These attitudes can be refined as soon as L contains at least one value differing from 0 or 1 leading to situations where $0 < N(A) < 1$ or $0 < \Pi(A) < 1$.

It is easy to verify that possibility and necessity measures saturate inequalities (2) verified by capacities:

$$\Pi(A \cup B) = \max(\Pi(A), \Pi(B)). \tag{9}$$

$$N(A \cap B) = \min(N(A), N(B)). \tag{10}$$

Possibility measures are said to be *maxitive* and are fully characterized by the maxitivity property (9) in the finite case; necessity measures are said to be *minitive* and are fully characterized by the minitivity property (10) in the finite case, including when these functions take values in [0, 1].

In general possibility and necessity measures do not coincide. It is impossible for a set function to be at the same time maxitive and minitive for all events, except in case of complete knowledge ($E = \{s_0\}$). Then $N = \Pi$ also coincide with a Dirac $\{0, 1\}$-valued probability measure.

Observe that we only have

$$N(A \cup B) \geq \max(N(A), N(B)) \text{ and } \Pi(A \cap B) \leq \min(\Pi(A), \Pi(B)),$$

and it may occur that the difference is maximal. Indeed in the $\{0, 1\}$-valued case, if it is not known whether A is true or false (namely, $A \cap E \neq \emptyset$ and $\overline{A} \cap E \neq \emptyset$), then $\Pi(A) = \Pi(\overline{A}) = 1$ and $N(A) = N(\overline{A}) = 0$; however, by definition $\Pi(A \cap \overline{A}) = \Pi(\emptyset) = 0$ and $N(A \cup \overline{A}) = N(S) = 1$.

4.1.2 Two Decreasing Set Functions. Bipolarity

Yet another set function Δ and its dual companion ∇ (first introduced in 1991, see (Dubois and Prade 1998)) can be naturally associated with the possibility distribution π_v in the possibilistic framework:

$$\Delta(A) = \min_{s \in A} \pi_v(s); \quad \nabla(A) = 1 - \Delta(\overline{A}) = \max_{s \notin A} 1 - \pi_v(s) \tag{11}$$

Observe first that in contrast with Π and N, Δ and ∇ are decreasing functions with respect to set inclusion (hence to the logical consequence relation). Function Δ is called *strong possibility* or *guaranteed possibility* since inside set A, the degree of possibility is never less than $\Delta(A)$ (while Π is only a weak possibility degree that just measures consistency); dually, function ∇ is a measure of weak necessity, while N is a measure of strong necessity. Besides, the following inequality hold:

$$\forall A, \max(\Delta(A), N(A)) \leq \min(\Pi(A), \nabla(A))$$

provided that both π_v and $1 - \pi_v$ are normalised.

Characteristic properties of Δ and ∇ are:

$$\Delta(A \cup B) = \min(\Delta(A), \Delta(B)); \quad \Delta(\emptyset) = 1. \tag{12}$$

$$\nabla(A \cap B) = \max(\nabla(A), \nabla(B)); \quad \nabla(S) = 0. \tag{13}$$

From the standpoint of knowledge representation, it is interesting to consider the case when the possibility distribution π_v only takes a finite number of distinct values $\alpha_1 = 1 > \cdots > \alpha_n > \alpha_{n+1} = 0$. It can then be described by n nested subsets $E_1 \subseteq \cdots \subseteq E_i \subseteq \cdots \subseteq E_n$ where $\pi_v(s) \geq \alpha_i \Leftrightarrow s \in E_i$. One can then verify that $\Delta(E_i) \geq \alpha_i$, while $N(E_i) \geq 1 - \alpha_{i+1}$ for $i = 1, \ldots, n$, and that $\pi_v(s) = \max_{E_i \ni s} \Delta(E_i) = \min_{E_i \not\ni s}(1 - N(E_i))$ (with conventions $\max_\emptyset = 0$ et $\min_\emptyset = 1$). A distribution π_v can thus be seen as a weighted disjunction of sets E_i, from the point of view of Δ, and as a weighted conjunction of sets E_i from the point of view of N. The reading of π_v viewed from Δ offers a positive understanding of the possibility distribution, expressing to which extent each value is possible, while viewed from N, π_v expressed to what extent each value is not impossible (since each value s is all the more impossible as it belongs to fewer subsets E_i).

These positive and negative flavors respectively attached to Δ and N lay the foundation of a *bipolar* representation of information in possibility theory (Benferhat et al. 2008). The idea of bipolarity refers to an explicit handling of positive or negative features of information items (Dubois and Prade, eds. 2008). There are several forms of bipolarity and we only focus on the case when it comes from the existence of distinct sources of information. In the possibilistic setting, *two* possibility distributions δ_v and π_v are instrumental to respectively represent values that are guaranteed possible for v and values that are just known to be not-impossible (because not ruled out). The concept of bipolarity applies to representing knowledge as well as preferences. These distributions are differently interpreted: when representing knowledge $\delta_v(s) = 1$ means that s is certainly possible because this value or state has been actually observed, and, when representing preferences, s is an ideal choice. Moreover, when representing knowledge, $\delta_v(s) = 0$ just means that nothing is known about this value that has not been observed, and, when representing preferences, that the choice s is not at all attractive. In contrast, when representing knowledge, $\pi_v(s) = 1$ means that s is not impossible (just feasible when representing preference), but $\pi_v(s) = 0$ means that s is completely ruled out (or not acceptable for preferences). Intuitively, any state that is guaranteed possible should be among the non-impossible situations. So there is a coherence condition to be required: $\delta_v \leq \pi_v$. It corresponds to a standard fuzzy set inclusion). In possibilistic logic presented further on, the distribution π_v stems from constraints of the form $N(A_i) \geq \eta_i$, and distribution δ_v from statements of the form $\Delta(B_j) \geq \delta_j$ where $A_i \subseteq S$, $B_j \subseteq S$, and $\eta_i \in L$, $\delta_j \in L$. The idea of bipolar representation is not limited to possibility theory, even if it was not often considered in other frameworks (see Dubois et al. (2000a)).

4.1.3 Possibility and Necessity of Fuzzy Events

The set functions Π, N, Δ et ∇ can be extended to fuzzy sets. The (weak) possibility of a fuzzy event F is defined by $\Pi(F) = \sup_s \min(F(s), \pi_v(s))$ (Zadeh 1978); still using duality, the necessity of a fuzzy event then reads $N(F) = 1 - \Pi(\overline{F}) = \inf_s \max(F(s), 1 - \pi_v(s))$. Functions Π and N still satisfy, respectively, maxitivity (9) and minitivity (10) properties. The values $\Pi(F)$ and $N(F)$ turn out to be spe-

cial cases of Sugeno integrals (see chapter "Multicriteria Decision Making" in this volume). Possibility and necessity of fuzzy events are instrumental to evaluate the extent to which a flexible condition is satisfied by an ill-known piece of data (Cayrol et al. 1982); in particular, if $\pi_v = F$, only $N(F) \geq 1/2$ obtains, which at first glance may be questionable. To get $N(F) = 1$, the condition $\forall s\ \pi_v(s) > 0 \Rightarrow F(s) = 1$ is needed, which means the inclusion of the support of π in the core of F so that any value that is possible even to a very low extent be fully in agreement with F. Such evaluations have been applied to fault diagnosis problems using a qualitative handling of uncertainty, where one may separate anomalies that more or less certainly appear when a failure occurs, from anomalies that more or less possibly appear (Cayrac et al. 1996; Dubois et al. 2001). Functions Δ and ∇ extend similarly to fuzzy events as $\Delta(F) = \inf_s \max(1 - F(s), \pi_v(s))$, letting $\nabla(F) = 1 - \Delta(\overline{F})$ by duality, while preserving respective properties (12) and (13).

Set functions N and Δ on fuzzy events are also very useful to represent fuzzy if-then rules (see also chapter "Case-Based Reasoning, Analogical Reasoning, and Interpolation" in this volume) of the form *the more v is F, the more it is* sure *that y is G*, and *the more v is F, the more it is* possible *that y is G* respectively, where F (but possibly G as well) are gradual properties represented by fuzzy sets (Dubois and Prade 1996). Indeed, the first type of rule expresses a constraint of the form $N(G) \geq F(s)$ while the second one is better modeled by the inequality $\Delta(G) \geq F(s)$. However, the first type of rule, where $1 - F(s)$ is viewed as the degree of possibility that the conclusion G is false, while in the second type of rule $F(x)$ is the minimal degree of possibility that the conclusion G holds, which corresponds to the following possibility distributions on the joint domain of (x, y):

$$\pi_{x,y}(s,t) \leq \max(1 - F(s), G(t)) \text{ and } \pi_{x,y}(s,t) \geq \min(F(s), G(t)).$$

Definitions of the strong necessity and possibility functions compatible with these inequalities are not the ones based on Zadeh's weak possibility of a fuzzy event. Based on the following equivalence: $c \leq \max(a, 1 - b) \Leftrightarrow (1 - a) \to (1 - c) \geq b$, where $\to$ is Gödel implication

$$u \to v = \begin{cases} 1 \text{ si } u \leq v, \\ v \text{ otherwise,} \end{cases}$$

the following extensions of strong necessity and possibility of fuzzy events N and Δ must be used: $N(G) = \inf_s (1 - F(s)) \to (1 - \pi_v(s))$ and $\Delta(G) = \inf_s F(s) \to \pi_v(s)$. These evaluations do reduce to strong necessity and possibility of standard events, like the ones in the previous paragraph, but the necessity function satisfies $N(G) = 1$ when $\pi_v = G$ (since we expect some equivalence between statements such as *it is sure that John is young* and *John is young*). Likewise, $\Delta(G) = 1$ when $\pi_v = G$. See Dubois et al. (2017a) for a systematic analysis of extensions of the four set functions of possibility theory to fuzzy events. The two types of fuzzy rules reflect a bipolar view of a standard rule *if* $v \in A$ *then* $y \in B$, which, on a Cartesian product of domains $S \times T$ can be represented either by the relational constraint

$R(s,t) \geq (A \times B)(s,t)$ pointing out examples, or by the relational constraint $\overline{R}(s,t) \geq (A \times \overline{B})(s,t) \Leftrightarrow R(s,t) \leq (\overline{A} + B)(s,t)$ excluding counter-examples, where the overbar means complementation and where $A + B = \overline{\overline{A} \times \overline{B}}$. The view of an if-then rule as a conditional event $B|A$ is thus retrieved.

4.1.4 Conditioning in Possibility Theory: Qualitative Versus Quantitative Settings

Since the basic properties in possibility are based on minimum, maximum and an order-reversing map on the uncertainty scale $(1 - (\cdot)$ on the unit interval, and $1 - \alpha_k = \alpha_{m-k})$ on a bounded chain $\{\alpha_0, \cdots, \alpha_m\})$, it is not imperative to use a numerical setting for the measurement of possibility and necessity. When the set functions take values in the unit interval, we speak of *quantitative possibility theory*. When they take values in a bounded chain, we speak of *qualitative possibility theory* (Dubois and Prade 1998). In both cases, possibility theory offers a simple, but non trivial, approach to non-probabilistic uncertainty. The two versions of possibility theory diverge when it comes to conditioning. In the qualitative case, there is no product operation, and the counterpart of Bayes rule is naturally expressed replacing it by the minimum operation on the bounded chain L:

$$\Pi(A \cap B) = \min(\Pi(A \mid B), \Pi(B)). \tag{14}$$

This equation has no unique solution. In the spirit of possibility theory, one is led to select the least informative solution, according to minimal commitment, namely when $B \neq \emptyset$, and $A \neq \emptyset$:

$$\Pi(A \mid B) = \begin{cases} 1 & \text{if } \Pi(A \cap B) = \Pi(B), \\ \Pi(A \cap B) & \text{otherwise.} \end{cases} \tag{15}$$

This is just like conditional probability, except that we no longer make a division by $\Pi(B)$. When $\Pi(B) = 0$, $\Pi(A \mid B) = 1$ as soon as $A \cap B \neq \emptyset$. It reflects the idea than you may destroy available information when conditioning on an impossible event. Conditional necessity is defined by duality as[5]:

$$N(A \mid B) = 1 - \Pi(\overline{A} \mid B) = \begin{cases} 0 & \text{if } \Pi(\overline{A} \cap B) = \Pi(B); \\ N(A \cup \overline{B}) & \text{otherwise.} \end{cases}$$

[5]The Bayesian-like rule in terms of necessity measures, $N(A \cap B) = \min(N(A \mid B), N(B))$, is trivial. Its least specific solution, minimizing necessity degrees, is $N(A \mid B) = N(A \cap B) = \min(N(A), N(B))$, which defines in turn $\Pi(A \mid B) = \Pi(\overline{B} \cup A)$. It comes down to interpreting a conditional event as a material implication.

The least specific solution to equation (14) does capture an ordinal form of conditioning due to the following result:

$$N(A \mid B) > 0 \iff \Pi(A \cap B) > \Pi(\overline{A} \cap B)$$

when $\Pi(B) > 0$. Intuitively, it means that a proposition A is an accepted belief in context B if it is more plausible than its negation in this context. Like with probability, one may have that $\Pi(A \cap B) > \Pi(\overline{A} \cap B)$ while $\Pi(\overline{A} \cap B \cap C) > \Pi(A \cap B \cap C)$ in a more restricted context $B \cap C$. An alternative approach to conditional possibility is the one of Coletti and Vantaggi (2006), in which coherent possibility assessments on conditional events are defined based on Eq. (14), in the style of De Finetti's conditional probability.

In the case of quantitative possibility theory, the lack of continuity of the set function $\Pi(A \mid B)$ in Eq. (15) (de Cooman 1997) has led to replace minimum by product in this equation, mimicking conditional probability:

$$\Pi(A \mid B) = \frac{\Pi(A \cap B)}{\Pi(B)} \text{ provided that } \Pi(B) \neq 0.$$

As we shall see, it coincides with Dempster's rule of conditioning in evidence theory (see the next chapter in this volume). More generally, on the unit interval, the product can be extended to a triangular norm, and this general setting has been studied by Coletti and Vantaggi (2009) under the coherence approach in the style of De Finetti.

A major difference between possibility and probability theories concern independence. While stochastic independence between events with positive probability is a symmetric, negation-invariant, notion, since $Prob(B|A) = Prob(B)$ is equivalent to $Prob(A \cap B) = Prob(A) \cdot Prob(B)$ and to $Prob(B|\overline{A}) = Prob(B)$, this is no longer the case for possibilistic independence, several versions of which exist. For instance, in qualitative possibility theory, the equality $N(B|A) = N(B) > 0$ expresses that learning A does not question the accepted belief B and is not equivalent to $N(A|B) = N(A) > 0$ nor to $N(B|\overline{A}) = N(B) > 0$. Another form of independence is $N(B|A) = N(B) = N(\overline{B}|A) = N(\overline{B}) = 0$, which means that learning A leaves us ignorant about B; see (Dubois et al. 1999) for a complete study. There exist several definitions of conditional possibilistic independence between variables, in qualitative possibility theory, one being symmetric ($\Pi(x, y|z) = \min(\Pi(x|z), \Pi(y|z))$) and one being asymmetric ($\Pi(x|z) = \Pi(x|z, y)$); see Ben Amor et al. (2002). In the quantitative setting, independence between variables ($\forall x, y, z, \Pi(x|y, z) = \Pi(x|z)$) is symmetric since it is equivalent to $\forall x, y, z$, $\Pi(x, y|z) = \Pi(x|z) \cdot \Pi(y|z)$. The notion of possibilistic independence has also been studied by Coletti and Vantaggi (2006).

Conditional probability is the basis of representation of uncertain information in the form of Bayesian networks. There also exist graphical possibilistic representations in quantitative possibility theory, and in qualitative possibility theory as well (see chapter "Belief Graphical Models for Uncertainty Representation and Reasoning" in Volume 2) and some variants of possibilistic independence are useful to develop local uncertainty propagation methods.

4.2 Qualitative Possibility Theory

The main application of qualitative possibility theory is the development of possibilistic logic, an extension of classical logic that handles qualitative uncertainty, and is useful for encoding non monotonic reasoning and dealing with inconsistency. Besides, the basic setting of formal concept analysis can be seen as a set-valued counterpart of possibility theory, which leads to an interesting parallel between the two theories. We first present possibilistic logic. Note that qualitative possibility theory can be used for decision under uncertainty. Decision-theoretic foundations of qualitative possibility theory are presented in chapter "Decision under Uncertainty" of this volume.

4.2.1 Possibilistic Logic

The building blocks of possibilistic logic (Dubois et al. 1994; Dubois and Prade 2004) are pairs made of a (well-formed) formula of classical logic (propositional, or first order), and a weight (or level) which may be qualitative or numerical, but qualitatively handled. The weights usually belong to a totally ordered scale, but may only belong to a lattice structure with a smallest and a greatest element).

Necessity-based possibilistic logic In its basic version, possibilistic logic only allows to consider conjunctions of pairs of the form (p, α) where p is a propositional logic formula associated with a weight α belonging to the interval $(0, 1]$ (or to a finite totally ordered scale). The weight α is understood as a lower bound of a necessity mesure, i.e., the pair (p, α) encodes a constraint of the form $N(p) \geq \alpha$. It either corresponds to a piece of information (one is certain at level α that p is true), or a preference (p then represents a goal to be reached with priority α). The decomposability property of necessity mesures (10) ensures that we make no difference between $(p \wedge q, \alpha)$ and $(p, \alpha) \wedge (q, \alpha)$, and thus possibilistic bases, which are sets of such possibilistic pairs, can be expressed as conjunctions of weighted clauses.

Let $B^N = \{(p_j, \alpha_j) \mid j = 1, \ldots, m\}$ be a possibilistic base. At the semantic level, a possibility distribution π over the set of interpretations satisfies B^N (denoted by $\pi \models B^N$) if and only if $N(p) \geq \alpha_j$, $j = 1, \ldots, m$. The least specific possibility distribution that satisfies B^N exists and is of the form

$$\pi_B^N(s) = \min_{j=1,\ldots,m} \pi_{(p_j,\alpha_j)}(s) = \min_{j:\, s \models \neg p_j} 1 - \alpha_j,$$

where $\pi_{(p_j,\alpha_j)}(s) = 1$ if $s \models p_j$ and $1 - \alpha_j$ otherwise. Thus an interpretation s is all the more possible as it does not violate any formula p_j with a high priority level α_j, and $\pi \models B^N$ if and only if $\pi \leq \pi_B^N$.

The possibility distribution π_B^N provides a description "from above" (each pair (p_j, α_j) combined by min restricts the set of interpretations regarded as possible to some extent). It takes the form of a min-max combination, since $\pi_{(p_j,\alpha_j)}(s)$ is of

the form $\max(M(p_j)(s), 1-\alpha_j)$, where $M(p)$ denotes the characteristic function of the set of models of p.

Basic possibilistic logic possesses the cut rule

$$(\neg p \vee q, \alpha); (p \vee r, \beta) \vdash (q \vee r, \min(\alpha, \beta)).$$

This rule is sound and complete for refutation, with respect to possibilistic semantics. It should be noticed that the probabilistic counterpart to this rule, namely

$$Prob(\neg p \vee q) \geq \alpha; Prob(p \vee r) \geq \beta) \vdash Prob(q \vee r) \geq \max(0, \alpha + \beta - 1)$$

is sound, but not complete with respect to probabilistic semantics. Note that the deductive closure of possibilistic base $\{(p_j, \beta_j)$ with $\beta_j \geq \alpha\}_{j=1,n}$ only contains formulas with weights at least α, while this is wrong in general for the set of probabilistic formulas $\{p_j | Prob(p_j) \geq \alpha\}_{j=1,n}$ after closure with the corresponding resolution rule (except if $\alpha = 1$).

Dual possibilistic logic with guaranteed possibility weights A dual representation for possibilistic logic bases relies on guaranteed possibility functions. A formula is then a pair $[q, \beta]$, understood as the constraint $\Delta(q) \geq \beta$, where Δ is a guaranteed possibility (anti-)measure. It thus expresses that *all* the models of q are at least possible, at least satisfactory at level β. A Δ-base $B^\Delta = \{[q_i, \beta_i] \mid i = 1, \ldots, n\}$ is then associated with the distribution

$$\pi_B^\Delta(s) = \max_{i=1,\ldots,n} \pi_{[q_i,\beta_i]}(s) = \max_{i:\, s \models q_i} \beta_i,$$

with $\pi_{[q_i,\beta_i]}(s) = \min(M(q_i)(s), \beta_i)$. We define $\pi \models B^\Delta$ if and only if $\Delta(q_i) \geq \beta_i, \forall i = 1, \ldots, n$, which is equivalent to $\pi \geq \pi_B^\Delta$. So, π_B^Δ provides a description "from below" of the distribution representing an epistemic state. Taking advantage of decomposability property (12) of guaranteed possibility measures, it is easy to see that the set $\{[p, \alpha], [q, \alpha]\}$ is equivalent to the formula $[p \vee q, \alpha]$. Then putting classical logical formulas in disjunctive normal form, we can always rewrite a dual possibilistic base B^Δ into an equivalent base where all formulas q_i are conjunctions of literals.

A base B^Δ in dual possibilistic logic can always be rewritten equivalently in terms of a standard possibilistic logic N-base B^N (Benferhat and Kaci 2003; Benferhat et al. 2008), and conversely, in such a way that $\pi_B^N = \pi_B^\Delta$. However, note that Δ-based possibilistic logic obeys an inference rule different from the above resolution rule for N-bases: $[\neg p \wedge q, \alpha]; [p \wedge r, \beta] \vdash [q \wedge r, \min(\alpha, \beta)]$. It propagates guaranteed possibility levels in agreement with the decreasingness of set function Δ (indeed, if $r = \top$, and $q \vdash p$, then $\alpha = 1$ since $\Delta(\bot) = 1$, and the rule concludes $[q, \beta]$ from $[p, \beta]$).

A set of pieces of possibilistic Boolean information (with a finite number of possibility levels) can thus be represented by a possibility distribution on interpretations, but also in a more compact manner under the form of a finite set of formulas associated

either with a certainty (resp. priority) level, or with a level of guaranteed possibility (resp. satisfaction) when modeling knowledge (resp. preferences). Moreover, graphical representations of possibilistic bases in terms of possibilistic networks (either based on qualitative or on quantitative conditioning) have been proposed, with exact translations from one type of representation to the other (Benferhat et al. 2002). For an introduction to possibilistic networks and their algorithms, the reader is referred to chapter "Belief Graphical Models for Uncertainty Representation and Reasoning" in Volume 2. Possibilistic networks are also useful for preference modeling (Ben Amor et al. 2018) (see also chapter "Compact Representation of Preferences" in this volume).

There exist different variants of possibilistic logic where a logical formula is, in particular, associated with lower bounds of (weak) possibility measures. They can express different forms of ignorance by asserting that two opposite events are both at least somewhat possible). Other kinds of weights can be attached to logical formulas such as time slots where one is more or less certain that the formula is true, or subsets of sources or agents that are certain to various extents that the formula is true; see (Dubois and Prade 2004, 2014) for references. For further developments on multiple agent possibilistic logic, see (Belhadi et al. 2013).

Generalized possibilistic logic Another type of extension allows for negations or disjunctions of basic possibilistic formulas (and not only conjunctions as in standard possibilistic logic). It then results into a two-tiered logic, named "*generalized possibilistic logic*" (GPL) (Dubois et al. 2017c), where Boolean connectives can be placed inside or outside basic possibilistic formulas. Its semantics is in terms of *subsets* of possibility distributions. Indeed, elementary formulas in the logic GPL encode lower or upper bounds on the necessity or the possibility of logical formulas. GPL is both a generalization of the minimal epistemic logic MEL (Banerjee and Dubois 2014) (where weights are only 1 or 0), and of standard possibilistic logic, in full agreement with possibility theory. GPL has been axiomatized and inference in GPL has been shown sound and complete w.r.t. semantics in terms of subsets of possibility distributions.

GPL appears as a powerful unifying framework for various knowledge representation formalisms. Among others, logics of comparative certainty, and reasoning about explicit ignorance can be modeled in GPL. There also exists a close connection between GPL and various existing knowledge representation formalisms. It includes possibilistic logic with partially ordered formulas (Touazi et al. 2015), the logic of conditional assertions of Kraus et al. (1990), three-valued logics (Ciucci and Dubois 2013), and the 5-valued "equilibrium logic" of Pearce (2006) as well as answer set programming (Dubois et al. 2012) (see chapter "Logic Programming" in Volume 2). More specifically, the intended meaning of answer-set programs can be made more explicit through a translation in GPL (using a 3-level scale for the possibility distributions).

Lastly, in the same way as imprecise probabilities (see next chapter in this volume) are of interest, one may think of imprecise possibilities. In that respect, the following result is particularly worth noticing: any capacity (i.e., any monotonic increasing set

function) on a finite domain can be characterized by a set of possibility mesures; then capacities offer a semantics to non normal modal logics (useful for the handling of paraconsistency) (Dubois et al. 2015b), and it may provide a unifying framework for multiple source information processing in the spirit of Belnap logic.

4.2.2 Inconsistency and Non Monotonic Reasoning

An important feature of possibilistic logic is its ability to deal with inconsistency. The inconsistency level $inc(B)$ of a possibilistic base B is defined as

$$inc(B) = \max\{\alpha |\ B \vdash (\bot, \alpha)\}.$$

No formula whose level is strictly greater than $inc(B)$ contributes to inconsistency. It can be shown that $1 - inc(B)$ is the height $h(\pi_B)$ of π_B, defined by $h(\pi_B) = \max_s \pi_B(s)$ (π_B being the possibility distribution induced by B). Moreover, $inc(B) = 0$ if and only if the set of logical formulas appearing in B, irrespective of the weights, is consistent in the classical sense. All the formulas in B whose level is smaller or equal to $inc(B)$ are ignored in the standard possibilistic inference mechanism; they are said to be "drowned". However, there exist other extensions of possibilistic inference that take into account formulas at the inconsistency level or below, especially those not involved in any inconsistent subset of formulas (called free formulas), see (Benferhat et al. 1999a) for a complete overview of these inferences.

The application of default rules having potential exceptions (for instance, "birds fly") to particular situations (e.g., "Tweety is a bird") about which information is incomplete, may lead to tentative conclusions (here, "Tweety flies") that become inconsistent with the new conclusions obtained when more information becomes available on such particular situations (e.g., "Tweety is a penguin"). The non monotonic nature of conditional qualitative possibility enables us to handle this problem. Indeed it allows $N(B \mid A) > 0$ and $N(\overline{B} \mid A \cap A') > 0$ to simultaneously hold, i.e., the arrival of the piece of information A' may lead to reject a previously accepted proposition B in the context where we only knew A.

Indeed, a default rule "if A_i then generally B_j" can be represented by the possibilistic constraint $\Pi(B_j \cap A_i) > \Pi(\overline{B_j} \cap A_i)$ expressing that it is more possible to have B_i true than B_i false in the context where A_i is true. A base of default rules is then represented by a set of such constraints, which in turn determines a set of possibility measures that satisfy them. From such a rule base, two types of inference are natural in order to deduce new rules applicable to the situation where one *exactly* knows A (i.e., the rules of the form "if A then generally B", which will allow us to conclude B (tentatively) in this situation).

A first type of inference, which is cautious, requires that the inequality constraint $\Pi(A \cap B) > \Pi(A \cap \overline{B})$ associated with $B|A$ be satisfied by par *all* possibility measures that agree with the set of constraints (supposed to be consistent) associated with the set of default rules. A second, bolder, inference only considers the largest

(the least specific) possibility distribution that is a solution of the latter constraints (it can be shown that this distribution is unique when it exists). It can be established that the first inference relation basically corresponds the so-called preferential inference (system P (Kraus et al. 1990)) obeying basic postulates for non monotonic plausible inference (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume), while the second one is nothing but the "rational closure" inference of Lehmann and Magidor (1992). These two types of inference can be justified also using other semantics such as conditional objects (Dubois and Prade 1994), infinitesimal probabilities, systems Z and Z^+ (Pearl 1990; Goldszmidt and Pearl 1991), conditional modal logic (Boutilier 1994), Halpern's plausibility measures (Halpern 2001); see Benferhat et al. (1997) for an overview and references. There are also semantics in terms of big-stepped probabilities (Benferhat et al. 1999b), or conditional probabilities in De Finetti's sense (Coletti and Scozzafava 2002). In this latter case the rule "if A then generally B" simply corresponds to a constraint $Prob(B|A) = 1$ where $Prob(B|A)$ still makes sense when $Prob(A) = 0$ (0 does not mean impossible here, but rather something as "negligible at first glance"), thanks to a prioritized handling of constraints induced by a partitioning of the set of interpretations (Biazzo et al. 2002). The setting of possibilistic logic thus enables us to practically handle a form of default reasoning (Benferhat et al. 1998), as well as reasoning from qualitative uncertain information; il is even possible to combine both (Dupin de Saint-Cyr and Prade 2008).

Belief revision theory (Gärdenfors 1988) (see chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" in this volume), which is closely related to non monotonic reasoning, relies on the notion of epistemic entrenchment, used by the revision process for ordering the way pieces of information are called into question. It is interesting to note that an epistemic entrenchment relation is nothing but a qualitative necessity relation (Dubois and Prade 1991) (whose unique counterpart on a totally ordered scale is a necessity measure (Dubois 1986)). Moreover the possibilistic setting can make sense of the intuition that propositions in the belief base that are independent of the input information should remain after revision (Dubois et al. 1999). Besides, updating and revision can be combined, in the style of Kalman (1960) filtering, in the qualitative setting of possibilistic logic (Benferhat et al. 2000).

Let us also mention a model of causal *ascription* where an agent, in the presence of a sequence of events that took place, is supposed to assert causal relations between some of these events on the basis of his beliefs on the normal course of things (Bonnefon et al. 2008). The normal course of things is represented by default rules (obeying system P postulates). The possibilistic framework for causal ascription favors "abnormal" events as potential causes which may be adopted by the agent; a detailed comparison of this approach with the probabilistic modeling of causation is presented in Bonnefon et al. (2012). The reader is referred to chapter "A Glance at Causality Theories for Artificial Intelligence" in this volume for an overview of approaches to causality modeling.

4.2.3 Possibility Theory and Formal Concept Analysis

Formal concept analysis (FCA) is a knowledge representation formalism at the basis of a data mining methodology (see chapters. "Designing Algorithms for Machine Learning and Data Mining" and "Formal Concept Analysis: From Knowledge Discovery to Knowledge Processing" of Volume 2). It provides a theoretical setting for learning hierarchies of concepts. Strong similarities between this representation framework and possibility theory have been pointed out in the last decade (and also to some extent with rough set theory (Pawlak and Skowron 2007)). This is the reason for the presence of this – maybe unexpected – subsection in this chapter.

In FCA (Barbut and Montjardet 1970; Ganter and Wille 1999), one starts with a binary relation $\mathscr{R}$, called *formal context*, between a set of objects $\mathscr{O}$ and a set of properties $\mathscr{P}$; $x\mathscr{R}y$ means that object x possesses property y. Given an object x and a property y, let $R(x) = \{y \in \mathscr{P} \mid x\mathscr{R}y\}$ be the set of properties possessed by object x and let $R(y) = \{x \in \mathscr{O} \mid x\mathscr{R}y\}$ be the set of objects having property y. In FCA correspondences are defined between the sets $2^{\mathscr{O}}$ and $2^{\mathscr{P}}$. These correspondences are Galois derivation operators. The Galois operator at the basis of FCA, here denoted by $(.)^{\Delta}$ (for a reason made clear in the following), enables us to describe the set of properties satisfied by *all* the objects in $X \subseteq \mathscr{O}$ as

$$X^{\Delta} = \{y \in \mathscr{P} \mid \forall x \in \mathscr{O}\ (x \in X \Rightarrow x\mathscr{R}y)\} = \{y \in \mathscr{P} \mid X \subseteq R(y)\} = \bigcap_{x \in X} R(x).$$

In a dual manner, the set of objects satisfying all the properties in Y is defined by

$$Y^{\Delta} = \{x \in \mathscr{O} \mid \forall y \in \mathscr{P}\ (y \in Y \Rightarrow x\mathscr{R}y)\} = \{x \in \mathscr{O} \mid Y \subseteq R(x)\} = \bigcap_{y \in Y} R(y).$$

The pair of operators $((.)^{\Delta}, (.)^{\Delta})$ applied respectively to $2^{\mathscr{O}}$ and $2^{\mathscr{P}}$ constitutes a Galois connection that induces formal concepts. Namely, a *formal concept* is a pair (X, Y) such that

$$X^{\Delta} = Y \text{ and } Y^{\Delta} = X.$$

In other words, X is a maximal set of objects, and Y a maximal set of properties such that each object in X satisfies all the properties in Y. Then the set X (resp. Y) is called *extension* (resp. *intension*) of the concept. In an equivalent way, (X, Y) is a formal concept if and only if it is a maximal pair for the inclusion

$$X \times Y \subseteq \mathscr{R}.$$

The set of all formal concepts is naturally equipped with an order relation (denoted by $\preccurlyeq$) and defined by: $(X_1, Y_1) \preceq (X_2, Y_2)$ iff $X_1 \subseteq X_2$ (or $Y_2 \subseteq Y_1$). This set equipped with the order relation $\preccurlyeq$ forms a complete lattice. Then association rules between

properties can be found by exploiting this lattice, see Guigues and Duquenne (1986), Pasquier et al. (1999).

Note that $X^{\Delta} = \bigcap_{x \in X} R(x)$ mirrors the definition of a guaranteed possibility measure $\Delta(F) = \min_{s \in F} \pi(s)$ (where π is a possibility distribution), changing L into 2^Y and π into a set-valued map (R is viewed as a Boolean lattice-valued map). On the basis of this parallel with *possibility theory*, other operators can be introduced (Dubois and Prade 2012). Namely, the possibility operator (denoted by $(.)^{\Pi}$) and its dual necessity operator (denoted by $(.)^{N}$), as well as the operator $(.)^{\nabla}$ dual to the operator $(.)^{\Delta}$ on which FCA is based. They are defined as follows:

- X^{Π} is the set of properties satisfied by at least one object in X:

$$X^{\Pi} = \{y \in \mathscr{P} \mid \exists x \in X,\ x \mathscr{R} y\} = \{y \in \mathscr{P} \mid X \cap R(y) \neq \emptyset\} = \bigcup_{x \in X} R(x);$$

- X^{N} is the set of properties that only the objects in X have:

$$X^{N} = \{y \in \mathscr{P} \mid \forall x \in \mathscr{O}\ (x \mathscr{R} y \Rightarrow x \in X)\} = \{y \in \mathscr{P} \mid R(y) \subseteq X\} = \bigcap_{x \notin X} \overline{R}(x),$$

  where $\overline{R}(x)$ is the set of properties that x has not;
- X^{∇} is the set of properties that are not satisfied by at least one object outside X:

$$X^{\nabla} = \{y \in \mathscr{P} \mid \exists x \in \overline{X},\ x \overline{\mathscr{R}} y\} = \{y \in \mathscr{P} | R(y) \cup X \neq \mathscr{O}\} = \bigcup_{x \notin X} \overline{R}(x).$$

The operators Y^{Π}, Y^{N}, Y^{∇} are obtained similarly. While the equalities $X^{\nabla} = Y$ and $Y^{\nabla} = X$ provide another characterization of usual formal concepts, it can be shown that pairs (X, Y) such that $X^{N} = Y$ and $Y^{N} = X$ (equivalently, $X^{\Pi} = Y$ and $Y^{\Pi} = X$) characterize independent sub-contexts (i.e., that have no object or property in common) inside the initial context (Dubois and Prade 2012). The pairs (X, Y) such that $X^{N} = Y$ and $Y^{N} = X$ are such that:

$$(X \times Y) \cup (\overline{X} \times \overline{Y}) \supseteq \mathscr{R}.$$

It can be checked that the four sets X^{Π}, X^{N}, X^{Δ}, X^{∇} are complementary pieces of information, all necessary for a complete analysis of the situation of X in the formal context $\mathscr{K} = (\mathscr{O}, \mathscr{P}, \mathscr{R})$. In practice, one supposes that both $R(x) \neq \emptyset$ and $R(x) \neq \mathscr{P}$, i.e., each object possesses at least one property in $\mathscr{P}$, but no object possesses all the properties in $\mathscr{P}$. Under this hypothesis of *bi-normalisation*, the following inclusion relation holds: $R^{N}(Y) \cup R^{\Delta}(Y) \subseteq R^{\Pi}(Y) \cap R^{\nabla}(Y)$, which is a counterpart of a relation that holds as well in possibility theory (provided that distributions π and $1 - \pi$ are both normalized).

Finally, let us also mention that there exists an extension of FCA to graded properties (Belohlavek 2002), as well as an extension to formal contexts displaying incom-

plete or uncertain information (Burmeister and Holzer 2005; Ait-Yakoub et al. 2017). Another extension deals with the capability of associating objects no longer with simple properties, but with structured descriptions, possibly imprecise, or with logical descriptions, thanks to so-called *patron structures* (Ganter and Kuznetsov 2001; Ferré and Ridoux 2004). They remain in agreement with the possibilistic paradigm (Assaghir et al. 2010).

4.3 Quantitative Possibility and Bridges to Probability

In the quantitative version of possibility theory, it is natural to relate possibility and probability measures. It can be done in several independent ways. In the following, we outline the three main ones: namely, a possibility distribution can be viewed as a likelihood function in non-Bayesian statistics, possibility (resp. necessity) degrees of events can be viewed either as upper (resp. lower) probability bounds, or as a suitable transformation of exponents of infinitesimal probabilities.

4.3.1 Possibility Distributions as Likelihood Functions

The idea of casting likelihood functions inside the framework of possibility theory was suggested by Smets (1982), but it has roots in considerations relating statistical inference and consonant belief functions (another name for necessity measures) in Shafer (1976)'s book; see also (Denœux 2014) on this topic. The connection was formalized in (Dubois et al. 1997), and further studied in the coherence framework of De Finetti in (Coletti and Scozzafava 2003). Consider an estimation problem where the value of a parameter $\theta \in \Theta$ that governs a probability distribution $P(\cdot \mid \theta)$ on S is to be determined from data. Suppose the obtained data is described by the information item A. The function $\ell(\theta) = P(A \mid \theta), \theta \in \Theta$ is not a probability distribution, it is a *likelihood* function: a value θ is all the more plausible as $P(A \mid \theta)$ is greater, while this value can be ruled out if $P(A \mid \theta) = 0$ (in practice, less that a small relevance threshold). Such a function is often renormalized so that its maximal value is 1, since a likelihood function is defined up to a positive multiplicative constant. There are some good reasons why one may see $\ell(\theta)$ as a degree of possibility of θ, and let $\pi(\theta) = P(A \mid \theta)$ (up to renormalizing). First, it can be checked that, in the absence of prior probability on $\Theta, \forall B \subseteq \Theta$, $P(A \mid B)$ is upper and lower bounded as follows:

$$\min_{\theta \in B} P(A \mid \theta) \leq P(A \mid B) \leq \max_{\theta \in B} P(A \mid \theta)$$

It suggests that we can apply the maxitivity axiom to get an optimistic estimate of $P(A \mid B)$ from $\{P(A \mid \theta), \theta \in B\}$. However, insofar as $\ell(b)$ is the likelihood of $\theta = b$, and we extend it to all subsets B of Θ, we should have that $\ell(B) \geq \ell(b)$, for all $b \in B$. Hence, in the absence of prior probability, we can identify $\ell(B)$ as

a possibility measure with distribution $\pi(\theta) = P(A \mid \theta)$ (Coletti and Scozzafava 2003). Considering the lower bound of $P(A \mid B)$ would yield a guaranteed possibility measure.

However, note that under this view, possibility degrees are known in relative values, which means that not all basic notions of possibility theory apply (e.g., comparing the informativeness of π and π' using fuzzy set inclusion, by checking if $\pi \leq \pi'$ becomes questionable).

4.3.2 Possibility as Upper Probability

Alternatively, possibility degrees valued on [0, 1] viewed as an absolute scale can be exactly defined as upper probability bounds as Zadeh (1978) had the intuition from the start. The generation process can be described as follows: consider an increasing sequence of nested sets $E_1 \subset E_2 \subset, \ldots, \subset E_k$ and let $\alpha_1 \leq \alpha_2 \leq, \ldots, \leq \alpha_k \in [0, 1]$, such that α_i is a lower bound on the probability $P(E_i)$. This type of information is typically provided by an expert estimating a quantity v by means of set E_k with confidence α_k that E_k contains v. Consider the probability family $\mathscr{P} = \{P : P(E_i) \geq \alpha_i, \forall i = 1, \ldots, k\}$. It is easy to check (Dubois and Prade 1992) that the function $P_*(A) = \inf_{P \in \mathscr{P}} P(A)$ is a necessity measure and the function $P^*(A) = \sup_{P \in \mathscr{P}} P(A)$ is a possibility measure induced by the possibility distribution:

$$\forall s \in S, \quad \pi(s) = \min_{i=1,\ldots,k} \max(E_i(s), 1 - \alpha_i). \tag{16}$$

where $E_i(s) = 1$ if $s \in E_i$ and 0 otherwise. See de Cooman and Aeyels (1999) for the extension of this result to infinite settings. Each pair (E_k, α_k), made of a set and its confidence level is encoded by the possibility distribution $\max(E_i(s), 1 - \alpha_i)$, where $1 - \alpha_i$ is an upper bound on the probability that $v \notin E_k$. Equation (16) just performs the conjunction of these local distributions. It is clear that π is a very concise encoding of the probability family $\mathscr{P}$. Conversely, the (convex set) of probability measures encoded by a possibility distribution π can be retrieved as

$$\mathscr{P}(\pi) = \{P, P(A) \leq \Pi(A), \forall A \text{ measurable}\} = \{P, P(A) \geq N(A), \forall A \text{ mesurable}\},$$

and it can be checked that $\Pi(A) = \sup_{P \in \mathscr{P}(\pi)} P(A)$. In the case where the sets E_i are not nested, the above formula (which is in agreement with possibilistic logic semantics of Sect. 4.2.1) only yields an approximation of the probability family $\mathscr{P}$; better approximations can be obtained by means of pairs of possibility distributions enclosing $\mathscr{P}$ (Destercke et al. 2008). This view of possibility measures cast them in the landscape of imprecise probability theory studied in the next chapter of this volume.

Nested shortest dispersion intervals can be obtained from a given probability distribution (or density) p, letting $E_\alpha = \{s : p(s) \geq \alpha\}$, and $\alpha_\alpha = P(E_\alpha)$. The obtained possibility distribution, that covers p as tightly as possible, is called opti-

mal probability-possibility transform of p (Dubois et al. 2004) and is instrumental for comparing probability distributions in terms of their peakedness or entropies (by comparing their possibility transforms in terms of relative specificity) (Dubois and Hüllermeier 2007).

4.3.3 Possibility as Infinitesimal Probability

Ranking functions, originally called ordinal conditional functions (OCF), have been proposed by Spohn (1988, 2012) to represent the notion of belief in a setting that is basically equivalent to possibility theory, but for the direction and nature of its value scale. Each state of the world $s \in S$ is assigned a degree $\kappa(s)$ not in $[0, 1]$, but in the set of non-negative integers $\mathbb{N}$, (sometimes even ordinals). The convention for ranking functions is opposite to the one in possibility theory, since the smaller $\kappa(s)$ the more possible s. It is more in agreement with a degree of potential surprise suggested by Shackle (1961): $\kappa(s) = +\infty$ means that s is impossible, while $\kappa(s) = 0$ means that nothing opposes to s being the true state of the world. Set functions expressing disbelief, similar to possibility measures, are then built in the same style as Shackle (1961):

$$\kappa(A) = \min_{s \in A} \kappa(s) \text{ and } \kappa(\emptyset) = +\infty.$$

More specifically, Spohn (1990) interprets $\kappa(A)$ as the integer exponent of an infinitesimal probability $P(A) = \varepsilon^{\kappa(A)}$, which is indeed in agreement with the union-minitivity property $\kappa(A \cup B) = \min(\kappa(A), \kappa(B))$ of ranking functons.

Conditioning is defined by Spohn (1988) as follows:

$$\kappa(s \mid B) = \begin{cases} \kappa(s) - \kappa(B) & \text{si } s \in B \\ +\infty & \text{sinon} \end{cases}$$

It is obvious that $\kappa(s \mid B)$ is the exponent of the infinitesimal conditional probability $P(s \mid B) = \varepsilon^{\kappa(s)} / \varepsilon^{\kappa(B)}$.

Casting ranking functions in possibility theory is easy, due to the following transformations (Dubois and Prade 1991):

$$\pi_\kappa(s) = 2^{-\kappa(s)}, \Pi_\kappa(A) = 2^{-\kappa(A)}.$$

As a consequence possibility distributions π_κ and functions Π_κ take values on a subset of rational numbers in $[0, 1]$. Function Π_κ is indeed a possivility measure since

$$\Pi_\kappa(A \cup B) = 2^{-\min(\kappa(A), \kappa(B))} = \max(\Pi_\kappa(A), \Pi_\kappa(B)).$$

Moreover, for the conditional ranking function one obtains $\forall s$,

$$\pi_{\kappa(s|B)} = 2^{-\kappa(s)+\kappa(B)} = \frac{2^{-\kappa(s)}}{2^{-\kappa(B)}} = \frac{\pi_\kappa(s)}{\Pi_\kappa(B)},$$

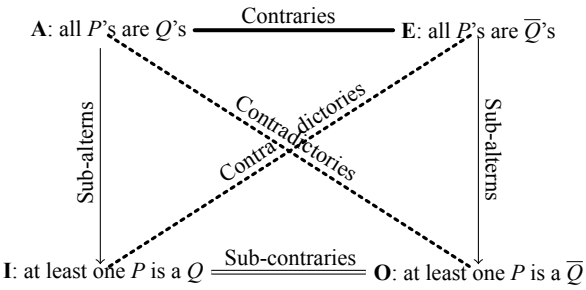


Fig. 1 Square of opposition

which is the product-based conditioning of possibility theory. The converse (logarithmic) transformation of a possibility distribution into a ranking function is only possible if it maps real numbers to non-negative integers. More on the comparison between possibility theory and ranking functions can be found in (Dubois and Prade 2016).

Note that this approach is often presented as qualitative while it is a numerical one. In some applications, or when modeling expert opinions, it may be more convenient to describe degrees of (dis)belief by means of integers rather than by real numbers in [0, 1]. However it is easier to introduce intermediary grades with a continuous scale. The integer scale of ranking functions has been used recently by Kern-Isberner and Eichhorn (2014) to encode non-monotonic inferences and applied in (Eichhorn and Kern-Isberner 2015) to belief networks.

5 The Cube of Opposition: A Structure Unifying Representation Frameworks

Many knowledge representation formalisms, although they look quite different at first glance and aim at serving diverse purposes, share a common structure where involutive negation plays a key role. This structure can be summarized under the form of a square or a cube of opposition. This in particular true for frameworks able to represent incomplete information. It can be observed that the properties of non empty intersection and of inclusion related by negation are at the basis of possibility theory, formal concept analysis, as well as rough set theory. It is still true for belief functions presented in the next chapter in this volume. This section first introduces the square and the cube of opposition, and indicates the formalisms to which it applies.

The traditional square of opposition (Parsons 2008), which dates back to Aristotle time, is built with universally and existentially quantified statements in the following way. Consider four statement of the form (**A**): "all P's are Q's", (**O**): "at least one P is not a Q", (**E**): "no P is a Q", and (**I**): "at least one P is a Q". They can be displayed on a square whose vertices are traditionally denoted by the letters **A**, **I** (affirmative half) and **E**, **O** (negative half), as pictured in Fig. 1 (where $\overline{Q}$ stands for "not Q").

As can be checked, noticeable relations hold in the square provided that there a non empty set of P's to avoid existential import problems:

1. **A** and **O** (resp. **E** and **I**) are the negation of each other;
2. **A** entails **I**, and **E** entails **O** (it is assumed that there is at least one P);
3. **A** and **E** cannot be true together;
4. **I** and **O** cannot be false together.

Another classical example of such a square is obtained with modal logic operators by taking **A** as $\Box p$, **E** as, $\Box\neg p$, **I** as $\Diamond p$, and **O** as $\Diamond\neg p$. This structure, largely forgotten with the advent of modern logic after G. Boole, was rediscovered by Blanché (1966) and then by Béziau (2003) who both advocate its interest. In particular, Blanché noticed that adding two vertices **U** and **Y** defined respectively as the disjunction of **A** and **E**, and as the conjunction of **I** and **O**, leads to a hexagon that includes three squares of opposition in the above sense. Such a hexagon is obtained each time we start with three mutually exclusive statements, such as **A**, **E**, and **Y**, and it turns out that this structure is often encountered when representing relationships between concepts on the same domain (e.g., deontic notions such as permission, obligation, interdiction, etc.).

Switching to first order logic notations (e.g., **A** becomes $\forall x,\ P(x) \rightarrow Q(x)$), and negating the predicates, i.e., changing P into $\neg P$, and Q in $\neg Q$ leads to another similar square of opposition **aeoi**, where we also assume that the set of "not-P's" is non-empty. Altogether, we obtain eight statements that may be organized in what may be called a *cube of opposition* (Reichenbach 1952). The front facet and the back facet of the cube are traditional squares of opposition, and the two facets are related by entailments.

Such a structure can be extended to graded notions (Ciucci et al. 2016), using an involutive negation such as $1-(\cdot)$, and where the mutual exclusiveness of **A** and **E** translates into a sum of degrees less or equal to 1, while entailments are translated by inequalities between degrees (in agreement with residuated implications). An example of a graded cube is given by possibility theory. Indeed, assuming a normalized possibility distribution $\pi : S \rightarrow [0, 1]$, and also assuming that $1-\pi$ is normalized (i.e., $\exists s \in S, \pi(s) = 0$), we obtain a cube of opposition on Fig. 2, linking $\Pi(A)$, $N(A)$, $\Delta(A)$, $\nabla(A)$, $\Pi(\overline{A})$, $N(\overline{A})$, $\Delta(\overline{A})$, and $\nabla(\overline{A})$. The front and back facets form two squares of opposition, while the side facets express a different property, namely inequalities such as $\min(\Pi(A), \nabla(A)) \geq \max(N(A), \Delta(A))$. Since these set functions rely on ideas of graded inclusion and degrees of non-empty intersections, the fact that they fit with a graded structure of cube of opposition should not be too surprizing.

In fact, the structure of cube of opposition is quite general. As noticed by Ciucci et al. (2016), any binary relation R on a Cartesian product $X \times Y$ (one may have $Y = X$) gives birth to a cube of opposition, when applied to a subset. Indeed, we assume $R \neq \emptyset$. Let $R(x) = \{y \in Y \mid (x, y) \in R\}$. $\overline{R}$ denotes the complementary relation ($(x, y) \in \overline{R}$ iff $(x, y) \notin R$), and R^t the transposed relation ($(y, x) \in R^t$ if and only if $(x, y) \in R$); let $R(y) = \{x \in X \mid (x, y) \in R\} = R^t(y)$. Moreover, it is assumed that $\forall x,\ R(x) \neq \emptyset$, which means that the relation R is *serial*, namely $\forall x, \exists y$

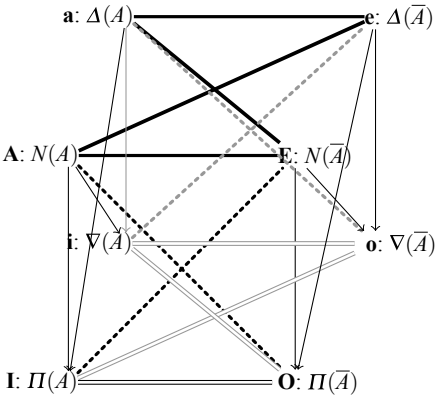


Fig. 2 Cube of opposition of possibility theory

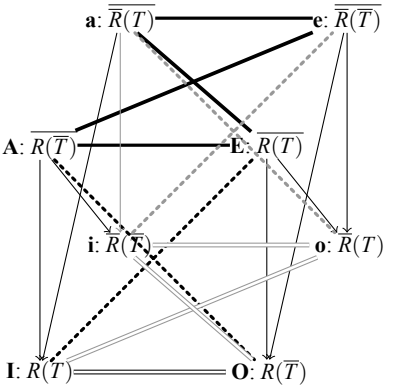


Fig. 3 Cube induced by a relation R and a subset T

such that $(x, y) \in R$. Similarly, R^t is also supposed to be serial, i.e., $\forall y,\ R(y) \neq \emptyset$, as well as $\overline{R}$ and its transpose, i.e. $\forall x,\ R(x) \neq Y$ and $\forall y,\ R(y) \neq X$.

Let T be a subset of Y and $\overline{T}$ its complement. We assume $T \neq \emptyset$ and $T \neq Y$. The composition is defined in the usual way $R(T) = \{x \in X \mid \exists t \in T, (x, t) \in R\}$. From the relation R and the subset T, one can define the four following subsets of X (and their complements):

$$R(T) = \{x \in X \mid T \cap R(x) \neq \emptyset\} \tag{17}$$

$$\overline{R(\overline{T})} = \{x \in X \mid R(x) \subseteq T\} \tag{18}$$

$$\overline{\overline{R}(T)} = \{x \in X \mid T \subseteq R(x)\} \tag{19}$$

$$\overline{R}(\overline{T}) = \{x \in X \mid T \cup R(x) \neq X\}. \tag{20}$$

These four subsets and their complements can be nicely organized into a cube of opposition (Fig. 3). Some of the required conditions for the cube hold thanks to seriality (which plays the role of normalization in possibility theory).

The front facet of the cube fits well with the modal logic reading of the square where R is viewed as an accessibility relation, and T as the set of models of a proposition p. Indeed, $\Box p$ (resp. $\Diamond p$) is true in world x means that p is true at *every* (resp. at *some*) possible world accessible from x; this corresponds to $\overline{R(\overline{T})}$ (resp. $R(T)$) which is the set of worlds where $\Box p$ (resp. $\Diamond p$) is true.

Other than the semantics of modal logics, there are a number of AI formalisms that exploit a relation and to which the cube of opposition of Fig. 3 applies: formal concept analysis, as seen in Sect. 4.2.3, rough sets induced by an equivalence relation (see Sect. 2.6), or abstract argumentation based on an attack relation between arguments (Amgoud and Prade 2013). Graded squares or cubes also apply to belief functions (Dubois et al. 2015a) and to upper and lower probabilities (Pfeifer and Sanfilippo 2017) presented in the next chapter in this volume, as well as to aggregation functions such as Sugeno integrals (Dubois et al. 2015a) used in multiple criteria aggregation and qualitative decision theory, or yet Choquet integrals (Dubois et al. 2017b), both presented in chapter "Multicriteria Decision Making" in this volume.

This common structure is deeply related to the interplay of three negations as revealed by the relational cube. In contrast the square and the cube collapse to a segment in the case of probabilities since they are autodual.

The cube of opposition lays bare common features underlying many knowledge representation formalisms. It exhibits fruitful parallelisms between them, which may even lead to highlight some missing components present in one formalism and currently absent from another.

6 Conclusion

In this chapter, we have tried to show that while probability theory properly captures uncertainty due to the randomness of precisely observed phenomena, the representation of uncertainty due to incomplete information requires a different setting having roots in classical and modal logics, where incompleteness is a usual feature. The corresponding uncertainty framework is possibility theory, which allows for a qualitative representation of uncertainty as well as a quantitative one. It has been shown that numerical possibility theory is appropriate provided that the available information items, although imprecise, are consonant, i.e., do not contradict each other. The joint handling of imprecise and possibly conflicting information items require joint extensions of probability and quantitative possibility theory studied in the next chapter.

References

Abadi M, Halpern JY (1994) Decidability and expressiveness for first-order logics of probability. Inf Comput 112(1):1–36

Ait-Yakoub Z, Djouadi Y, Dubois D, Prade H (2017) Asymmetric composition of possibilistic operators in formal concept analysis: Application to the extraction of attribute implications from incomplete contexts. Int J Intell Syst 32(12):1285–1311

Amgoud L, Prade H (2013) A formal concept view of abstract argumentation. In: van der Gaag LC (ed) Proceedings of the 12th European conference on symbolic and quantitative approaches to reasoning with uncertainty (ECSQARU 2013). Utrecht, pp 1–12, Springer, vol 7958 in LNCS

Assaghir Z, Kaytoue M, Prade H (2010) A possibility theory-oriented discussion of conceptual pattern structures. In: Deshpande A, Hunter A (eds) Proceedings of international conference on scalable uncertainty management (SUM'10). Toulouse, pp 70–83 Sept. 27–29, Springer, vol 6379 in LNCS

Bacchus F (1991) Representing and reasoning with probabilistic knowledge: a logical approach to probabilities. MIT Press, Cambridge

Banerjee M, Dubois D (2014) A simple logic for reasoning about incomplete knowledge. Int J Approx Reason 55:639–653

Barbut M, Montjardet B (1970) Ordre et Classification: Algèbre et Combinatoire. Hachette

Belhadi A, Dubois D, Khellaf-Haned F, Prade H (2013) Multiple agent possibilistic logic. J Appl Non-Class Log 23(4):299–320

Belnap ND (1977a) How a computer should think. In: Contemporary aspects of philosophy. Oriel, Boston, pp 30–56

Belnap ND (1977b) A useful four-valued logic. In: Epstein G (ed) Modern uses of multiple-valued logic. Reidel, pp 8–37

Belohlavek R (2002) Fuzzy relational systems: foundations and principles. Kluwer

Ben Amor N, Benferhat S, Dubois D, Mellouli K, Prade H (2002) A theoretical framework for possibilistic independence in a weakly ordered setting. Int J Uncertain, Fuzziness Knowl-Based Syst 10(2):117–155

Ben Amor N, Dubois D, Gouider H, Prade H (2018) Possibilistic preference networks. Inf Sci 460–461:401–415

Benferhat S, Dubois D, Prade H (1997) Nonmonotonic reasoning, conditional objects and possibility theory. Artif Intell 92:259–276

Benferhat S, Dubois D, Prade H (1998) Practical handling of exception-tainted rules and independence information in possibilistic logic. Appl Intell 9:101–127

Benferhat S, Dubois D, Prade H (1999a) An overview of inconsistency-tolerant inferences in prioritized knowledge bases. In: Dubois D, Prade H, Klement E (eds) Fuzzy sets, logic and reasoning about knowledge, applied logic series, vol 15. Kluwer, Dordrecht, pp 395–417

Benferhat S, Dubois D, Prade H (1999b) Possibilistic and standard probabilistic semantics of conditional knowledge bases. J Log Comput 9:873–895

Benferhat S, Dubois D, Prade H (2000) Kalman-like filtering in a possibilistic setting. In: Horn W (ed) Proceedings of 14th European conference on artificial intelligence (ECAI'00). Springer, Berlin, pp 8–12 Aug. 20–25

Benferhat S, Dubois D, Garcia L, Prade H (2002) On the transformation between possibilistic logic bases and possibilistic causal networks. Int J Approx Reason 29:135–173

Benferhat S, Dubois D, Kaci S, Prade H (2008) Modeling positive and negative information in possibility theory. Int J Intell Syst 23:1094–1118

Benferhat S, Kaci S (2003) Logical representation and fusion of prioritized information based on guaranteed possibility measures: application to the distance-based merging of classical bases. Artif Intell 148:291–333

Benferhat S, Smaoui S (2011) Inferring interventions in product-based possibilistic causal networks. Fuzzy Sets Syst 169(1):26–50

Béziau JY (2003) New light on the square of oppositions and its nameless corner. Log Investig 10:218–233

Biazzo V, Gilio A, Lukasiewicz T, Sanfilippo G (2002) Probabilistic logic under coherence, model-theoretic probabilistic logic, and default reasoning in system P. J Appl Non-Class Log 12(2):189–213

Blanché R (1966) Structures Intellectuelles Essai sur l'Organisation Systématique des Concepts. Vrin, Paris
Bolt JH, van der Gaag LC, Renooij S (2005) Introducing situational signs in qualitative probabilistic networks. Int J Approx Reason 38:333–354
Bonnefon JF, Da Silva Neves R, Dubois D, Prade H (2008) Predicting causality ascriptions from background knowledge: model and experimental validation. Int J Approx Reason 48:752–765
Bonnefon JF, Da Silva Neves R, Dubois D, Prade H (2012) Qualitative and quantitative conditions for the transitivity of perceived causation - theoretical and experimental results. Ann Math Artif Intell 64:311–333
Boutilier C (1994) Modal logics for qualitative possibility theory. Int J Approx Reason 10(2):173–201
Buchanan BG, Shortliffe EH (eds) (1984) Rule-based expert systems. Addison- Wesley, Reading
Burmeister P, Holzer R (2005) Treating incomplete knowledge in formal concepts analysis. In: Ganter B (ed) Formal concept analysis : foundations and applications, LNCS, vol. 3626. Springer, Berlin, pp 114–126
Cayrac D, Dubois D, Prade H (1996) Handling uncertainty with possibility theory and fuzzy sets in a satellite fault diagnosis application. IEEE Trans on Fuzzy Syst 4(3):251–269
Cayrol M, Farreny H, Prade H (1982) Fuzzy pattern matching. Kybernetes 11(2):103–116
Cheeseman P (1988) An inquiry into computer understanding. Computational Intelligence 4:58–66, with comments by R. Aleliunas, A. Bundy, N. C. Dalkey, A. P. Dempster, D. Dubois and H. Prade, M. L. Ginsberg, R. Greiner, P. J. Hayes, D. Israel, L. Kanal and D. Perlis, H. Kyburg, D. McDermott, D. L. McLeish, C. G. Morgan, E. Neufeld and D. Poole, J. Pearl, L. Rendell, E. H. Ruspini, L.K. Schubert, G. Shafer, D. J. Spiegelhalter, R. R. Yager, L. A. Zadeh (67–128), and a reply by P. Cheeseman (129–142)
Choquet G (1953) Theory of capacities. Ann. l'Institut Fourier 5:131–295
Ciucci D, Dubois D (2013) A modal theorem-preserving translation of a class of three-valued logics of incomplete information. J Appl Non-Class Log 23(4):321–352
Ciucci D, Dubois D, Prade H (2016) Structures of opposition induced by relations - the Boolean and the gradual cases. Ann Math Artif Intell 76(3–4):351–373
Coletti G, Scozzafava R (2002) Probabilistic logic in a coherent setting. Kluwer Academic Publication, Dordrecht
Coletti G, Scozzafava R (2003) Coherent conditional probability as a measure of uncertainty of the relevant conditioning events. In: Nielsen TD, Zhang NL (eds)) Proceedings of 7th European conference on symbolic and quantitative approaches to reasoning with uncertainty (ECSQARU'03), vol. 2711. LNCS, Springer, pp 407–418
Coletti G, Vantaggi B (2006) Possibility theory: conditional independence. Fuzzy Sets Syst 157(11):1491–1513
Coletti G, Vantaggi B (2009) T-conditional possibilities: coherence and inference. Fuzzy Sets Syst 160(3):306–324
Cox RT (1946) Probability, frequency, and reasonable expectation. Am J Phys 14:1–13
de Cooman G (1997) Possibility theory. Part I: measure- and integral-theoretic ground- work; Part II: conditional possibility; Part III: possibilistic independence. Int J General Syst 25:291–371
de Cooman G, Aeyels D (1999) Supremum preserving upper probabilities. Inf Sci 118(1–4):173–212
De Finetti B (1936) La logique de la probabilité. Congrès International de Philosophie Scientifique. Hermann et Cie, Paris, pp 1–9
De Finetti B (1974) Theory of probability. Wiley, New York
Denœux T (2014) Likelihood-based belief function: justification and some extensions to low-quality data. Int J Approx Reason 55(7):1535–1547
Destercke S, Dubois D, Chojnacki E (2008) Unifying practical uncertainty representations: I. Generalized p-boxes II. Clouds. Int J Approx Reason 49:649–663, 664–677
Domotor Z (1985) Probability kinematics - conditional and entropy principles. Synthese 63:74–115

Dubois D (1986) Belief structures, possibility theory and decomposable confidence measures on finite sets. Comput Artif Intell 5(5):403–416
Dubois D (2012) Reasoning about ignorance and contradiction: many-valued logics versus epistemic logic. Soft Comput 16(11):1817–1831
Dubois D, Esteva F, Godo L, Prade H (2007) Fuzzy-set based logics an history-oriented presentation of their main developments. In: Gabbay DM, Woods J (eds) Handbook of the history of logic, vol 8. Elsevier, pp 325–449
Dubois D, Fariñas del Cerro L, Herzig A, Prade H (1999) A roadmap of qualitative independence. In: Dubois D, Prade H, Klement E (eds) Fuzzy sets, logics and reasoning about knowledge, applied logic series, vol 15. Kluwer Academic Publication, Dordrecht, pp 325–350
Dubois D, Fodor J, Prade H (2010) Conditional measures: an alternative to Cox functional equation. In: Cintula P, Klement EP, Stout LN (eds) Proceedings of 31st linz seminar on fuzzy set theory, Linz, Austria, pp 43–46
Dubois D, Foulloy L, Mauris G, Prade H (2004) Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliab Comput 10:273–297
Dubois D, Godo L, de Mántaras RL, Prade H (1993) Qualitative reasoning with imprecise probabilities. J Intell Inf Syst 2(4):319–363
Dubois D, Grabisch M, de Mouzon O, Prade H (2001) A possibilistic framework for single-fault causal diagnosis under uncertainty. Int J General Syst 30(2):167–192
Dubois D, Hajek P, Prade H (2000a) Knowledge-driven versus data-driven logics. J Log Lang Inf 9:65–89
Dubois D, Lang J, Prade H (1994) Possibilistic logic. In: Gabbay D, Hogger C, Robinson J, Nute D (eds) Handbook of logic in artificial intelligence and logic programming, vol 3. Oxford University, Oxford, pp 439–513
Dubois D, Hüllermeier E (2007) Comparing probability measures using possibility theory: a notion of relative peakedness. Int J Approx Reason 45:364–385
Dubois D, Moral S, Prade H (1997) A semantics for possibility theory based on likelihoods. J Math Anal Appl 205:359–380
Dubois D, Pap E, Prade H (2000b) Hybrid probabilistic-possibilistic mixtures and utility functions. In: De Baets B, Perny P, Fodor J (eds) Preferences and decisions under incomplete knowledge, Physica-Verlag, pp 51–73
Dubois D, Prade H (1982) A class of fuzzy measures based on triangular norms. a general framework for the combination of uncertain information. Int J General Syst 8(1):43–61
Dubois D, Prade H (1988) Possibility theory: an approach to computerized processing of uncertainty. Plenum, USA
Dubois D, Prade H (1989) Handling uncertainty in expert systems: pitfalls, difficulties, remedies. In: Hollnagel E (ed) The Reliability of Expert Systems. Ellis Horwood, Chichester, U.K., pp 64–118
Dubois D, Prade H (1991) Epistemic entrenchment and possibilistic logic. Artif Intell 50:223–239
Dubois D, Prade H (1992) Putting rough sets and fuzzy sets together. In: Slowinski R (ed) Intelligent decision support - handbook of applications and advances of the rough sets theory. Kluwer Academic Publication, Dordrecht, pp 203–232
Dubois D, Prade H (1994) Conditional objects as nonmonotonic consequence relationships. IEEE Trans Syst Man Cybern 24(12):1724–1740
Dubois D, Prade H (1996) What are fuzzy rules and how to use them. Fuzzy Sets Syst 84:169–185
Dubois D, Prade H (1998) Possibility theory: Qualitative and quantitative aspects. In: Gabbay DM, Smets P (eds) Quantified representation of uncertainty and imprecision, handbook of defeasible reasoning and uncertainty management systems, vol 1. Kluwer Academic Publication, Dordrecht, pp 169–226
Dubois D, Prade H (2001) Possibility theory, probability theory and multiple-valued logics : a clarification. Ann Math Artif Intell 32:35–66
Dubois D, Prade H (2004) Possibilistic logic: a retrospective and prospective view. Fuzzy Sets Syst 144:3–23

Dubois D, Prade H (2009) Formal representations of uncertainty. In: Bouyssou D, Dubois D, Pirlot M, Prade H (eds) Decision-making process - concepts and methods. Wiley, New York, pp 85–156
Dubois D, Prade H (2012) Possibility theory and formal concept analysis: characterizing independent sub-contexts. Fuzzy Sets Syst 196:4–16
Dubois D, Prade H (2014) Possibilistic logic - An overview. In: Siekmann JH (ed) Computational logic, handbook of the history of logic, vol. 9. Elsevier, pp 283–342
Dubois D, Prade H (2016) Qualitative and semi-quantitative modeling of uncertain knowledge - A discussion. In: Brewka G, Thimm M, Beierle C (eds) Computational models of rationality. College Publications, pp 280–296
Dubois D, Prade H (eds.) (2008) Bipolar Representations of Information and Preference. Part 1A & Part 1B: Cognition and Decision; Part 2: Reasoning and Learning. Special issue, Int. J. of Intelligent Systems, 23(8,9,10), Wiley
Dubois D, Prade H, Rico A (2015a) The cube of opposition: a structure underlying many knowledge representation formalisms. In: Yang Q, Wooldridge M (eds) Proceedings of 24th international joint conference on artificial intelligence, IJCAI'15, AAAI, pp 2933–2939
Dubois D, Prade H, Rico A (2015b) Representing qualitative capacities as families of possibility measures. Int J Approx Reason 58:3–24
Dubois D, Prade H, Rico A (2017a) Graded cubes of opposition and possibility theory with fuzzy events. Int J Approx Reason 84:168–185
Dubois D, Prade H, Rico A (2017b) Organizing families of aggregation operators into a cube of opposition. In: Kacprzyk J, Filev D, Beliakov G (eds) Granular soft and fuzzy approaches for intelligent systems, Springer, Berlin, pp 27–45
Dubois D, Prade H, Schockaert S (2012) Stable models in generalized possibilistic logic. In: Brewka G, Eiter T, McIlraith SA (eds) Proceedings of 13th international conference principles of knowledge representation and reasoning(KR'12), Morgan Kaufmann, pp 519–529
Dubois D, Prade H, Schockaert S (2017c) Generalized possibilistic logic: foundations and applications to qualitative reasoning about uncertainty. Artif Intell 252:139–174
Dupin de Saint-Cyr F, Prade H (2008) Handling uncertainty and defeasibility in a possibilistic logic setting. Int J Approx Reason 49:67–82
Eichhorn C, Kern-Isberner G (2015) Using inductive reasoning for completing OCF-networks. J Appl Log 13(4):605–627
Ferré S, Ridoux O (2004) Introduction to logical information systems. Inf Process Manag 40(3):383–419
Fine T (1983) Theories of probability. Academic Press, New York
Gaifman H, Snir M (1982) Probabilities over rich languages, testing and randomness. J Symb Log 47(3):495–548
Ganter B, Kuznetsov SO (2001) Pattern structures and their projections. In: Delugach HS, Stumme G (eds) Proceedings of 9th International conference on conceptual structures (ICCS'01), vol 2120. Springer, LNCS, pp 129–142
Ganter B, Wille R (1999) Formal concept analysis. mathematical foundations. Springer, Berlin
Gärdenfors P (1988) Knowledge in flux, 2nd edn. Modeling the dynamics of epistemic states. MIT, College Publications, 2008
Ginsberg ML (1990) Bilattices and modal operators. J Log Comput 1:1–41
Goldszmidt M, Pearl J (1991) System Z^+: a formalism for reasoning with variable-strength defaults. In: Proceedings of 9th national conference on artificial intelligence (AAAI'91), vol 1, pp 339–404
Greco S, Inuiguchi M, Slowinski R (2006) Fuzzy rough sets and multiple-premise gradual decision rules. Int J Approx Reason 41(2):179–211
Grzymala-Busse JW (1988) Knowledge acquisition under uncertainty - a rough set approach. J Intell Robot Syst 1(1):3–16
Guigues JL, Duquenne V (1986) Familles minimales d'implications informatives résultant d'un tableau de données binaires. Mathématiques et Sciences Humaines 95:5–18
Hájek P (1998) The metamathematics of fuzzy logics. Kluwer Academic, Dordrecht
Halpern JY (1990) An analysis of first-order logics of probability. Artif Intell 46:311–350

Halpern JY (1999a) A counterexample to theorems of Cox and Fine. J Artif Intell Res (JAIR) 10:67–85
Halpern JY (1999b) Technical addendum, Cox's theorem revisited. J Artif Intell Res (JAIR) 11:429–435
Halpern JY (2001) Plausibility measures: A general approach for representing uncertainty. In: Nebel B (ed) Proceedings of 17th international joint conference on artificial intelligence (IJCAI'01), Morgan Kaufmann, pp 1474–1483
Halpern JY (2003) Reasoning about uncertainty. MIT, Cambridge
Halpern JY, Pucella R (2002) A logic for reasoning about upper probabilities. J Artif Intell Res (JAIR) 17:57–81
Halpern JY, Pucella R (2006) A logic for reasoning about evidence. J Artif Intell Res (JAIR) 26:1–34
Hempel CG (1945) Studies in the logic of confirmation, I and II. Mind, pp 1–26, 97–121
Higashi M, Klir GJ (1982) Measures of uncertainty and information based on possibility distributions. Int J General Syst 8:43–58
Hong T, Tseng L, Wang S (2002) Learning rules from incomplete training examples by rough sets. Expert Syst Appl 22(4):285–293
Horvitz E, Heckerman D, Langlotz C (1986) A framework for comparing alternative formalisms for plausible reasoning. In: Kehler T (ed) Proceedings of 5th national conference on artificial intelligence. vol 1. Morgan Kaufmann, pp 210–214
Jaeger M (2001) Automatic derivation of probabilistic inference rules. Int J Approx Reason 28:1–22
Jaeger M (2006) Probabilistic role models and the guarded fragment. Int J Uncertain, Fuzziness Knowl-Based Syst 14(1):43–60
Jaynes ET (1979) Where do we stand on maximum entropy. In: Tribus M, Levine I (eds) The maximum entropy formalism, MIT, pp 15–118
Jaynes ET (2003) Probability theory: the logic of science. Cambridge University, Cambridge, p 1996 preprint version, 1996
Kalman RE (1960) A new approach to linear filtering and prediction problems. J Basic Eng Trans ASME, Ser D 82:35–45
Kern-Isberner G, Eichhorn C (2014) Structural inference from conditional knowledge bases. Studia Logica 102(4):751–769
Kleene SC (1952) Introduction to metamathematics. North Holland
Klement E, Mesiar R, Pap E (2000) Triangular norms. Kluwer Academic Publication, Boston
Kraus S, Lehmann D, Magidor M (1990) Nonmonotonic reasoning, preferential models and cumulative logics. Artif Intell 44:167–207
Kyburg HE Jr, Teng CM (2012) The logic of risky knowledge, reprised. Int J Approx Reason 53:274–285
Lehmann DJ, Magidor M (1992) What does a conditional knowledge base entail? Artif Intell 55:1–60
Lindley DV (1982) Scoring rules and the inevitability of probability. Int Stat Rev 50:1–26
Liu W (2001) Propositional, probabilistic and evidential reasoning: integrating numerical and symbolic approaches. Physica Verlag, Springer
Lucas P, van der Gaag L (1991) Principles of expert systems. Addison-Wesley
Marchioni E, Godo L (2004) A logic for reasoning about coherent conditional probability: a modal fuzzy logic approach. In: Alferes JJ, Leite JA (eds) Proceedings of 9th European conference on logics in artificial intelligence (JELIA'04), vol 3229. Springer, LNCS, pp 213–225
Martin T (2006) Logique du probable de Jacques Bernoulli à J.-H. Lambert. Journ@l Electronique d'Histoire des Probabilités et de la Statistique 2(1b). http://www.jehps.net/Novembre2006/Martin3.pdf
Milch B, Russell SJ (2007) First-order probabilistic languages: into the unknown. In: Muggleton S, Otero RP, Tamaddoni-Nezhad A (eds) Revised selected papers of the 16th international conference on inductive logic programming (ILP'06), vol 4455. Springer, LNCS, pp 10–24

Mongin P (1994) Some connections between epistemic logic and the theory of nonadditive probability. In: Humphreys P (ed) Patrick Suppes: scientific philosopher. vol 1: probability and probabilistic causality, vol 234. Springer, Synthese Library, pp 135–171
von Neumann J, Morgenstern O (1944) Theory games and economic behavior. Princeton University, Princeton
Nilsson NJ (1993) Probabilistic logic revisited. Artif Intell 59:39–42
Paris J (1994) The uncertain reasoner' companion. Cambridge University, Cambridge
Parsons S (2001) Qualitative approaches for reasoning under uncertainty. MIT Press Cambrige, Mass
Parsons T (2008) The traditional square of opposition. In: Zalta EN (ed) The stanford encyclopedia of philosophy. Stanford University
Pasquier N, Bastide Y, Taouil R, Lakhal L (1999) Efficient mining of association rules using closed itemset lattices. Inf Syst 24:25–46
Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publication, Dordrecht
Pawlak Z, Skowron A (2007) 1. Rudiments of rough sets. 2. Rough sets: Some extensions. 3. Rough sets and Boolean reasoning. Inf Sci 177(1):3–73
Pearce D (2006) Equilibrium logic. Ann Math Artif Intell 47:3–41
Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publication
Pearl J (1990) System Z: a natural ordering of defaults with tractable applications for default reasoning. In: Proceedings of theoretical aspects of reasoning about knowledge TARK90, pp 121–135
Pearl J (2000) Causality. Models, Reasoning and Inference. Cambridge University, Cambridge
Pfeifer N, Sanfilippo G (2017) Probabilistic squares and hexagons of opposition under coherence. Int J Approx Reason 88:282–294
Reichenbach H (1952) The syllogism revised. Philos Sci 19(1):1–16
Renooij S, van der Gaag L (1999) Enhancing QPNs for trade-off resolution. In: Laskey KB, Prade H (eds) Proceedings of 15th conference on uncertainty in artificial intelligence (UAI '99), Morgan Kaufmann, pp 559–566
Renooij S, van der Gaag LC (2008) Enhanced qualitative probabilistic networks for resolving trade-offs. Artif Intell 172:1470–1494
Ruspini EH (1970) Numerical methods for fuzzy clustering. Inf Sci 2:319–350
Schweizer B, Sklar A (1963) Associative functions and abstract semi-groups. Publ Math Debrecen 10:69–180
Shackle GLS (1961) Decision, order and time in human affairs, 2nd edn. Cambridge University, UK
Shafer G (1976) A mathematical theory of evidence. Princeton University, Press
Shafer G (1978) Non-additive probabilities in the work of Bernoulli and Lambert. Arch Hist Exact Sci 19(4):309–370
Shortliffe EH (1976) Computer-based medical consultations MYCIN. Elsevier
Smets P (1982) Possibilistic inference from statistical data. In: Proceedings of 2nd World Conference on mathematics at the service of man, Las Palmas, pp 611–613
Snow P (1999) Diverse confidence levels in a probabilistic semantics for conditional logics. Artif Intell 113:269–279
Spohn W (1988) Ordinal conditional functions: a dynamic theory of epistemic states. In: Harper WL, Skyrms B (eds) Causation in decision, belief change, and statistics, vol 2. Kluwer, pp 105–134
Spohn W (1990) A general, nonprobabilistic theory of inductive reasoning. In: Shachter RD, Levitt TS, Kanal LN, Lemmer JF (eds) Uncertainty in Artificial Intelligence 4. North Holland, Amsterdam, pp 149–158
Spohn W (2012) The laws of belief: ranking theory and its philosophical applications. Oxford University, Oxford

Sugeno M (1977) Fuzzy measures and fuzzy integrals - A survey. In: Gupta MM, Saridis GN, Gaines BR (eds) Fuzzy Automata and Decision Processes. North Holland, Amsterdam, pp 89–102
Touazi F, Cayrol C, Dubois D (2015) Possibilistic reasoning with partially ordered beliefs. J Appl Log (4, Part 3) 13:770–798
Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
Zadeh LA (1968) Probability measures of fuzzy events. J Math Anal Appl 23(2):421–427
Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci 8:199–249
Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28

Representations of Uncertainty in AI: Beyond Probability and Possibility

Thierry Denœux, Didier Dubois and Henri Prade

Abstract This chapter completes the survey of the existing frameworks for representing uncertain and incomplete information, started in the previous chapter of this volume. The theory of belief functions and the theory of imprecise probabilities are presented. The latter setting is mathematically more general than the former, and both include probability theory and quantitative possibility theory as particular cases. Their respective knowledge representation capabilities are highlighted.

1 Introduction

Usually items of information are neither precise nor always coherent with one another. This chapter presents two uncertainty theories that generalize probability theory while being capable of handing incomplete information in an explicit way, by including possibility theory as a special case. There are two ways of building such a generalized framework.

The first idea is to introduce probability theory on top of the basic set-valued representation of incomplete information. Dempster imagined a set equipped with a probability distribution and a one-to-many mapping from this set to a space of interest. Such probabilities can be subjective or frequentist. Upper and lower probabilities are then obtained on the second space. Dempster considered this set-up as an extension of the fiducial paradigm for statistical inference, while Shafer interpreted these upper and lower probabilities as plausibility and belief functions without reference to an underlying probability space with a one-to-many mapping. The approach

T. Denœux (✉)
Université de Technologie de Compiègne,
CNRS, Heudiasyc (UMR 7253), Compiègne, France
e-mail: tdenoeux@utc.fr

D. Dubois · H. Prade
IRIT, CNRS and Université Paul Sabatier, Toulouse, France
e-mail: dubois@irit.fr

H. Prade
e-mail: prade@irit.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_4

so-obtained was called theory of evidence by Shafer. It is tailored for the representation and merging of unreliable pieces of evidence. In contrast, upper and lower probabilities in Dempster set-up may also model ill-known probabilities due to incomplete observations of random variables.

The second idea is to work with (convex) sets of probabilities, either because the statistical model is ill-known, or because the usual protocol for generating subjective probabilities is altered, admitting that buying and selling prices of lotteries pertaining to risky events may differ. The latter is the basis of Walley theory of lower previsions and imprecise probabilities. It turns out that the framework of Walley is mathematically more general than the theory of Dempster-Shafer. This chapter provides an account of these generalizations of Bayesian probability theory.

2 Theory of Belief Functions

The belief function model (Shafer 1976, 1990; Yager and Liu 2008) adds probabilities on top of the set-based approach to imprecision. It replaces a representation of the form $v \in A$, where A is a set of possible values of v, by a discrete probability distribution over possible statements of the form $v \in A$ (assuming the universe, called *frame of discernment* by Shafer, S is finite). We denote by m such a probability distribution on the power set 2^S of S (the set of all subsets of S). As m is a probability distribution, the condition $\sum_{A \subseteq S} m(A) = 1$ is verified. Function m is called a *mass function*, and $m(A)$ is called the *belief mass* assigned to subset A. Any subset A of S such that $m(A) > 0$ is called a *focal set* of m. We denote by $\mathscr{F}$ the family of focal sets. In general, we do not assign any positive mass to the empty set, i.e., we assume that $m(\emptyset) = 0$; mass function m is then said to be *normalized*. However, the Transferable Belief Model (TBM) (Smets and Kennes 1994) relaxes this constraint: the mass $m(\emptyset)$ then represents the degree of internal contradiction of the mass function.

In this hybrid representation of uncertainty, it is important to understand the meaning of the mass function. In particular, the belief mass $m(A)$ should not be confused with a probability of occurrence of A. According to Shafer (1976), $m(A)$ is "the measure of the belief committed exactly to A". More precisely, we can say that $m(A)$ is the probability that the agent *only knows* that $v \in A$. There is thus an implicit epistemic modality in $m(A)$, which is absent from $P(A)$. This is the reason why function m may be non-monotonic with respect to inclusion: we may have $m(A) > m(B) > 0$ when $A \subset B$, if the agent is sure enough that what is known is of the form $v \in A$. In particular, $m(S)$ is the probability that the agent does not know anything. The *vacuous* mass function $m^?$ defined by $m^?(S) = 1$ thus represents total ignorance. This epistemic interpretation of mass functions is in line with Shafer (1981)'s *random code* metaphor outlined in the next section.

2.1 Random Code Semantics

A mass function can be interpreted by considering that the information provided by a source (a piece of evidence) can be assimilated to a coded message whose meaning is random (Shafer 1981). More precisely, assume that the source sends an encrypted message using a code chosen at random from a set $C = \{c_1, \ldots, c_n\}$ with probabilities $p_1, \ldots, p_n$. We know the set of codes as well as the chances of each code to be selected. If we decode the message using code c_i, we get a decoded message of the form $v \in \Gamma(c_i) = A_i$, where Γ is a multivalued mapping from C to 2^S. The probability that the meaning of the original message is $v \in A$ is thus

$$m(A) = \sum_{\{1 \leq i \leq n : A_i = A\}} p_i. \tag{1}$$

In particular, the probability that the message is empty, i.e., that it contains no information about v, is $m(S)$. The triple (C, P, Γ), where P is a probability measure on C, defines a random set (Nguyen 2006). The formal equivalence between random sets and belief functions has been proved for the first time by Nguyen (1978). However, in random set theory, sets A with $m(A) > 0$ do not necessarily represent states of knowledge. They can be objects taking the form of sets (Couso et al. 2014), contrary to the case of evidence theory illustrated in the following example.

> **Example**: Consider a watch that may be out of order with some known probability ε. The set C describes the set of states of the watch, $C = \{\mathsf{working}, \mathsf{broken}\}$. Assume that the watch shows time h. In that case, the multivalued mapping Γ is $\Gamma(\mathsf{working}) = \{h\}$ (if the watch is working, it shows the right time), and $\Gamma(\mathsf{broken}) = S$ (if it is out of order, we do not know what time it is). The mass function induced by S is thus $m(\{h\}) = 1 - \varepsilon$ and $m(S) = \varepsilon$.

The mass function obtained in the previous example is said to be *simple* because the belief mass is shared between a single subset A of S, and S itself. Such a mass function arises when a non-reliable source states that $v \in A$, and the agent believes that the source is irrelevant with probability ε. This probability is committed to S whereas $m(A) = 1 - \varepsilon$.

This way of generating a mass function from a multivalued mapping was first proposed by Dempster (1967) in the context of statistical inference. Shafer (1976) renamed the upper and lower probabilities of Dempster *plausibility and belief functions*, respectively. To quote Shafer (2016b)'s recent intellectual autobiography:

> My thought was to surrender the word *probability* to the objective concept and to build a new subjective theory using mainly the word *belief*.

A mass function m models a state of knowledge, whereas the underlying triple (C, P, Γ) represents a body of evidence with uncertain meaning. Among theories of uncertainty, the theory of belief functions has the particularity of putting emphasis on the evidence that generates a state of knowledge, as shown by the title of Shafer (1976)'s seminal book: *A Mathematical Theory of Evidence*.

2.2 Basic Set Functions

A mass function m induces two set functions: a belief function Bel (for "*belief*") and a plausibility function Pl, defined, respectively, by

$$Bel(A) = \sum_{E \subseteq A, E \neq \emptyset} m(E); \quad Pl(A) = \sum_{E \cap A \neq \emptyset} m(E). \tag{2}$$

Observe that $\forall A$, $Bel(A) \leq Pl(A)$. When $m(\emptyset) = 0$, it is clear that $Bel(S) = Pl(S) = 1$, $Pl(\emptyset) = Bel(\emptyset) = 0$, and $Bel(A) = 1 - Pl(\overline{A})$. Consequently, these two functions are dual, as are necessity and possibility functions. The degree of belief $Bel(A)$ can be interpreted as the probability of provability of A from the available knowledge represented by m. In the language of modal logic, we should write $Bel(A) = P(\Box A)$, where $\Box$ represents the modality of provability (Pearl 1990). In the same way, $Pl(A)$ can be seen as the probability of logical consistency of A with m.

Belief functions Bel are *completely monotone*, i.e., for any $k \geq 2$ and any family $(A_1, \ldots, A_k)$ of subsets of S, the following inequality holds,

$$Bel\left(\bigcup_{i=1,\ldots,k} A_i\right) \geq \sum_{i=1}^{k} (-1)^{i+1} \sum_{I:|I|=i} Bel\left(\bigcap_{j \in I} A_j\right). \tag{3}$$

For Shafer (2016b), these inequalities play for belief functions the same role as Kolmogorov axioms for probability theory. Plausibility functions verify a similar property (they are *completely alternating*), changing the direction of the inequality and switching the $\cap$ and $\cup$ operations.

A *commonality function*

$$Q(A) = \sum_{E \supseteq A} m(E) \tag{4}$$

was also introduced by Shafer (1976), essentially for computational reasons. It later appeared that the commonality function is an extension of the guaranteed possibility function in possibility theory (Dubois et al. 2001) (see the previous chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" in this volume).

Conversely, knowing function Bel, we can uniquely recover function m by the Möbius transform as:

$$m(E) = \sum_{A \subseteq E} (-1)^{|E \setminus A|} Bel(A).$$

Similar identities make it possible to recover m from Pl or Q. The fast Möbius transform (Kennes 1992) can perform these operations efficiently.

Belief functions are often defined on finite universes. Yet, thanks to the formal identity between belief functions and random sets, it is easy to define belief functions

on the real line (Dempster 1968; Strat 1984; Smets 2005; Denœux 2009), or even on more abstract topological spaces (Shafer 1973, 1979; Nguyen 1978, 2006). We can also extend belief and plausibility functions to fuzzy events (Smets 1981) by means of Choquet integrals:

$$Bel(F) = \sum_{E \subseteq S} m(E) \cdot \min_{s \in E} F(s) \tag{5}$$

and

$$Pl(F) = \sum_{E \subseteq S} m(E) \cdot \max_{s \in E} F(s), \tag{6}$$

for the finite case. It is also possible to "fuzzify" the theory of belief functions by allowing either the focal sets to be fuzzy sets (Zadeh 1979; Yen 1990), or the belief masses to be intervals or fuzzy numbers (Denœux 1999, 2000a).

Two Special Cases

Two remarkable special kinds of belief functions are worth noticing:

1. Probability functions are obtained by assuming the focal sets to be singletons. It is clear that, if $m(A) > 0$ implies $\exists s \in S, A = \{s\}$, then $Bel(A) = Pl(A) = P(A)$ is the probability function such that $P(\{s\}) = m(\{s\}), \forall s \in S$. Conversely, Bel is a probability function if and only of $Bel(A) = Pl(A), \forall A \subseteq S$.
2. Plausibility functions are possibility measures (or, dually, belief functions are necessity measures) if and only of the focal sets are nested, i.e., if $\forall E \neq F \in \mathscr{F}, E \subset F$ or $F \subset E$. In that case, $Pl(A \cup B) = \max(Pl(A), Pl(B))$ and $Bel(A \cap B) = \min(Bel(A), Bel(B))$. For instance, a simple mass function, as in the above watch example, yields possibility and necessity measures.

We can associate to m the mapping $\varphi_m : S \to [0, 1]$ called *contour function* of m defined by $\varphi_m(s) = \mathrm{Pl}(\{s\})$, i.e.,

$$\forall s \in S, \quad \varphi_m(s) = \sum_{s \in E} m(E). \tag{7}$$

It is easy to see that function φ_m is normalized in the sense of possibility theory ($\varphi_m(s) = 1$ for some state $s \in S$) whenever the focal sets have a nonempty intersection (which is the case if they are nested). Recovering the mass function m from φ_m is only possible when the focal sets are either nested or disjoint. In particular, if Bel is a probability measure, φ_m coincides with m and is a probability distribution. Now assume that the focal sets are nested and form an increasing sequence $E_1 \subset E_2 \subset, \ldots, \subset E_n$, where $E_i = \{s_1, \ldots, s_i\}$; then φ_m is indeed a possibility distribution π, and (7) reduces to $\pi(s_i) = \sum_{j=i}^{n} m(E_j)$. The possibility measure Π and the necessity measure N defined from π coincide, respectively, with the plausibility and belief functions induced by m. The mass function can be recomputed from π as follows (with the notation $\pi(s_{n+1}) = 0$) (Dubois and Prade 1982):

$$m_\pi(E_i) = \pi(s_i) - \pi(s_{i-1}), \quad i = 1, \ldots, n. \tag{8}$$

2.3 Combination Rules

The combination of information or evidence from different sources plays a fundamental role in the theory of belief functions. The basic tool is *Dempster's rule of combination* (Dempster 1967; Shafer 1976), which makes it possible to combine independent pieces of information. This tool, as well as the precise definition of independence in this context can be introduced using the random code metaphor introduced in Sect. 2.1.

2.3.1 Dempster's Rule of Combination

Let m_1 and m_2 be two mass functions on S induced by random sets (C_1, P_1, Γ_1) and (C, P_2, Γ_2), where C_1 and C_2 are, as before, interpreted as sets of codes. Assume both codes are selected independently at random. For each pair $(c_1, c_2) \in C_1 \times C_2$, the probability that c_1 and c_2 are jointly selected is $P_1(\{c_1\})P_2(\{c_2\})$; we can then deduce that $v \in \Gamma_1(c_1) \cap \Gamma_2(c_2)$. If moreover we assume the two bodies of evidence pertain to the same message, we have to restrict to cases where $\Gamma_1(c_1) \cap \Gamma_2(c_2) \neq \emptyset$. Consequently, the joint probability distribution on $C_1 \times C_2$ should be conditioned on the set $\{(c_1, c_2) \in C_1 \times C_2 | \Gamma_1(c_1) \cap \Gamma_2(c_2) \neq \emptyset\}$. This line of reasoning leads to the following combination rule, called Dempster's rule:

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{9}$$

for any $A \subseteq S$, $A \neq \emptyset$ and $(m_1 \oplus m_2)(\emptyset) = 0$, where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{10}$$

is called the *degree of conflict* between m_1 and m_2. If $\kappa = 0$, the two bodies of evidence are said to be *non-conflicting*, i.e., each focal set of m_1 intersects all focal sets of m_2. If $\kappa = 1$, the two bodies of evidence are logically contradictory and, consequently, they cannot be combined. Mass function $m_1 \oplus m_2$ is called the *orthogonal sum* of m_1 and m_2. The unnormalized version of this rule, which corresponds to a random set intersection (Dubois and Prade 1986), was introduced by Smets (1990a). A general definition of Dempster's rule in infinite spaces was given by Shafer (1973, 2016a).

Dempster's rule is commutative, associative and it admits the vacuous mass function $m^?$ as neutral element. It can be easily computed using the commonality function (4). Denoting by Q_1, Q_2 and $Q_1 \oplus Q_2$ the commonality functions associated, respectively to m_1, m_2 and $m_1 \oplus m_2$, the following relation holds,

$$Q_1 \oplus Q_2 = \frac{1}{1-\kappa} Q_1 \cdot Q_2. \tag{11}$$

2.3.2 Dempster's Rule of Conditioning

Conditioning in evidence theory, referred to as *Dempster's rule of conditioning*, was proposed by Shafer (1976). It is a special case of Dempster's rule of combination (cf. Sect. 2.3.1), mass function m being combined with a logical mass function m_C such that $m_C(C) = 1$. The idea is to transfer all the mass from each focal set E to $E \cap C \neq \emptyset$, since m_C states that the truth lies in C, and to renormalize the obtained result. The new information C can then be viewed as a revision of the original belief function so as to ensure that $Pl(\overline{C}) = 0$: the situations in which C is false are now considered as impossible. Denoting by $Pl(A \parallel C)$ the revised plausibility, we have

$$Pl(A \parallel C) = \frac{Pl(A \cap C)}{Pl(C)}, \tag{12}$$

which clearly constitutes an extension of probabilistic conditioning. The conditional belief function is then obtained dually as $Bel(A \parallel C) = 1 - Pl(\overline{A} \parallel C)$. We can remark that, with this rule of conditioning, the size of focal sets decreases: consequently, information becomes more precise, and the intervals $[Bel(A), Pl(A)]$ become narrower (up to the normalization factor). Especially, when $Bel(C) = 0$ and $Pl(C) = 1$ (total ignorance about C), conditioning on C by Dempster's rule increases the precision of the resulting mass function. Indeed, Dempster's conditioning corresponds to a revision process leading to information enrichment. Revision is here viewed as the combination between a body of uncertain evidence and a sure piece of information.

2.3.3 Other Combination Rules

Dempster's rule tends to concentrate belief masses on smaller focal sets: it thus has a conjunctive behavior. We can define a *disjunctive* counterpart to Dempster's rule (Dubois and Prade 1986; Smets 1993) as follows,

$$\forall A \subseteq S, \quad (m_1 \cup m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C). \tag{13}$$

This combination rule assumes that at least one of the two information sources is reliable, contrary to Dempster's rule, which assumes that they both are reliable. The disjunctive rule is commutative, associative, and admits as neutral element the mass function m such $m(\emptyset) = 1$. It can be expressed from belief functions using product:

$$(Bel_1 \cup Bel_2) = Bel_1 \cdot Bel_2, \tag{14}$$

which can be compared to (11). Note that the weighted average of belief functions is still a belief function. It offers yet another alternative combination rule. The set of belief functions is thus closed under product and weighted average.

2.3.4 Approximations by Reducing the Number of Focal Sets

Both Dempster's rule (9) and its dual disjunctive rule (13) have the effect of increasing the number of focal sets. To avoid combinatorial explosion, a useful strategy is to approximate each mass function by a simpler one with fewer focal sets. Several methods with different degrees of complexity have been proposed for this purpose (Lowrance et al. 1986; Tessem 1993; Bauer 1997; Harmanec 1999; Denœux 2001). The simplest, yet quite effective approach, is the *Summarization* algorithm (Lowrance et al. 1986), which works as follows. Let $F_1, \ldots, F_n$ be the focal sets of m ranked by decreasing mass, i.e., $m(F_1) \geq m(F_2) \geq \cdots \geq m(F_n)$. If n exceeds some the maximum allowed number k of focal sets, then the $n-k$ focal sets F_i, $i = k+1, \ldots, n$ with the smallest masses are replaced by their union, and m is approximated by the mass function m' defined as

$$m'(F_i) = m(F_i), \quad i = 1, \ldots, k, \tag{15a}$$

$$m'\left(\bigcup_{i=k+1}^{n} F_i\right) = \sum_{i=k+1}^{n} m(F_i). \tag{15b}$$

A more sophisticated algorithm for grouping focal sets while minimizing information loss, based on the principle of hierarchical clustering, was proposed by Denœux (2001).

When Eq. (11) or (14) are used, the complexity depends no longer on the number of focal sets, but on the cardinality of the universe S. An efficient approximation algorithm based on the search for a coarsening (grouping of focal sets) minimizing information loss was proposed by Denœux and Ben Yaghlane (2002). Using a completely different approach, the combination of several belief functions can also be performed by Monte-Carlo simulation (see, e.g., Moral and Wilson 1994, 1996).

2.3.5 Conflict Management

The management of conflict between information sources in an important practical problem, which has drawn a lot of attention over the years (Lefèvre et al. 2002; Smets 2007; Martin et al. 2008; Destercke and Burger 2013). When a high conflict between pieces of information is detected, two strategies are possible: we can either revise the way information has been formalized, or we can use a *robust* combination rule yielding a consistent result in case of conflict.

An example of such rule is the (Dubois and Prade 1988) rule defined as follows:

$$(m_1 \circledast m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C) + \sum_{\{B \cap C = \emptyset, B \cup C = A\}} m_1(B) m_2(C), \qquad (16)$$

for any $A \subseteq \Omega$, $A \neq \emptyset$, and $(m_1 \circledast m_2)(\emptyset) = 0$. When the degree of conflict κ between m_1 and m_2 is zero, we get $m_1 \circledast m_2 = m_1 \oplus m_2$: in the absence of conflict, the Dubois–Prade rule is equivalent to Dempster's rule. In contrast, when the degree of conflict is equal to 1, we have $m_1 \circledast m_2 = m_1 \cup m_2$: in that case, the Dubois–Prade rule boils down to the disjunctive rule. In all other cases, the behavior of the $\circledast$ operator is intermediate between conjunctive and disjunctive modes: it is an *adaptive* combination rule. We can remark that this rule is commutative but it is not associative. However, an n-ary version can easily be defined, based on maximal consistent subsets of focal sets. More complex ways of distributing the conflict among focal sets have been proposed (see, e.g., Lefèvre et al. 2002; Martin et al. 2008). See also chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" in this volume, for more details on fusion operations.

2.3.6 Combination of Dependent Information

Dempster's rule (9) and its disjunctive counterpart (13) both make an independence assumption about the pieces of information to be combined. While it is often possible to break down a body of evidence into independent pieces (Shafer 2016c), this is not always the case, especially in sensor fusion applications. It is then useful to have a well-justified rule allowing us to combine non independent pieces of evidence.

Such a rule, called the *cautious rule*, was proposed by Denœux (2008). It is based on the weight function representation, which we will now introduce. A mass function m is said to be *separable* (Shafer 1976) if it is the orthogonal sum of simple mass functions (see Sect. 2.1). Denoting a simple mass function with focal sets A and S as $A^{w(A)}$, where $w(A)$ is the mass committed to S (so, $1 - w(A)$ is committed to A), a separable mass function can thus be written as

$$m = \bigoplus_{\emptyset \neq A \subset S} A^{w(A)}. \qquad (17)$$

Considering the negation $\overline{m}$ of a mass function m, defined by $\forall A, \overline{m} = m(\overline{A})$ (Dubois and Prade 1986), there is a De Morgan duality between the disjunctive rule (14) and the non-normalized variant of Dempster's rule (9) that has been exploited by Denœux (2008) to define a disjunctive decomposition of belief functions.

Given a separable mass function m with commonality function Q such that $m(S) > 0$, the weights $w(A)$ can be recovered from Q as

$$\ln w(A) = -\sum_{B \supseteq A} (-1)^{|B|-|A|} \ln Q(B), \quad \forall A \subset S,\ A \neq \emptyset. \qquad (18)$$

The mapping $w : 2^S \setminus \{\emptyset, S\} \rightarrow [0, 1]$ defined by (18) is called the *weight function* associated to m. When m is not separable but still verifies $m(S) > 0$ (it is then said to be *non dogmatic*), we can still define the weight function w from (18), but we can now have $w(A) > 1$ for some A (Smets 1995). Mass function m can then still computed from w using (17), where $A^{w(A)}$ with $w(A) > 1$ is no longer a proper mass function but a *generalized mass function* assigning "masses" $w(A) > 1$ to S and $1 - w(A) < 0$ to A.

Given two non dogmatic mass function m_1 and m_2 with weight functions w_1 and w_2, their orthogonal sum can be written as:

$$m_1 \oplus m_2 = \bigoplus_{\emptyset \neq A \subset \Omega} A^{w_1(A) w_2(A)},$$

i.e., the weight function of $m_1 \oplus m_2$ is the product of those of m_1 and m_2. In contrast, the cautious rule is defined as

$$m_1 \circledcirc m_2 = \bigoplus_{\emptyset \neq A \subset \Omega} A^{\min(w_1(A), w_2(A))}, \tag{19}$$

i.e., the weight function of $m_1 \circledcirc m_2$ is the minimum of those of m_1 and m_2. The cautious rule is commutative, associative *and idempotent*, which makes it suitable to combine dependent pieces of evidence. It can be justified by the Least Commitment Principle (see Sect. 2.4). A disjunctive counterpart of $\circledcirc$, called the bold disjunctive rule, can also be defined (Denœux 2008). With Dempster's rule and the disjunctive rule (13), the cautious and bold rules can be seen as particular elements of infinite families of rules based on triangular norms and on uninorms (Pichon and Denœux 2010). Other idempotent, but non associative rules have been defined and studied by Destercke and Dubois (2011) and Cattaneo (2011).

2.3.7 Taking into Account Metaknowledge About Sources

When combining information from several sources, it is often useful to take into account not only the information provided by the sources, but also metaknowledge about their properties (such as their reliability or truthfulness). The *discounting* operation (Shafer 1976; Smets 1993) makes it possible to account for the reliability of a source by transferring a fraction α of each mass $m(A)$ for $A \subset S$ to S. The discounted mass function, denoted by ${}^{\alpha}m$, is then given by

$${}^{\alpha}m = (1 - \alpha)\, m + \alpha\, m^{?},$$

where, as before $m^{?}$ denotes the vacuous mass function and α is called the *discount rate*. The *contextual discounting* operation, introduced by Mercier et al. (2008), generalizes discounting by allowing one to take into account the source's reliability in different contexts. Pichon et al. (2012) have proposed a very general mechanism for "correcting" and combining mass functions, taking into account the relevance

and truthfulness of information sources; they have shown that all connectives of Boolean logic can be interpreted in the light of these two properties. Other belief function correction mechanisms have been proposed by Mercier et al. (2012, 2016), and Pichon et al. (2016).

2.4 Imprecision, Specialization and Information Measures

Like any information items, it is interesting to compare belief functions according to their information content. This makes it possible, in particular, to apply the *maximum uncertainty* (Klir and Wierman 1999) or *least commitment* (Smets 1993) principle, which serves the same purpose as the maximum entropy principle in probability theory and the principle of minimal specificity in possibility theory. According to this principle, when several belief functions are compatible with a set of constraints, the least committed should be selected. In order to apply this principle, we need to define a partial order on the set of belief functions. For that purpose, we may either define a degree of imprecision or of uncertainty of a belief function, or we may adopt a more qualitative approach and directly define an informational ordering relation on the set of belief functions.

2.4.1 Quantitative Approach

As belief functions model both imprecise and uncertain information, we may be willing to measure imprecision and uncertainty separately. A natural measure of imprecision is the *expected cardinality* of the random set defined by the mass function,

$$\text{Imp}(m) = \sum_{E \subseteq S} m(E) \cdot \text{card}(E). \tag{20}$$

It is clear that $\text{Imp}(m^?) = \text{card}(S)$, where $m^?$ is the vacuous mass function, and $\text{Imp}(m) = 1$ when m is a probability mass function. It can be checked that $\text{Imp}(m) = \sum_{s \in S} Pl(\{s\})$. An alternative measure of imprecision is *nonspecificity* (Dubois and Prade 1985), defined for a normalized mass function m as

$$N(m) = \sum_{E \subseteq S} m(E) \log_2 \text{card}(E). \tag{21}$$

Nonspecificity was shown by Ramer (1987) to be the only measure of imprecision satisfying some rationality requirements.

The degree of uncertainty of a belief function can be measured by generalizing the well-known Shannon entropy of a probability measure P defined by

$$H(P) = -\sum_{i=1}^{\text{card}(S)} p_i \cdot \ln p_i. \tag{22}$$

Several extensions of $H(p)$ to belief functions have been proposed, of the form

$$D(m) = -\sum_{E \subseteq S} m(E) \cdot \ln g(E), \tag{23}$$

where g can be, e.g., Pl or Bel (Dubois and Prade 1987; Klir and Wierman 1999). For $g = Pl$, we get a measure of *dissonance* (or internal conflict), which is maximized by uniform probability measures, and reaches its minimum (zero) when the intersection of focal sets is non empty : $\bigcap\{E : m(E) > 0\} \neq \emptyset$. For $g = Bel$, we rather have a measure of *confusion*, which is minimal (zero) for logical mass functions verifying $m(E) = 1$ for some unique focal set E (imprecise but certain and clear information), but high for uniform mass functions over subsets of S with cardinality $\text{Card}(S)/2$ (Dubois and Ramer 1993). See also Ramer and Klir (1993), Klir and Wierman (1999).

Another approach, proposed by Smets (1983), is to define a measure I of information content that relies on the pivotal role of Dempster's rule in the theory of belief functions, namely, it is natural to impose an additivity property with respect to this rule, such as $I(m_1 \oplus m_2) = I(m_1) + I(m_2)$ for any two non-conflicting mass functions m_1 and m_2. As shown by Smets (1983), this requirement, together with a few additional natural conditions, lead to the following definition[1]:

$$I(m) = -\sum_{E \subseteq S} \ln Q(E). \tag{24}$$

Other quantitative criteria attempt to measure imprecision and uncertainty simultaneously. For instance, *aggregate uncertainty* (Maeda and Ichihashi 1993; Harmanec and Klir 1994) is defined as follows, for a normalized mass function m:

$$AU(m) = \max_{P \in \mathscr{P}(m)} H(P), \tag{25}$$

where H is the Shannon entropy, and $\mathscr{P}(m)$ is the set of probability measures on S compatible with m:

$$\mathscr{P}(m) = \{P, P(A) \leq Pl(A), \forall A \subseteq S\}. \tag{26}$$

It is clear that $AU(m)$ is maximal both for the vacuous mass function $m = m^?$ and for the uniform Bayesian mass function m such that $m(\{s\}) = 1/\text{card}(S)$ for all $s \in S$; these two mass functions correspond, respectively, to maximal imprecision and to maximal uncertainty. Aggregate uncertainty can be shown to meet a number of reasonable requirements (Klir and Wierman 1999). However, the debate on what

[1] Considering the disjunctive rule instead of the conjunctive rule would lead to replace Q by Bel in (24).

should be a "natural" measure of total uncertainty in the theory of belief functions is not settled: see, for instance, the recent proposal and discussion by Jiroušek and Shenoy (2016).

2.4.2 Comparative Approach

The second approach to comparing the informational contents of belief functions consists of directly defining a partial order on the set of belief functions. Given two normalized mass functions m_1 and m_2, m_1 is said to be *more precise* than m_2 (denoted by $m_1 \sqsubseteq_{Pl} m_2$) iff, for any subset A of S, the interval $[Bel_1(A), Pl_1(A)]$ is included in the interval $[Bel_2(A), Pl_2(A)]$. Because of the duality of Bel and Pl, this condition can be simplified to: $\forall A$, $Pl_1(A) \leq Pl_2(A)$. In terms of imprecise probabilities, the condition $m_1 \sqsubseteq_{Pl} m_2$ means that $\mathscr{P}(m_1)$ is a subset of $\mathscr{P}(m_2)$ (Dubois and Prade 1986; Yager 1986). Mass function m is thus maximally precise when it coincides with a single probability measure, and minimally precise if $m = m^?$. It is also clear that, if $m_1 \sqsubseteq_{Pl} m_2$, then $AU(m_1) \leq AU(m_2)$. Note that this approach is in agreement with the imprecise probability interpretation of belief functions explained in Sect. 3.

An alternative method for comparing the informativeness of belief functions consists in generalizing relative specificity, viewed as set inclusion, to random sets. A normalized mass function m_1 is a *specialization* of a normalized mass function m_2 (denoted by $m_1 \sqsubseteq_s m_2$) if and only of the following three conditions hold:

1. Any focal set of m_2 contains at least one focal set of m_1;
2. Any focal set of m_1 is included in at least one focal set of m_2;
3. There exists a stochastic matrix W whose element w_{ij} is the proportion of the mass $m_1(E_i)$ assigned to $F_j \supseteq E_i$ in order to reconstruct mass $m_2(F_j)$, i.e., $m_2(F_j) = \sum_i w_{ij} \cdot m_1(E_i)$.

This relation is stronger than the previous one: if m_1 is a specialization of m_2, then m_1 is also more precise than m_2, but the converse is not true in general, see Dubois and Prade (1986). It is also obvious that, if m_1 is specialization of m_2, then $\text{Imp}(m_1) \leq \text{Imp}(m_2)$.

As noted in Sect. 2.2, in the consonant case, m_π and π contain the same information, i.e., $Pl = \Pi$ and $Bel = N$. Accordingly, for possibility measures, the precision and specialization orderings both coincide with the specificity ordering for possibility distributions: m_{π_1} is a specialization of m_{π_2} iff $\Pi_1(A) \leq \Pi_2(A), \forall A \subseteq S$ iff $\pi_1(s) \leq \pi_2(s), \forall s \in S$ (Dubois and Prade 1986).

Other informational orderings have been proposed. For instance, m_1 is said to be more informative than m_2 according to commonalities (denoted by $m_1 \sqsubseteq_Q m_2$) iff $Q_1 \leq Q_2$ (Dubois and Prade 1986; Yager 1986). This property can be interpreted from Eq. (11): as numbers $Q_1(A)$ get closer to 1, the influence of m_1 when combined by Dempster's rule with another mass function m_2 becomes smaller, which means that m_1 becomes less informative. Relation $\sqsubseteq_Q$ is weaker than $\sqsubseteq_s$, but it is not comparable with $\sqsubseteq_{Pl}$. Obviously, $m_1 \sqsubseteq_Q m_2$ implies that $I(m_1) \geq I(m_2)$.

Yet another ordering relation was proposed by Denœux (2008), based on the weight function (18). Mass function m_1 is said to be more informative than m_2 according to the weights (denoted by $m_1 \sqsubseteq_w m_2$) iff $w_1 \leq w_2$. This means that m_1 is the orthogonal sum of m_2 and a separable mass function m that has no conflict with m_2: $m_1 = m_2 \oplus m$. The cautious rule (19) can be derived from the least commitment principle based on relation $\sqsubseteq_w$.

2.5 Criteria for Decision Under Uncertainty

Consider a set $\mathscr{A} = \{a_1, \ldots, a_r\}$ of acts, a set $S = \{s_1, \ldots, s_n\}$ of states of nature, and a payoff matrix U of size $r \times n$, whose element u_{ij} is the utility of choosing act a_i if state s_j occurs. Assuming the uncertainty about the state of nature to be modeled by a mass function m on S, which act should be chosen? To answer this question, the classical Maximum Expected Utility (MEU) principle (von Neumann and Morgenstern 1944) can be generalized in a number of ways in the belief function setting (see also chapter "Decision Under Uncertainty" in this volume).

2.5.1 Lower and Upper Expected Utilities

According to Dempster (1967) and Shafer (1981), the *lower* and *upper expected utilities* of act a_i are defined, respectively, as the following Choquet integrals (further studied in chapter "Decision Under Uncertainty" of this volume) similar to (5):

$$\underline{\mathrm{EU}}(a_i) = \sum_{E \subseteq S} m(E) \min_{s_j \in E} u_{ij} \tag{27a}$$

$$\overline{\mathrm{EU}}(a_i) = \sum_{E \subseteq S} m(E) \max_{s_j \in E} u_{ij}. \tag{27b}$$

The lower and upper expected utilities can be shown to be, respectively, the lower and upper bounds of the expected utility with respect to all probability measures P on S compatible with m (Shafer 1981). An optimistic decision-maker (DM) will typically maximize the upper expected utility, while a pessimistic DM will maximize the lower expected utility. These two decision rules can be generalized by considering a convex sum of the lower and upper expected utility (Jaffray 1989; Strat 1990), which generalizes Hurwicz criterion (see chapter "Decision Under Uncertainty" in this volume for a detailed discussion of its axiomatization due to Jaffray):

$$\mathrm{EU}_\alpha(a_i) = \sum_{E \subseteq S} m(E) \left(\alpha \min_{s_j \in E} u_{ij} + (1-\alpha) \max_{s_j \in E} u_{ij} \right) \tag{28a}$$

$$= \alpha \underline{\mathrm{EU}}(a_i) + (1-\alpha) \overline{\mathrm{EU}}(a_i), \tag{28b}$$

where α can be seen as a pessimism index. An even more general approach, proposed by Yager (1992), combines the utilities in each set $\{u_{ij} \mid s_j \in E\}$ by an Ordered Weighted Average (OWA) operator.

2.5.2 Pignistic Probability

Following a different line of reasoning and putting emphasis on the avoidance of Dutch books (i.e., sequences of bets ensuring a sure loss), Smets (1990b) advocated a two-level mental model: the *credal* level, where uncertainty is represented by a belief function, and the *pignistic* level, where belief functions are transformed to probabilities for decision-making. The *pignistic transformation* (Smets 1990b) consists in distributing each mass $m(E)$ equally to all elements of E, resulting in the probability distribution *betp* defined as

$$betp(s) = \sum_{E:s\in E} \frac{m(E)}{\text{card}(E)}. \tag{29}$$

This transformation had been earlier proposed by Dubois and Prade (1982) as a generalization of Laplace's principle of insufficient reason to belief functions. Smets (1990b) justified it axiomatically, by imposing a linearity property (the pignistic probability of a convex sum of belief functions should be the convex sum of the pignistic probabilities) and an anonymity property (the pignistic probability of an event E should not change after permuting the elements of E). In fact, the pignistic probability was already known in the theory of cooperative games since the 1950s as the *Shapley value*, and Smets' axioms are mathematically the same as those proposed by Shapley (1953), albeit in a different context. The pignistic probability is also the center of gravity of the convex set of probabilities that dominate the belief function.

We can also search for the least informative belief function, according to the commonality ordering $\sqsubseteq_Q$ defined in Sect. 2.4.2, corresponding to a given pignistic probability distribution. As shown by Dubois et al. (2008), it is unique and consonant; consequently, it induces a possibility distribution.

Having defined the pignistic distribution *betp*, we can evaluate each act a_i by its expected utility with respect to *betp*,

$$\text{EU}_{betp}(a_i) = \sum_{s_j\in S} betp(s_j)u_{ij} = \sum_{E\subseteq S} m(E)\left(\frac{1}{\text{card}(E)}\sum_{s_j\in E} u_{ij}\right), \tag{30}$$

which can be compared to (27) and (28a). The pignistic criterion is a special case of Yager's OWA criterion (Yager 1992), as the average is a particular OWA operator.

2.6 Applications to Statistical Learning and Data Analysis

In Artificial Intelligence, the theory of belief functions has been used, until the early 1990's, to model uncertainty in expert systems (Shafer 1987; Shenoy 1989). Since 1990, we have seen the development of another application area: statistical learning (see chapter "Designing Algorithms for Machine Learning and Data Mining" of Volume 2). The theory of belief functions has proved to be an efficient formalism for combining models, modeling uncertainty in the outputs of classifiers or clustering algorithms, and learning from uncertain data. In the following, we review some of the recent developments in this area.

2.6.1 Classifier Combination

A first way of applying the theory of belief functions to classification is to consider classifier outputs as items of evidence and to merge them using Dempster's rule, or any other combination rule (see Sect. 2.3). Given the flexibility of the belief function formalism, this approach can be applied to classifiers of various types, the outputs of with can be converted into belief functions.

For instance, Xu et al. (1992) proposed to use a confusion matrix to convert a classifier's decision into a mass function. They obtained good results for a handwriting recognition problem. A similar approach was used by Mercier et al. (2009) for postal address recognition. More recently, Bi et al. (2008) proposed to represent classifier scores as "triplet" mass functions with three focal sets. Bi (2012) studied the influence of classifier diversity and the combination rule on the accuracy of the ensemble. Quost et al. (2011) considered a parametrized family of combination rules, including Dempster's rule and the cautious rule (see Sect. 2.3.6), and proposed a method to find the best rule in this family.

From a different perspective, Quost et al. (2007) considered the combination of binary classifiers as a way to solve multi-class classification problems. For instance, in the "one-against-one" approach, binary classifiers are trained using data from only two classes; consequently, their outputs can be interpreted as conditional mass functions. The problem is then to construct an unconditional mass function on the whole set of classes, as consistent as possible with the conditional mass functions provided by the binary classifiers.

2.6.2 Evidential Classifiers

An *evidential classifier* is a classifier whose output is a mass function over a set of classes $\Omega = \{\omega_1, \ldots, \omega_c\}$. Two main approaches have been proposed for constructing such a classifier from training data.

The first approach, first introduced and justified axiomatically by Appriou (1991, 1998), is to construct a mass function m on Ω from the likelihoods $p(x|\omega_k)$, where

x is the feature vector. One of the two methods proposed by Appriou is identical to the solution resulting from the application the *Generalized Bayes Theorem* (GBT) introduced by Smets (1993). The mass function has the following expression:

$$m = \bigoplus_{k=1}^{c} \overline{\{\omega_k\}}^{\alpha_k p(x|\omega_k)}, \tag{31}$$

where the α_k's are coefficients ensuring that $\alpha_k p(x|\omega_k) \leq 1$ for $k = 1, \ldots, c$, and the notation A^w stands for the simple mass function μ such that $\mu(A) = 1 - w$ and $\mu(\Omega) = w$ (see Sect. 2.3.6). A major advantage of this method is that it can be used without prior class probabilities, or with only weak prior information encoded as a belief function. However, when prior probabilities are given, the GBT yields the same solution as the Bayesian approach. Appriou (1991) showed the robustness of this method, in particular when the test data distribution differs from the learning distribution due, e.g., to different data acquisition methods or to sensor malfunction.

Another approach, referred to as the *evidential k-nearest neighbor (NN) rule*, was introduced by Denœux (1995). It consists in considering each training instance (or only each of the k nearest instances in the training set) as a piece of evidence about the class of the new object to be classified. The different pieces of evidence are represented by mass functions and are combined using Dempster's rule. In the most general form of this method, we consider a training set

$$\mathscr{L} = \{(x^{(1)}, m^{(1)}), \ldots, (x^{(N)}, m^{(N)}),$$

where $x^{(i)}$ is the feature vector of instance i and $m^{(i)}$ is a mass function on Ω representing partial knowledge about the class of that example. In the fully supervised case, each mass function $m^{(i)}$ is certain, i.e., we have $m^{(i)}(\{\omega_j\}) = 1$ for some element ω_j of Ω. In the general case, we have a *partially supervised* learning problem. Partial knowledge about the class of training instances may be provided by an expert or derived from indirect observation. We also assume a distance or dissimilarity measure δ between feature vectors.

The mass function representing the evidence of the training example $e^{(i)} = (x^{(i)}, m^{(i)})$ is defined as

$$m(A \mid e^{(i)}) = \varphi\left(\delta(x, x^{(i)})\right) m^{(i)}(A), \quad \forall A \subset \Omega \tag{32a}$$

$$m(\Omega \mid e^{(i)}) = 1 - \sum_{A \subset \Omega} m(A|e^{(i)}), \tag{32b}$$

where φ is a decreasing function verifying $\varphi(0) \leq 1$ and $\lim_{d\to\infty} \varphi(d) = 0$. Mass function $m(\cdot|e^{(i)})$ is thus obtained by discounting $m^{(i)}$ (see Sect. 2.3.7), with a discount rate that gets closer to one when the dissimilarity between vectors x and $x^{(i)}$ goes to infinity. The condition $\lim_{d\to\infty} \varphi(d) = 0$ ensures that mass function $m(\cdot|e^{(i)})$ becomes vacuous when the dissimilarity between vectors x and $x^{(i)}$ goes to infinity.

Let us now consider a new object described by a known feature vector $\hat{x}$ and an unknown class label $y \in \Omega$. Having computed mass functions (32) for each of the K nearest neighbors of $\hat{x}$, the combined mass function on Ω is

$$m(\cdot \mid \mathscr{L}) = \bigoplus_{\{i \mid x_i \in \mathscr{N}_K(\hat{x})\}} m(\cdot \mid e^{(i)}), \tag{33}$$

where $\mathscr{N}_K(\hat{x})$ denotes the set of the K nearest neighbors of $\hat{x}$. The choice of a best class $\hat{y} \in \Omega$ can then be made using one of the decision rules described in Sect. 2.5 and in chapter "Decision Under Uncertainty" of this volume. Denœux (1997) describes several decision strategies with different reject options.

Zouhal and Denœux (1998) have proposed a method for choosing function φ within a parametric family by minimizing an error function. The *evidential neural network classifier* introduced by Denœux (2000b) is a variant of this method, in which the training set is summarized as a set of prototypes. Both the evidential k-NN rule and the evidential neural network classifier have been implemented in the R package `evclass` (Denœux 2017). Denœux and Zouhal (2001) have studied another variant of the evidential k-NN rule in which partial information about the class of training instances is given as possibility distributions. Petit-Renaud and Denœux (2004) have extended the approach to regression problems, where variable y is numerical. Recently, Lian et al. (2015) proposed a feature selection method based on the evidential k-NN rule, and Lian et al. (2016) described an algorithm for learning the distance function δ in (32).

The evidential k-NN rule has also been extended to multi-label classification problems, in which each object may belong simultaneously to several classes (Denœux et al. 2010). In this case, the universe is the power set 2^{Ω} of the set of classes. To prevent double exponential complexity in the manipulation of mass functions, belief functions can then be defined on a lattice of subsets of Ω (the intervals with respect to the ordering relation $\subseteq$). A general presentation of this approach (with applications not only to classification, but also to preference elicitation and to clustering) can be found in Denœux and Masson (2012). See also Grabisch (2009) for the general theory of belief functions on lattices.

The likelihood-based and distance-based evidential classification methods outlined above have been compared experimentally by Fabre et al. (2001), and theoretically by Denœux and Smets (2006), who showed that they can both be derived from the GBT.

2.6.3 Evidential Clustering

The theory of belief functions has also been applied to clustering, which consists in finding groups (or clusters) in data (see chapters "Designing Algorithms for Machine Learning and Data Mining" and "Constrained Clustering: Current and New Trends" of Volume 2). Here, belief functions can be used to quantify the uncertainty

about the group membership of each particular object. Given a set of n objects $\mathcal{O} = \{o_1, \ldots, o_n\}$ and a set of c clusters $\Omega = \{\omega_1, \ldots, \omega_c\}$, Denœux and Masson (2004) defined a *credal partition* as an n-tuple $M = (m_1, \ldots, m_n)$ of (not necessarily normalized) mass functions on Ω, where m_i quantifies the uncertainty about the cluster membership of object o_i. A credal partition boils down to a hard partition when all mass functions are precise (i.e., when they focus on only one singleton). Most other "soft" clustering notions such as fuzzy, possibility and rough clustering are also recovered as special cases (Denœux and Kanjanatarakul 2016). For instance, if all mass functions correspond to probability distributions (i.e., their focal sets are singletons), then we can identify each mass $m_i(\{\omega_k\})$ with the degree of membership u_{ik} of object o_i to cluster ω_k, and we have a fuzzy partition (Bezdek 1981). If each mass function m_i is categorical (i.e., it has only one focal set A_i), then we can define the *lower approximation* of cluster ω_k as the set of objects o_i that surely belong to ω_k, i.e., such that $A_i = \{\omega_k\}$, and the *upper approximation* of cluster ω_k as the set of objects o_i that may belong to ω_k, i.e., such that $\omega_k \in A_i$. We then have a *rough partition* as defined by Lingras and Peters (2012). A general credal partition can also easily be summarized into a hard partition or any type of soft partition. For instance, we obtain a fuzzy partition by replacing each mass m_i by its pignistic probability distribution (29), and we get a rough partition by selecting, for each mass function m_i, the focal set with the largest mass (Denœux and Kanjanatarakul 2016).

An *evidential clustering* algorithm is a procedure that constructs a credal partition from a dataset. Several such algorithms have been proposed over the years:

- The *EVCLUS* algorithm, introduced by Denœux and Masson (2004), applies ideas from multidimensional scaling to clustering: given a dissimilarity matrix, it finds a credal partition such that the degrees of conflict (10) between mass functions match the dissimilarities, dissimilar objects being represented by highly conflicting mass functions; this is achieved by iteratively minimizing a stress function. A variant of EVCLUS allowing one to use prior knowledge in the form of pairwise constraints was later introduced by Antoine et al. (2014), and several improvements to the original algorithm making it capable of handling large dissimilarity datasets have been reported by Denœux et al. (2016) and Li et al. (2018).
- The *Evidential c-means* (ECM) algorithm (Masson and Denœux 2008) is a c-means-like algorithm that minimizes a cost function by searching alternatively the space of prototypes and the space of credal partitions. Unlike the hard and fuzzy c-means algorithms, ECM associates a prototype not only to each cluster, but also to each nonempty set of clusters. The prototype associated to a set of clusters is defined as the barycenter of the prototypes of each single cluster in the set. The cost function to be minimized insures that objects close to a prototype have a high mass assigned to the corresponding set of clusters. A variant with adaptive metrics and binary constraints was introduced by Antoine et al. (2012), and a relational version for dissimilarity data (called RECM) has been proposed by Masson and Denœux (2009). A version of ECM taking into account spatial constraints and suitable for image segmentation was introduced by Lelandais et al. (2014).

- The Ek-NNclus algorithm (Denœux et al. 2015) is a decision-directed clustering procedure based on the evidential k-NN rule described in Sect. 2.6.2. Starting from an initial partition, the algorithm iteratively reassigns objects to clusters using the evidential k-NN rule, until a stable partition is obtained. After convergence, the cluster membership of each object is described by a mass function on Ω assigning a mass to each cluster and to the whole set of clusters. The mass assigned to the set of clusters can be used to identify outliers. The procedure can be seen as searching for the most plausible partition of the data.

All these algorithms have been implemented in the R package `evclust` (Denœux 2016).

3 Imprecise Probabilities

Imprecise probability theory (Walley 1991) relies on an approach opposite to the one of belief functions. Instead of randomizing the set-based approach to incomplete information, incompleteness is injected in probability theory. Under the frequentist view, epistemic uncertainty goes on top of a probabilistic model. Under the subjectivist view, the betting protocol is relaxed, by no longer enforcing the equality between buying and selling prices. In the area of economics, Gilboa and Schmeidler (1989) already showed that by suitably relaxing Savage axioms for decision under uncertainty, it is possible to formally justify the idea that an agent's epistemic state consists of a set of probability distributions on the set S of possible states of the world: in order to hedge against uncertainty, when evaluating a decision, the cautious agent picks the probability distribution that minimizes its expected utility.

3.1 Basic Definitions and Interpretations

An imprecise probability model comes down to specifying a family $\mathscr{P}$ of probability functions over S. However, there are several approaches to come up with this family according to the understanding of probability (frequentist or subjectivist), and to the available data in the specific application context.

3.1.1 Incomplete Information About Frequentist Probability

Under a frequentist view, $\mathscr{P}$ is an epistemic set reflecting incomplete information about an otherwise precise mathematical model of a random process: a probability distribution in $\mathscr{P}$ is the right one. The family $\mathscr{P}$ thus represents an imprecise probabilistic model. There are several situations that lead to such a model:

- The most common situation is when several probability measures are compatible with the available information, for instance in the case of scarce data. In the parametric case, the parameters of the model are ill-known, because the confidence intervals for these parameters are too wide. Bayesians then often assume a prior probability distribution on the parameter range or the set of possible probability functions. This is precisely what is not assumed in the imprecise probability setting. Some authors may still use the Bayesian paradigm, but assume imprecision about the prior probability (they are called robust Bayesians Huber 1981; Berger 1994), resulting in an imprecise posterior distribution.
- Imprecise information can be obtained by an expert or from empirical data about statistical parameters (like support, mean, mode, median, some fractiles) but the type of probabilistic model is otherwise ill-known (Baudrit and Dubois 2006) (e.g., you know the empirical mean and variance but you do not know if the process is Gaussian or not). It may be that the expert provides probability bounds on some events (intervals, quantiles, etc.). In the finite case, an expert may assign a probability interval to each outcome instead of a precise value (de Campos et al. 1994).
- A usual setting for getting upper and lower probabilities is the one of imprecise statistical information, that corresponds to Dempster (1967)'s setting for belief functions. The mass value of a focal set is the frequency of observing this incomplete information item. In that case, belief and plausibility functions are lower and upper probabilities, respectively, with a frequentist flavor. See the book by Couso et al. (2014) for a presentation of this approach to imprecise statistics.
- Some authors have even questioned the basic assumptions that frequencies converge toward limit probabilities. For instance it is only known that frequencies eventually remain inside an interval (Walley and Fine 1982).

Suppose one comes up to a probability family $\mathscr{P}$ via some of the above scenarii. Then one can assign to each event lower and upper bounds for the probability of this event (Smith 1961):

$$P_*(A) = \inf_{P \in \mathscr{P}} P(A); \quad P^*(A) = \sup_{P \in \mathscr{P}} P(A). \tag{34}$$

Functions P_* and P^* are monotonic with respect to inclusion and satisfy the duality property $P^*(A) = 1 - P_*(\overline{A})$. We call set functions P_* and P^* *lower and upper envelopes* respectively, after (Walley 1991). The additivity property of P enforces the following conditions for such envelopes (Good 1962): $\forall A, B \subseteq S$, such that $A \cap B = \emptyset$,

$$P_*(A) + P_*(B) \leq P_*(A \cup B) \leq P_*(A) + P^*(B) \leq P^*(A \cup B) \leq P^*(A) + P^*(B). \tag{35}$$

The width of the interval $[P_*(A), P^*(A)]$ represents the amount of ignorance of the agent as to the truth of proposition A. Total ignorance is when this interval is $[0, 1]$. When this interval reduces to a singleton, full probabilistic knowledge is obtained.

Probability envelopes are more general than belief and plausibility functions, hence more general than necessity and possibility measures (Walley 1996).

It is important to notice that in general, it is impossible to reconstruct the original set $\mathscr{P}$ from the knowledge of these intervals $[P_*(A),\ P^*(A)]$ for all events A. Indeed, these intervals correspond to particular projections of $\mathscr{P}$. Namely, letting $\mathscr{P}(P_*) = \{P : \forall A \subseteq S,\ P(A) \geq P_*(A)\}$, it is easy to see that $\mathscr{P}(P_*)$ is convex (if $P_1 \in \mathscr{P}(P_*)$ and $P_2 \in \mathscr{P}(P_*)$ then, $\forall \lambda \in [0, 1]$, $\lambda \cdot P_1 + (1 - \lambda) \cdot P_2 \in \mathscr{P}(P_*)$) and contains the convex hull of $\mathscr{P}$, even if $\mathscr{P}$ and $\mathscr{P}(P_*)$ have the same upper and lower envelopes.

A characteristic property of an upper envelope (induced by a non-empty set of probabilities) was found by Giles (1982). Viewing a set A as its $\{0, 1\}$-valued characteristic function ($A(s) = 1$ if $s \in A$ and 0 otherwise). A set-function g is an upper envelope if and only if for any tuple $A_0, A_1, \ldots, A_k$ of subsets of S, and any pair of positive integers (r, s) such that $\sum_{i=1}^{k} A_i(\cdot) \geq r + s \cdot A_0(\cdot)$, it holds that

$$\sum_{i=1}^{k} g(A_i) \geq r + s \cdot g(A_0). \tag{36}$$

3.1.2 The Subjectivist Point of View

The subjectivist approach to imprecise probability was fully developed by Walley (1991). It is powerful enough to encompass all convex sets of probabilities. In this approach the agent proposes buying prices for gambles. A gamble is a function f from S to the real line that expresses losses ($f(s) < 0$) or gains ($f(s) > 0$). The gamble associated to an event is its characteristic function. The agent is not committed to selling such gambles at the same prices as the ones he or she accepts to buy them.

Informally, the approach relies on so-called *desirable gambles* (Walley 1991) that the agent would agree to buy for a positive price. The set of desirable gambles contains at least all positive gambles. Moreover the sum of two desirable gambles is desirable, and a desirable gamble remains desirable when multiplied by a positive constant. The lower prevision $LP(f)$ of a gamble f is the maximal value α such that $f - \alpha$ is desirable. It can be shown that given a set of gambles $f_i \in \mathscr{G}$ and their lower previsions $LP(f_i)$, there is a convex set of probabilities $\mathscr{P}$, called *credal set*, such that $LP(f_i)$ is the lower expectation of f_i according to $\mathscr{P}$, for all $f_i \in \mathscr{G}$. One important point is that any convex set of probabilities can be represented by lower previsions on some family of gambles.

In this setting, the upper prevision $UP(f)$ of a gamble f is provably equal to $-LP(-f)$. The value $LP(f)$ is thus the maximal buying price for a gamble f, and the upper prevision $UP(f) (\geq LP(f))$ is the minimal selling price of f. If the credal set attached to a set of gambles and its lower previsions is empty, then the proposal is inconsistent and the agent incurs a sure loss after buying and resolving these gambles. Moreover, due to the interaction between gambles, it may be that the consistent buying prices proposed by the agent for gambles $f_i \in \mathscr{G}$ are too low and could be raised without altering the credal set. A set of buying prices $pr(f_i)$, $f_i \in \mathscr{G}$ is said to be *coherent* if and only if $LP(f_i) = pr(f_i), \forall f_i \in \mathscr{G}$. In other words, letting $E_P(f)$ be

the expectation of f with respect to probability P, a set of buying prices for a set of gambles $\mathscr{G}$ is coherent if and only if for any $f_i \in \mathscr{G}$, $\inf\{E_P(f_i) : P \in \mathscr{P}\} = pr(f_i)$, where $\mathscr{P}$ is the credal set induced by the gambles $f_i \in \mathscr{G}$, and their buying prices. Clearly, Giles condition (36) is easily interpreted in terms of coherence of gambles. It expresses the coherence of a set of upper probabilities assigned to subsets of S (minimal selling prices of 0-1 gambles), protecting an agent who sells $k+1$ lottery tickets corresponding to events $A_0, A_1, \ldots, A_k$ from losing money while proposing optimal selling prices $g(A_i)$.

The gamble approach leads to a decision rule that is specific to the imprecise probability setting, namely a gamble f is preferred to a gamble g if and only the gamble $h = f - g$ is desirable, i.e., if the lower expectation of the latter gamble with respect to the corresponding credal set $\mathscr{P}$ is positive. It gives a partial ordering on gambles. It implies that $\forall P \in \mathscr{P}$, $E_P(f) \geq E_P(g)$. See chapter "Decision Under Uncertainty" in this volume for other decision rules with credal sets

3.1.3 Special Cases

A monotonic set-function $g : 2^S \to [0, 1]$ is said to be a Walley-coherent lower probability if the following property holds:

$$g(A) = \inf\{P(A) : P(A) \geq g(A), \forall A \subseteq S\}.$$

In that case, the credal set $\mathscr{P} = \{P : P(A) \geq g(A), \forall A \subseteq S\}$ is characterized by the set-function g, that is, it can be described by assigning optimal buying prices to events (viewed as 0-1 gambles) only. Mind that not all credal sets can be characterized in this way. They generally require the assignment of buying (or selling) prices to general gambles. A sufficient condition for a monotonic set function to be Walley-coherent is the supermodularity condition: $g(A \cup B) + g(A \cap B) \geq g(A) + g(B)$. Such a function g is a called a *convex capacity*. So it is clear that other set-functions met in this chapter and the previous one are Walley-coherent as well, such as belief functions (equivalently plausibility functions) and necessity measures (equivalently possibility measures), which can represent specific credal sets.

Interestingly, Walley-coherence can be viewed as a generalization of deductive closure to families of weighted propositions. Let $\mathscr{K}$ be a consistent set of propositions $A_0, A_1, \ldots, A_k$, and suppose we assign the buying prices $pr(A_i) = 1, i = 0, \ldots k$, then $P_*(A) = 1$ if and only if $\mathscr{K} \models A$.

More about imprecise probability theories can be found in Walley (1991)'s book and their relevance for uncertainty management in artificial intelligence is discussed in Walley (1996), where the position of belief functions and possibility measures in the landscape is pointed out. More recent books on the topics are the collection of introductory papers edited by Augustin et al. (2014), and the mathematically oriented monograph on lower previsions by de Cooman and Troffaes (2014).

3.2 Two Types of Conditioning

In the framework of imprecise probabilities, there are several ways of extending the Bayesian conditioning of probability theory. It reflects the fact that the two usual tasks performed by Bayes rule, that is prediction and revision, can no longer be performed by the same conditioning rule (Dubois and Prade 1997b).

3.2.1 Prediction

When a credal set represents generic knowledge, Bayesian prediction or plausible inference is achieved by performing a form of sensitivity analysis on probabilistic conditioning, a rule proposed by Walley (1991), Fagin and Halpern (1991). Let $\mathscr{P}$ be a credal set on S. It induces lower and upper bounds $P_*(A)$ and $P^*(A)$ of the probability of each proposition A. In the presence of new pieces of information about a singular case, summarized by the context C, the belief of the agent that proposition A holds for the case at hand is represented by the interval $[P_*(A \mid C), P^*(A \mid C)]$ defined by

$$P_*(A \mid C) = \inf\{P(A \mid C) \text{ s.t. } P(C) > 0, P \in \mathscr{P}\}$$

$$P^*(A \mid C) = \sup\{P(A \mid C) \text{ s.t. } P(C) > 0, P \in \mathscr{P}\}.$$

Note that it is possible that interval $[P_*(A \mid C), P^*(A \mid C)]$ is larger than $[P_*(A), P^*(A)]$, which means that there is a deficit of information given by the credal set $\mathscr{P}$ in the specific context C, while there is more in more general contexts. This is called the dilation effect (Seidenfeld and Wasserman 1993). It reflects the fact that in the presence of incomplete information, the more observations are available on a singular case, the less relevant to this case is generic information about the population of cases, because the less the new case can be viewed as representative of this population. In the case of Bayes rule applied to a known frequentist distribution, this dilation effect does not appear because a single number is always obtained. However, this value becomes all the more dubious as the number of cases similar to the one under study in the population justifying the frequentist distribution becomes smaller and smaller as we condition on a more specific context.

If $\mathscr{P}$ is the credal set associated to a convex capacity (hence, belief functions, necessity measures as well) the upper and lower conditional functions take the remarkable forms (Fagin and Halpern 1991):

$$P_*(A \mid C) = \frac{P_*(A \cap C)}{P_*(A \cap C) + P^*(\overline{A} \cap C)}; \quad P^*(A \mid C) = \frac{P^*(A \cap C)}{P^*(A \cap C) + P_*(\overline{A} \cap C)} \tag{37}$$

It is easy to see that $P^*(A \mid C) = 1 - P_*(\overline{A} \mid C)$, and these formula extend probabilistic conditioning, in the sense that $P_*(A \mid C)$ is a function of $P_*(A \cap C)$ and $P^*(\overline{C} \cup A)$ (and similarly for $P^*(A \mid C)$). It is clear that this form of conditioning

does not correspond to the idea of enriching generic information by new observations, i.e., the latter do not alter the credal set. We just extract from it information that fits the available evidence, in the spirit of De Finetti.

In the theory of belief functions, the above form of conditioning can be justified in terms of their mass functions, positive weights $m(E)$ assigned to subsets E of S. When a mass function represents generic knowledge, $m(E)$ may be, e.g., the proportion of individuals, for which only imprecise proposition E is known to hold, in the whole population. In this setting, prediction in context C consists in evaluating mass function $m(\cdot \mid C)$ induced by m in context C summarizing the available singular information. Three cases can be considered (de Campos et al. 1990):

1. $E \subseteq C$: in that case, $m(E)$ remains committed to E;
2. $E \cap C = \emptyset$: in that case, $m(E)$ is no longer relevant and is discarded;
3. $E \cap C \neq \emptyset$ and $\overline{E} \cap C \neq \emptyset$: in that case, a part $\alpha_E \cdot m(E)$ of $m(E)$ remains committed to $E \cap C$ and the rest, i.e., $(1 - \alpha_E) \cdot m(E)$, is committed to $\overline{E} \cap C$. But the proportion α_E is unknown.

The third case corresponds to incomplete information E which neither confirms, nor contradicts C. We do not have information to determine if, in each of the situations corresponding to these observations, C is true or not. Assume that one knows the proportions $\{\alpha_E, E \subseteq S\}$. We always have $\alpha_E = 1$ in the first case and $\alpha_E = 0$ in the second case. One thus constructs a mass function $m_\alpha(\cdot \mid C)$. We can remark that renormalization of the resulting mass function is necessary whenever $Pl(C) < 1$: each mass is then divided by $Pl(C)$. Denoting by $Bel_\alpha(A \mid C)$ and $Pl_\alpha(A \mid C)$ the belief and plausibility obtained by focalization on C with vector of proportions α, we can define the conditional degrees of belief and of plausibility given C as

$$Bel(A \mid C) = \inf_\alpha Bel_\alpha(A \mid C); \quad Pl(A \mid C) = \sup_\alpha Pl_\alpha(A \mid C). \tag{38}$$

These definitions yield the following special cases of Bayesian conditioning for imprecise probability (37):

$$Bel(A \mid C) = \inf\{P(A \mid C) \text{ s.t. } P(C) > 0, P \geq Bel\} = \frac{Bel(A \cap C)}{Bel(A \cap C) + Pl(\overline{A} \cap C)}; \tag{39}$$

$$Pl(A \mid C) = \sup\{P(A \mid C) \text{ s.t. } P(C) > 0, P \geq Bel\} = \frac{Pl(A \cap C)}{Pl(A \cap C) + Bel(\overline{A} \cap C)}. \tag{40}$$

We still obtain belief and plausibility functions[2] (see the non-trivial proofs by Jaffray 1992 and Paris 1994). Let us notice that if $Bel(C) = 0$ and $Pl(C) = 1$ (total ignorance about C) then all focal sets of m overlap C but C does not contain any of them. In that case, $Bel(A \mid C) = 0$ and $Pl(A \mid C) = 1, \forall A \neq S, \emptyset$: nothing can be inferred in context C.

[2]When applied to necessity and plausibility measures, these two formulas also preserve consonance and yield another form of conditional possibility and necessity (Dubois and Prade 1997a).

3.2.2 Revision

In the framework of imprecise probabilities, a simple brute force approach to revision of a credal set $\mathscr{P}$ by an information item C consists in enforcing the additional constraint $P(C) = 1$ to $\mathscr{P}$, namely restrict the latter, and update the upper and lower probabilities of events accordingly:

$$P_*(A \mid\mid C) = \inf\{P(A \mid C) \text{ s.t. } P(C) = 1, P \in \mathscr{P}\}; \tag{41}$$

$$P^*(A \mid\mid C) = \sup\{P(A \mid C) \text{ s.t. } P(C) = 1, P \in \mathscr{P}\}. \tag{42}$$

Clearly, it is supposed, in contrast with the assumption in the prediction problem, that the new item of information is of the same nature as the original credal set, and can be modelled by the credal set $\{P : P(C) = 1\}$ (it can be frequentist or subjectivist).

However, by doing so, it may be that the intersection of the two credal sets, i.e., $\{P \in \mathscr{P} \text{ s.t. } P(C) = 1\}$ is empty. This is for instance most of the time the case in the standard probabilistic setting since $\mathscr{P}$ reduces to a singleton. The way out is to apply the maximum likelihood principle (Gilboa and Schmeidler 1992), selecting the most likely probability functions in $\mathscr{P}$, replacing condition $P(C) = 1$ by $P(C) = P^*(C)$ in the above definition of conditioning:

$$P_*(A \mid\mid C) = \inf\{P(A \mid C) \text{ s.t. } P(C) = P^*(C), P \in \mathscr{P}\}; \tag{43}$$

$$P^*(A \mid\mid C) = \sup\{P(A \mid C) \text{ s.t. } P(C) = P^*(C), P \in \mathscr{P}\}. \tag{44}$$

For convex capacities, it holds that $P^*(A \mid\mid C) = \frac{P^*(A \cap C)}{P^*(C)}$, which generalizes Dempster rule of conditioning. In the belief function setting, this form of conditioning systematically assumes that $\alpha_E = 1$ whenever $E \cap C \neq \emptyset$ in $Bel_\alpha(A \mid C)$ and $Pl_\alpha(A \mid C)$. From the perspective of Shafer and Smets, mass function m does not represent generic information, but uncertain singular information, such as unreliable testimonies or inconclusive pieces of evidence about a specific situation. The existence of two forms of conditioning in the theory of belief functions can thus be explained by the difference between generic and singular information.

As a general setting for the numerical representation of uncertainty, liable of various interpretations, and encompassing other theories of uncertainty as formal particular cases, imprecise probabilities receive an increasing attention and foster a number of theoretical works; for instance, de Cooman and Hermans (2008) build bridges between Walley's approach to imprecise probabilities and the game-theoretic view of probability by Shafer and Vovk (2001). Practical representation methods in artificial intelligence are also studied, for instance the imprecise probability version of Bayesian nets, including dedicated uncertainty propagation algorithms (Cozman 2000; de Campos and Cozman 2005; Cozman and Mauá 2017).

4 Conclusion

Artificial Intelligence, when focusing on representation and reasoning with imperfect information, was naturally bound to realize that classical logic on the one hand, and precise probabilities on the other hand, were separately insufficient to deal with this issue. Alternative formal frameworks have emerged in the last 40 years or so to that effect, that this chapter partially accounts for. These frameworks are numerous and often complement each other rather than compete, even if research in this area remains fragmented. Nevertheless, these alternative theories of uncertain, incomplete or conflicting information offer a very rich range of formalisms. It is important to correctly understand their potentials and limitations prior to appropriately exploiting them. These frameworks can be qualitative (like possibilistic logic, discussed in the previous chapter) or quantitative (like belief functions and imprecise probabilities). A significant effort is still needed before a full-fledged unification of the various approaches is achieved, and the links with neighboring disciplines like statistics are fully established, in order to master their use in applications.

References

Antoine V, Quost B, Masson MH, Denoeux T (2012) CECM: constrained evidential c-means algorithm. Comput Stat Data Anal 56(4):894–914

Antoine V, Quost B, Masson MH, Denoeux T (2014) CEVCLUS: evidential clustering with instance-level constraints for relational data. Soft Comput 18(7):1321–1335

Appriou A (1991) Probabilités et incertitude en fusion de données multi-senseurs. Revue Scientifique et Technique de la Défense 11:27–40

Appriou A (1998) Uncertain data aggregation in classification and tracking processes. In: Bouchon-Meunier B (ed) Aggregation and fusion of imperfect information. Physica-Verlag, Heidelberg, pp 231–260

Augustin T, Coolen F, de Cooman G, Troffaes M (eds) (2014) Introduction to imprecise probabilities. Wiley, New York

Baudrit C, Dubois D (2006) Practical representations of incomplete probabilistic knowledge. Comput Stat Data Anal 51(1):86–108

Bauer M (1997) Approximation algorithms and decision making in the Dempster-Shafer theory of evidence - an empirical study. Int J Approx Reason 17:217–237

Berger JO (1994) An overview of robust Bayesian analysis. Test 3:5–124, with discussion

Bezdek J (1981) Pattern recognition with fuzzy objective function algorithm. Plenum Press, New York

Bi Y (2012) The impact of diversity on the accuracy of evidential classifier ensembles. Int J Approx Reason 53(4):584–607

Bi Y, Guan J, Bell D (2008) The combination of multiple classifiers using an evidential reasoning approach. Artif Intell 172(15):1731–1751

Cattaneo MEGV (2011) Belief functions combination without the assumption of independence of the information sources. Int J Approx Reason 52(3):299–315

Couso I, Dubois D, Sanchez L (2014) Random sets and random fuzzy sets as ill-perceived random variables. Springer briefs in computational intelligence. Springer, Berlin. http://www.springerlink.com

Cozman FG (2000) Credal networks. Artif Intell 120:199–233

Cozman FG, Mauá DD (2017) On the complexity of propositional and relational credal networks. Int J Approx Reason 83:298–319
de Campos CP, Cozman FG (2005) The inferential complexity of Bayesian and credal networks. In: Kaelbling LP, Saffiotti A (eds) Proceedings of the 19th international joint conference on artificial intelligence (IJCAI'05). AAAI Press, pp 1313–1318
de Campos LM, Huete JF, Moral S (1994) Probability intervals: a tool for uncertain reasoning. Int J Uncertain Fuzziness Knowl-Based Syst 2:167–196
de Campos LM, Lamata MT, Moral S (1990) The concept of conditional fuzzy measure. Int J Intell Syst 5:237–246
de Cooman G, Hermans F (2008) Imprecise probability trees: bridging two theories of imprecise probability. Artif Intell 172:1400–1427
de Cooman G, Troffaes M (2014) Lower previsions. Wiley, New York
Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Stat 28:325–339
Dempster AP (1968) Upper and lower probabilities generated by a random closed interval. Ann Math Stat 39(3):957–966
Denœux T (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans Syst Man Cybern 25(05):804–813
Denœux T (1997) Analysis of evidence-theoretic decision rules for pattern classification. Pattern Recognit 30(7):1095–1107
Denœux T (1999) Reasoning with imprecise belief structures. Int J Approx Reason 20:79–111
Denœux T (2000a) Modeling vague beliefs using fuzzy-valued belief structures. Fuzzy Sets Syst 116(2):167–199
Denœux T (2000b) A neural network classifier based on Dempster-Shafer theory. IEEE Trans Syst Man Cybern A 30(2):131–150
Denœux T (2001) Inner and outer approximation of belief structures using a hierarchical clustering approach. Int J Uncertain Fuzziness Knowl-Based Syst 9(4):437–460
Denœux T (2008) Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. Artif Intell 172:234–264
Denœux T (2009) Extending stochastic ordering to belief functions on the real line. Inf Sci 179:1362–1376
Denœux T (2016) EVCLUST: evidential clustering. https://CRAN.R-project.org/package=evclust, R package version 1.0.3
Denœux T (2017) EVCLASS: evidential distance-based classification. https://CRAN.R-project.org/package=evclass, R package version 1.1.1
Denœux T, Ben Yaghlane A (2002) Approximating the combination of belief functions using the fast Mœbius transform in a coarsened frame. Int J Approx Reason 31(1–2):77–101
Denœux T, Kanjanatarakul O (2016) Beyond fuzzy, possibilistic and rough: an investigation of belief functions in clustering. In: Soft methods for data science (Proceedings of SMPS 2016). Advances in intelligent and soft computing, vol AISC 456. Springer, Berlin, pp 157–164
Denœux T, Kanjanatarakul O, Sriboonchitta S (2015) EK-NNclus: a clustering procedure based on the evidential k-nearest neighbor rule. Knowl-Based Syst 88:57–69
Denœux T, Masson MH (2004) EVCLUS: evidential clustering of proximity data. IEEE Trans Syst Man Cybern B 34(1):95–109
Denœux T, Masson MH (2012) Evidential reasoning in large partially ordered sets. Application to multi-label classification, ensemble clustering and preference aggregation. Ann Oper Res 195(1):135–161
Denœux T, Smets P (2006) Classification using belief functions: the relationship between the case-based and model-based approaches. IEEE Trans Syst Man Cybern B 36(6):1395–1406
Denœux T, Sriboonchitta S, Kanjanatarakul O (2016) Evidential clustering of large dissimilarity data. Knowl-Based Syst 106:179–195
Denœux T, Younes Z, Abdallah F (2010) Representing uncertainty on set-valued variables using belief functions. Artif Intell 174(7–8):479–499

Denœux T, Zouhal LM (2001) Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets Syst 122(3):47–62
Destercke S, Burger T (2013) Toward an axiomatic definition of conflict between belief functions. IEEE Trans Cybern 43(2):585–596
Destercke S, Dubois D (2011) Idempotent conjunctive combination of belief functions: extending the minimum rule of possibility theory. Inf Sci 181(18):3925–3945
Dubois D, Prade H (1982) On several representations of an uncertain body of evidence. In: Gupta MM, Sanchez E (eds) Fuzzy information and decision processes. North-Holland, pp 167–181
Dubois D, Prade H (1985) A note on measures of specificity for fuzzy sets. Int J Gen Syst 10(4):279–283
Dubois D, Prade H (1986) A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. Int J Gen Syst 12:193–226
Dubois D, Prade H (1987) Properties of information measures in evidence and possibility theories. Fuzzy Sets Syst 24:161–182
Dubois D, Prade H (1988) Representation and combination of uncertainty with belief functions and possibility measures. Comput Intell (Canada) 4(4):244–264
Dubois D, Prade H (1997a) Bayesian conditioning in possibility theory. Fuzzy Sets Syst 92:223–240
Dubois D, Prade H (1997b) Focusing vs. belief revision: a fundamental distinction when dealing with generic knowledge. In: Gabbay DM, Kruse R, Nonnengart A, Ohlbach HJ (eds) Qualitative and quantitative practical reasoning (Proceedings of ECSQARU-FAPR'97). LNCS, vol 1244. Springer, Berlin, pp 96–107
Dubois D, Prade H, Smets P (2001) "Not impossible" vs. "guaranteed possible" in fusion and revision. In: Benferhat S, Besnard P (eds) Symbolic and quantitative approaches to reasoning with uncertainty (Proceedings of ECSQARU'01). LNCS, vol 2143. Springer, Berlin, pp 522–531
Dubois D, Prade H, Smets P (2008) A definition of subjective possibility. Int J Approx Reason 48:352–364
Dubois D, Ramer A (1993) Extremal properties of belief measures in the theory of evidence. Int J Uncertain Fuzziness Knowl-Based Syst 1(1):57–68
Fabre S, Appriou A, Briottet X (2001) Presentation and description of two classification methods using data fusion based on sensor management. Inf Fusion 2(1):49–71
Fagin R, Halpern J (1991) A new approach to updating beliefs. In: Bonissone PP, Henrion M, Kanal LN, Lemmer JF (eds) Uncertainty in artificial intelligence, vol 6. North-Holland, Amsterdam, pp 347–374
Gilboa I, Schmeidler D (1989) Maxmin expected utility with a non-unique prior. J Math Econ 18:141–153
Gilboa I, Schmeidler D (1992) Updating ambiguous beliefs. In: Moses Y (ed) Proceedings of the 4th conference on theoretical aspects of reasoning about knowledge (TARK'92). Morgan Kaufmann, pp 143–162
Giles R (1982) Foundations for a theory of possibility. In: Gupta MM, Sanchez E (ed) Fuzzy information and decision processes. North-Holland, pp 183–195
Good IJ (1962) Subjective probability as the measure of a non-measurable set. In: Nagel E, Suppes P, Tarski A (eds) Handbook of the history of logic. Stanford University Press, pp 319–329
Grabisch M (2009) Belief functions on lattices. Int J Intell Syst 24:76–95
Harmanec D (1999) Faithful approximations of belief functions. In: Laskey KB, Prade H (eds) Uncertainty in artificial intelligence (Proceedings of UAI99), Stockholm
Harmanec D, Klir GJ (1994) Measuring total uncertainty in Dempster-Shafer theory: a novel approach. Int J Gen Syst 22(4):405–419
Huber P (1981) Robust statistics. Wiley, New York
Jaffray JY (1989) Linear utility theory for belief functions. Oper Res Lett 8(2):107–112
Jaffray JY (1992) Bayesian updating and belief functions. IEEE Trans Syst Man Cybern 22:1144–1152

Jiroušek R, Shenoy PP (2016) Entropy of belief functions in the Dempster-Shafer theory: a new perspective. In: Vejnarová J, Kratochvíl V (eds) Belief functions: theory and applications (Proceedings of BELIEF 2016). Springer, Berlin, pp 3–13

Kennes R (1992) Computational aspects of the Möbius transformation of graphs. IEEE Trans Syst Man Cybern 22:201–223

Klir GJ, Wierman MJ (1999) Uncertainty-based information. Elements of generalized information theory. Springer, New York

Lefèvre E, Colot O, Vannoorenberghe P (2002) Belief function combination and conflict management. Inf Fusion 3(2):149–162

Lelandais B, Ruan S, Denœux T, Vera P, Gardin I (2014) Fusion of multi-tracer PET images for dose painting. Med Image Anal 18(7):1247–1259

Li F, Li S, Denœux T (2018) k-CEVCLUS: constrained evidential clustering of large dissimilarity data. Knowl-Based Syst 142:29–44

Lian C, Ruan S, Denœux T (2015) An evidential classifier based on feature selection and two-step classification strategy. Pattern Recognit 48:2318–2327

Lian C, Ruan S, Denœux T (2016) Dissimilarity metric learning in the belief function framework. IEEE Trans Fuzzy Syst 24(6):1555–1564

Lingras P, Peters G (2012) Applying rough set concepts to clustering. In: Peters G, Lingras P, Ślezak D, Yao Y (eds) Rough sets: selected methods and applications in management and engineering. Springer, London, UK, pp 23–37

Lowrance JD, Garvey TD, Strat TM (1986) A framework for evidential-reasoning systems. In: Proceedings of the national AI conference (AAAI'86), vol 2. AAAI Press, pp 896–903

Maeda Y, Ichihashi H (1993) An uncertainty measure with monotonicity under the random set inclusion. Int J Gen Syst 21(4):379–392

Martin A, Osswald C, Dezert J, Smarandache F (2008) General combination rules for qualitative and quantitative beliefs. J Adv Inf Fusion 3(2):67–82

Masson MH, Denœux T (2008) ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recognit 41(4):1384–1397

Masson MH, Denœux T (2009) RECM: relational evidential c-means algorithm. Pattern Recognit Lett 30:1015–1026

Mercier D, Quost B, Denœux T (2008) Refined modeling of sensor reliability in the belief function framework using contextual discounting. Inf Fusion 9(2):246–258

Mercier D, Cron G, Denœux T, Masson MH (2009) Decision fusion for postal address recognition using belief functions. Expert Syst Appl 36(3):5643–5653

Mercier D, Lefèvre E, Delmotte F (2012) Belief functions contextual discounting and canonical decompositions. Int J Approx Reason 53(2):146–158

Mercier D, Pichon F, Lefèvre E (2016) Corrigendum to "Belief functions contextual discounting and canonical decompositions" [Int J Approx Reason 53:146–158 (2012)]. Int J Approx Reason 70:137–139

Moral S, Wilson N (1994) Markov-chain Monte-Carlo algorithms for the calculation of Dempster-Shafer belief. In: Proceedings of the twelfth national conference on artificial intelligence (AAAI-94), vol 1, pp 269–274

Moral S, Wilson N (1996) Importance sampling Monte-Carlo algorithms for the calculation of Dempster-Shafer belief. In: Proceedings of international conference on information processing and management of uncertainty (Proceedings of IPMU'96), Granada, Spain, vol III, pp 1337–1344

Nguyen H (2006) An introduction to random sets. Chapman and Hall/CRC, Boca Raton

Nguyen HT (1978) On random sets and belief functions. J Math Anal Appl 65:531–542

Paris J (1994) The uncertain reasoner' companion. Cambridge University Press, Cambridge

Pearl J (1990) Reasoning with belief functions: an analysis of compatibility. Int J Approx Reason 4(5):363–389

Petit-Renaud S, Denœux T (2004) Nonparametric regression analysis of uncertain and imprecise data using belief functions. Int J Approx Reason 35(1):1–28

Pichon F, Denœux T (2010) The unnormalized Dempster's rule of combination: a new justification from the least commitment principle and some extensions. J Autom Reason 45(1):61–87
Pichon F, Denœux T, Dubois D (2012) Relevance and truthfulness in information correction and fusion. Int J Approx Reason 53(2):159–175
Pichon F, Mercier D, Delmotte F (2016) Proposition and learning of some belief function contextual correction mechanisms. Int J Approx Reason 72:4–42
Quost B, Denœux T, Masson MH (2007) Pairwise classifier combination using belief functions. Pattern Recognit Lett 28(5):644–653
Quost B, Masson MH, Denœux T (2011) Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. Int J Approx Reason 52(3):353–374
Ramer A (1987) Uniqueness of information measure in the theory of evidence. Fuzzy Sets Syst 24:183–196
Ramer A, Klir GJ (1993) Measures of discord in the Dempster-Shafer theory. Inf Sci 67:35–50
Seidenfeld T, Wasserman L (1993) Dilation for sets of probabilities. Ann Stat 21:1139–1154
Shafer G (1973) Allocations of probability: a theory of partial belief. PhD thesis, Princeton University
Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
Shafer G (1979) Allocations of probability. Ann Probab 7(5):827–839
Shafer G (1981) Constructive probability. Synthese 48(1):1–60
Shafer G (1987) Probability judgment in artificial intelligence and expert systems. Stat Sci 2(1):3–44
Shafer G (1990) Perspectives in the theory and practice of belief functions. Int J Approx Reason 4:323–362
Shafer G (2016a) Dempster's rule of combination. Int J Approx Reason 79:26–40
Shafer G (2016b) A mathematical theory of evidence turns 40. Int J Approx Reason 79:7–25
Shafer G (2016c) The problem of dependent evidence. Int J Approx Reason 79:41–44
Shafer G, Vovk V (2001) Probability and finance: it's only a game!. Wiley, New York
Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games, volume II. Annals of mathematical studies series, vol 28. Princeton University Press, Princeton, pp 307–317
Shenoy PP (1989) A valuation-based language for expert systems. Int J Approx Reason 3:383–411
Smets P (1981) The degree of belief in a fuzzy event. Inf Sci 25:1–19
Smets P (1983) Information content of an evidence. Int J Man-Mach Stud 19:33–43
Smets P (1990a) The combination of evidence in the transferable belief model. IEEE Trans Pattern Anal Mach Intell 12(5):447–458
Smets P (1990b) Constructing the pignistic probability function in a context of uncertainty. In: Henrion M, Shachter RD, Kanal LN, Lemmer J (eds) Uncertainty in artificial intelligence, vol 5. Elsevier Science Publication, Amsterdam, pp 29–39
Smets P (1993) Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. Int J Approx Reason 9:1–35
Smets P (1995) The canonical decomposition of a weighted belief. In: Proceedings of the international joint conference on artificial intelligence, Morgan Kaufman, San Mateo, CA, pp 1896–1901
Smets P (2005) Belief functions on real numbers. Int J Approx Reason 40:181–223
Smets P (2007) Analyzing the combination of conflicting belief functions. Inf Fusion 8(4):387–412
Smets P, Kennes R (1994) The transferable belief model. Artif Intell 66:191–234
Smith CAB (1961) Consistency in statistical inference and decision. J R Stat Soc B 23:1–37
Strat TM (1984) Continuous belief functions for evidential reasoning. In: Brachman RJ (ed) Proceedings of the national conference on artificial intelligence (AAAI'84), Austin, Aug 6–10, pp 308–313
Strat TM (1990) Decision analysis using belief functions. Int J Approx Reason 4(5–6):391–417
Tessem B (1993) Approximations for efficient computation in the theory of evidence. Artif Intell 61:315–329
von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton

Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, Boca Raton
Walley P (1996) Measures of uncertainty in expert systems. Artif Intell 83:1–58
Walley P, Fine T (1982) Towards a frequentist theory of upper and lower probability. Ann Stat 10:741–761
Xu L, Krzyzak A, Suen CY (1992) Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans Syst Man Cybern 22(3):418–435
Yager RR (1986) The entailment principle for Dempster-Shafer granules. Int J Intell Syst 1:247–262
Yager RR (1992) Decision making under Dempster-Shafer uncertainties. Int J Gen Syst 20(3):233–245
Yager RR, Liu LP (eds) (2008) Classic works of the Dempster-Shafer theory of belief functions. Springer, Heidelberg
Yen J (1990) Generalizing the Dempster-Shafer theory to fuzzy sets. IEEE Trans Syst Man Cybern 20(3):559–569
Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta MM, Ragade RK, Yager RR (eds) Advances in fuzzy sets theory and applications. North-Holland, Amsterdam, pp 3–18
Zouhal LM, Denœux T (1998) An evidence-theoretic k-NN rule with parameter optimization. IEEE Trans Syst Man Cybern C 28(2):263–271

Qualitative Reasoning

Jean-François Condotta, Florence Le Ber, Gérard Ligozat and Louise Travé-Massuyès

Abstract In this chapter, we discuss two research areas related to *qualitative reasoning*: firstly, *qualitative reasoning about dynamical systems*, or *qualitative physics*, that aims at providing qualitative descriptions of processes in the sense that they are characterized regardless of quantitative data (for instance, "the tank overflows", "temperature increases", etc.); and secondly *qualitative spatial and temporal reasoning* (QSTR), that aims at describing and reasoning about qualitative relationships between spatial regions ("the stadium is on the island", "the bike path crosses the river") or between time periods ("the minister's visit preceded the opening of the Olympic Games").

1 Introduction

At the very start of the 1980s — actually, in 1979 — the *Naive Physics Manifesto* by Hayes (1979) became the starting point of Qualitative Physics by claiming that an "intelligent machine" should have a model of the surrounding world and be able to anticipate what may or may not occur.

In that paper and in the revised version that followed in 1985 (Hayes 1985), the problem of modeling our *common sense* perception of the physical world was formulated and illustrated the same year by an axiomatization in first order logic of the "intuitive" behavior of liquids with *An Ontology for Liquids*. This was called

J.-F. Condotta (✉)
CRIL-CNRS and Université d'Artois, Lens, France
e-mail: condotta@cril.univ-artois.fr

F. Le Ber
ICube, Université de Strasbourg, ENGEES, CNRS, Strasbourg, France
e-mail: florence.leber@engees.unistra.fr

G. Ligozat
LIMSI-CNRS, Orsay, France
e-mail: gligozat@gmail.com

L. Travé-Massuyès
LAAS-CNRS, Toulouse, France
e-mail: louise@laas.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_5

more precisely *Naive Physics*. This project was very ambitious, more because of the amount of knowledge to be encoded it implied than for the complexity of the reasoning to be implemented, and its impact on our intelligent systems promised to be huge since they would have to be able to predict the *qualitative* features of the possible evolutions of the world.

Despite the attractiveness of such a project and the impact of these papers, the common sense reasoning of Naive Physics was quickly overtaken by what became known as *Qualitative Physics* or *Qualitative Reasoning*. Indeed, in 1977, in parallel to the programmatic call of Pat Hayes, the MIT launched a project aiming at the "creation" of an *artificial engineer*. The target was thus the knowledge and expertise of an engineer reasoning about an artifact or a natural system, halfway between Naive Physics and "classical" physics found in textbooks.

The area of *Qualitative spatial and temporal reasoning* can also avail itself of Hayes's work. Emphasizing the importance of the representation of space and change, Hayes had introduced the notion of *history*, a kind of space-time region. Hence, in his pioneering paper (Allen 1983), Allen refers to his own work as "describing a reasoning mechanism for the temporal aspect of Naive Physics".

The focus is indeed put on "common sense", in accordance with the qualification of "naive". Rather than representing the advanced knowledge of engineers, the goal is to represent the common sense knowledge that manifests itself especially in the use of natural language. It is also of importance to note that spatial qualitative reasoning predominantly prefers conceptualisations that reject the geometric concept of point — in the same way as Allen rejects that of instant — in favor of mereological visions of space. The research of the Leeds School (Randell et al. 1992b) and of the Toulouse School (Vieu 1991) typically adopts that point of view.

We now describe the two main directions of qualitative reasoning: *qualitative physics* and *qualitative spatial and temporal reasoning* (QSTR).

2 Qualitative Physics

Qualitative physics aims at automating the reasoning about the physical world, a central goal of Artificial Intelligence. Qualitative modeling and inferences about the behavior of a physical system where information is incomplete are two inputs of qualitative physics. The theoretical foundations have resulted in new mathematical tools that have had many practical spinoffs, resulting in several real world applications. We present different aspects, starting with theoretical work on qualitative algebras, up to diagnosis and interactive learning applications.

2.1 Historical Outline

One of the pioneering works is undoubtedly that of de Kleer who, in 1977, designed the *Newton* system, a system that was able to solve qualitatively simple mechan-

ical problems (de Kleer 1977). It was followed by programs able to reason about electrical circuits. De Kleer's *Local* system used its knowledge of normal and faulty behaviors of components of a circuit for detecting inconsistencies between observed and predicted behaviors and then locate the faults in the circuit (de Kleer 1979). Those ideas were to become the basic ideas of *model-based diagnosis* theory.

The concerns of Artificial Intelligence researchers thus met the body of work initiated by other scientific communities. Economists had already proposed qualitative approaches in the 1960s (Lancaster 1965). Their work highlighted the fact that the mere knowledge of the signs $(+, -, 0)$ of a few variables is sometimes sufficient to predict behavioral trends for a system. Qualitative analysis thus allows one to distinguish the purely structural causes of an evolution from those due to specific numerical configurations. Similarly, during the 1970s, some environmentalists developed formalisms to model a system in terms of signed graphs. That work was taken over by the automatic control community in the 1980s, which proposed extensions for the analysis of dynamical systems, in particular for assessing controllability and observability (Travé and Kaszkurewicz 1986).

Those communities took as a point of departure the observation that quantitative knowledge being often incomplete, building a numeric model is somehow arbitrary, while qualitative knowledge may provide relevant elements for analysis. The AI community, on the other hand, had a quite different agenda, its goal being to automate the kind of reasoning performed by humans and engineers, in particular when reasoning about physical phenomena. As a consequence, it has payed special attention to the issue of explanation, seeking not only to predict the behavior of a system but also to explain it. *Causal reasoning* has been a constant concern of qualitative reasoning. Similarly, model building and the formalization of the modeling process are its core activities and they constitute one of its major contributions.

In France, research related to qualitative reasoning developed as a craze, orchestrated by the working group *Qualitative Modeling and Decision* (MQD), which then had no less than fifty members. Under the impulse of MQD, a special issue of the journal *Revue d'Intelligence Artificielle* appeared in 1990 (Travé-Massuyès and Dormoy 1990) followed by a survey paper (Dague 1995) and some years later the collective book entitled *Qualitative Reasoning for Engineering Sciences* (Travé-Massuyès et al. 1997) whose second edition was published in the Hermès Collection in 2003 (Travé-Massuyés and Dague 2003). Echoing MQD, the European Network of Excellence MONET[1] gathered researchers and organized QR activities in Europe for about six years, from 1999 to 2004. These activities were punctuated by the annual workshop on Qualitative Reasoning that is still held every year.

French contributions are essentially theoretical, providing formalisms for reasoning about *orders of magnitude*, either *absolute* or *relative*, main contributors being J.-L. Dormoy and L. Travé-Massuyès for the former and O. Raiman and P. Dague for the latter.

[1] MONET: Network of Excellence on Model Based Systems and Qualitative Reasoning.

2.2 Different Aspects of Qualitative Reasoning

The underlying principles of qualitative reasoning provide several benefits over traditional analytical techniques often based on numerical simulation. Specifically, qualitative reasoning:

- *can deal with incomplete knowledge* that would be useless for conventional simulation methods;
- *produces inaccurate but correct predictions* that capture all possible behaviors consistent with the incomplete specification. This contrasts with numerical simulation that requires a specific value for each parameter, producing accurate but probably incorrect results (because precise values are often unknown);
- *enables an easy exploration of alternatives* which are obtained by qualitative simulation in a single execution. Exploring the same set of possible behaviors via numerical simulation requires a great number of executions, one for each different parameter value, with no guarantee of completeness;
- *provides an automatic interpretation of results* since evolutions are given as a function of the relevant qualitative characteristics used as input, whereas analyzing the result of a numerical simulation requires the identification of those characteristics by the user.

These properties are obtained thanks to specific conceptual approaches and mathematical formalisms specifically designed for qualitative reasoning.

2.2.1 Qualitative Abstractions

Qualitative reasoning is mainly concerned with systems with continuous dynamics for which a standard model would consist of differential equations. Qualitative abstractions aim at providing behavioral models that only retain the qualitative distinctions needed to solve a specific problem attached to a task. For this purpose, the two main abstractions used in qualitative reasoning are:

- *domain abstraction* that, for each continuous variable, consists in discretizing its value domain into a finite number of symbols, retaining its semantics in terms of *orders of magnitude*;
- *functional abstraction* that retains only some properties of the functions. A typical example is given by the qualitative operators $M+(y,x)$ and $M-(y,x)$ used by the QSIM simulator (Kuipers 1986), whose interpretation is that y is linked to x by an increasing, respectively decreasing, monotonic function, with no need to specify the function itself.

Domain abstraction has been the subject of much work. On the one hand, we find *qualitative algebras* or *Q-algebras* and *qualitative calculi* that makes use of a qualitative equality $\approx$ and of the two qualitative operators sum $\oplus$ and product $\otimes$. On the other hand, we have formal models that were proposed to represent relative

orders of magnitude. Actually, the two lines of work echo two types of knowledge about orders of magnitude:

- *absolute orders of magnitude* as used by the physicist when he/she approximates a number by the powers of 10. The most common absolute order of magnitude algebra is based on a partition of the real axis into 7 classes, corresponding to the symbols Negative Large (NL), Negative Medium (NM), Negative Small (NS), Zero (0), Positive Small (PS), Positive Medium (PM) and Positive Large (PL). Order of magnitude algebras generalize *sign algebra*, based on three classes Negative (–), Zero (0) and Positive (+), the most commonly used in qualitative inference systems. At the other end, interval algebra (Moore 1966) is also an order of magnitude algebra. Powerful algebraic properties have been identified for sign algebra (Travé and Dormoy 1988), whereas general order of magnitude algebras have weaker properties (Travé-Massuyès and Piera 1989).
- *relative orders of magnitude* as used by the physicist when he/she neglects a quantity relative to another. The first system to be proposed was the formal system FOG (Raiman 1991). Inspired by Non-Standard Analysis, it was based on three basic relationships representing the intuitive concepts "negligible with respect to" (Ne), "close to" (Vo) and "comparable to" (Co), and was axiomatized by 32 intuitive inference rules. The models ROM(K) and ROM($\Re$) proposed later by Dague (1993a, b) improved FOG's formalization and allowed the incorporation of quantitative information, producing valid results in the real world.

2.2.2 Generic and Modular Modeling

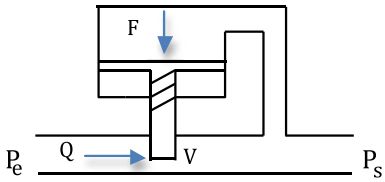
Qualitative reasoning explicitly represents the conceptual level of modeling. This feature is crucial to enable automatic model generation, the goal being to avoid developing the models by hand when considering different variants of a given system or specific purposes. Thus the two principles guiding knowledge representation in qualitative reasoning are *genericity* and *modularity*. Most qualitative reasoning systems make use of a model library of basic components or processes – if one refers to the two main qualitative reasoning ontologies respectively – that can be reused in different configurations and allow automatic modeling by composing these *model fragments* (Forbus 1984).

It must be noticed that the engineering world has a great expertise at building numerical models. Hence, mathematical foundations that would allow one to connect numerical models to qualitative models and help in the automatic model generation process are of extreme importance (Forbus 1984).

2.2.3 Qualitative Simulation

As already noticed, the information available in some areas is inherently qualitative. This is true of most natural systems. Moreover, the behavior of the system for precise

Fig. 1 A pressure regulator



parameter and initial condition values is not always of interest. It may be more interesting to know what types of qualitatively different behaviors are licensed by the structure of the system, given some constraints on the parameters and initial conditions. These objectives, at the core of qualitative reasoning, have given birth to *qualitative simulation*.

Three now classical approaches can be distinguished: the component-based approach by de Kleer and Brown (1984), the process-based approach by Forbus (1984) and the constraint-based approach by Kuipers (1986). The issues raised by these approaches have guided much of the work on qualitative simulation up to now.

In the ENVISION approach developed by de Kleer and Brown (1984), a qualitative model is expressed as a set of *confluences*, i.e. constraints on the qualitative values (here signs) of system variables. Confluences allow one to infer the qualitative values of unknown variables, leading to a specific representation of the behavior of the system.

This representation, called an *envisionment* is given by a graph that includes all the qualitative states of the system that satisfy the system confluences, as well as all the possible transitions between these states.

Consider, for example, the pressure regulator of Fig. 1, where Q is the flow rate of fluid in the pipe, P_e and P_s are respectively the pressure at the input and output of the regulator, V is the speed of the valve (when opening or closing), and F is the force exerted on the piston. The resulting envisionment is given in Fig. 2. The possible behaviors of the system, starting from a given initial state, are given by the different paths in the diagram as a sequence of states chronologically ordered. We distinguish between instantaneous states (symbolized by circles) and states with non-zero duration (denoted by squares).

For example, the sequence [5, 4, 5, 1, 2, 3] represents a possible qualitative behavior of the regulator. It indicates that, for a moment, the system can oscillate between zero and positive force F exerted on the piston with a negative speed (valve closing), to move to states in 1 and 2 where F is negative, then move to state 3, where the speed and force become simultaneously negative.

Qualitative Process Theory (QPT), developed by Forbus (1984), provides a process-centered view of the physical world. Given a scenario of a particular situation and a knowledge base of abstract model fragments describing objects, quantities, relations between objects and processes in the domain, QPT generates a model of the physical system under study. Basically, a model is a set of constraints, called *influences*, over the qualitative values of the variables. In a way similar to the

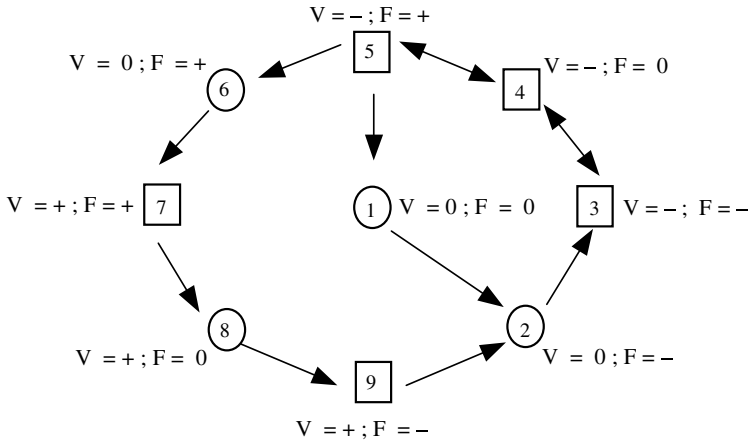


Fig. 2 Envisionment for the pressure regulator

ENVISION approach, qualitative simulation is used to build an envisionment describing the possible behaviors of the system.

Unlike ENVISION and QPT approaches, the QSIM approach by Kuipers (1986) ignores the aspects related to model building. The qualitative model of a system is a *qualitative differential equation*, which is an abstraction of a class of ordinary differential equations. Qualitative simulation makes use of continuity properties, given that variables are functions of time. It also considers the constraints over the values of qualitative variables involved in the qualitative differential equation. It produces the sequences of possible qualitative states, representing qualitative behaviors. QSIM is by far the most popular approach. The clear definition of concepts such as qualitative model, qualitative state and qualitative behavior as well as the explicit relationship with numerical simulation facilitated the adaptation and integration of the mathematical properties of ordinary differential equations. Kuipers's excellent book (Kuipers 1994) is also one of the key reasons of the success of the approach.

2.2.4 Causality

Causality (see chapter "A Glance at Causality Theories for Artificial Intelligence" of this volume) is one of the essential concepts when reasoning about physical systems. In many cases, the prediction of a given behavior is stated according to the cause-effect relations underlying this behavior. Causal knowledge allows one to infer the behavior but also to explain it. Therefore it is commonly used in tasks such as design or diagnosis. In the fields of engineering, there are often highly structured theories that underlie the behavior of systems. The Qualitative Reasoning community proposed automatic methods to discover causal relations with a view to enrich such theories with additional causal knowledge.

For a given physical system, for which there is a model defined by a set of constraints between a set of variables, the *causal ordering* problem is formulated as the problem of deriving the set of causal relationships, also called *influences*, between the variables involved in the model, given a subset of variables specified as exogenous. A variable is exogenous when it is controlled by the system's environment. The following three principles are generally accepted as underlying any causal influence: temporal order, locality and necessity.

The main methods that were proposed by the Qualitative Reasoning community are:

- mythical causality (de Kleer and Brown 1984, 1986);
- causal ordering (Iwasaki and Simon 1986, 1994);
- bond graphs (Top and Akkermans 1991; Dauphin-Tanguy et al. 2000).

Mythical causality can be seen as an intuitive method based on step by step propagation of the changes of direction variables will take in response to perturbation. For this purpose, it relies on heuristics. Causal ordering can be described as a computational method since causal chains coincide with paths of value computation of the variables, without reference to the underlying physical system. Finally, the bond graphs approach can be considered as methodological because it provides a unified modeling graphical language based on a categorization of physical phenomena.

The reader will find a detailed overview of causality as seen by the Qualitative Reasoning community in Dague and Travé-Massuyès (2004).

2.3 Evolutions and Trends

Qualitative Reasoning is a mature research field that fully contributes to the advancement of AI. Its influence in several related fields can no longer be denied, while the strengths and weaknesses of qualitative reasoning theories are perfectly identified. During the first ten years, Qualitative Reasoning produced ideas and novel theoretical results that are described in the "Readings" book (Weld and de Kleer 1989). Theoretical contributions as reported in Travé-Massuyès et al. (2003) then decreased. Let us notice however a contribution establishing the conditions under which an absolute order of magnitude model is consistent with a relative order of magnitude model in 2005 (Travé-Massuyès et al. 2005) and a qualitative information theory with the definition of the concept of entropy in absolute order of magnitude spaces in 2010 (Roselló et al. 2010). Less theoretical works resulted in many applications in the fields of engineering. Particular mention should be made of applications in the field of space autonomy (Williams and Nayak 1996; Muscettola et al. 1998), of diagnosis (Struss and Price 2003; Cascio et al. 1999; Travé-Massuyès and Milne 1997; Struss et al. 2014; Hofer et al. 2017), and of interactive learning (Bredeweg and Forbus 2003). An interactive learning environment based on qualitative simulation like DynaLearn allows for a novel form of active learning based on learning by modelling (Bredeweg et al. 2013).

The undeniable strength of qualitative simulation comes together with limitations that were quickly identified by the community. These are related to the prediction of spurious behaviors and has given rise to numerous works attempting to overcome the problem (Yilmaz and Say 2006). These limitations, however, correspond to the inherently incomplete nature of qualitative knowledge, a fact that the community now appears to have accepted (Kuipers 1985). Some relatively recent works about qualitative simulation define ad hoc algorithms tailored to specific systems. Let us notice in particular the applications to the simulation of genetic networks (de Jong et al. 2003; Ironi et al. 2008). However, Garp3 is at the forefront of nowadays qualitative simulators. Endowed with a diagrammatic visual language for representing qualitative models and a user friendly interface to inspect simulation results, Garp3 allows modellers to articulate and refine their conceptual domain knowledge and analyse this knowledge through simulation (Bredeweg et al. 2009).

The community has also faced the challenge of modeling, hence some researchers proposed to learn qualitative models automatically (Bratko and Suc 2003). In areas like ecology, biology, medecine or economics, challenges come from the fact that knowledge is not much formalized (Guerrin 1991; Ndiaye et al. 2009; Kansou and Bredeweg 2014). In technological fields that are typical engineering fields, qualitative models must get along with numerical models and the practices and know-how linked to them. Generating automatically a qualitative model from a numerical model is a critical issue that has been the focus of some work, particularly in the European project IDD (Picardi et al. 2002; Struss 2002), with no really satisfactory results. To move in this direction, the modeling practices should progress towards more modularity and the explicit modeling of validity assumptions of the models.

For a long time, most of the work of the Qualitative Reasoning community has been guided by technological applications (Iwasaki 1997; Travé-Massuyès and Milne 2009). This resulted in a significant overlapping of researchers attached to the Qualitative Reasoning community and the Model-Based Diagnosis community (see chapter "Diagnosis and Supervision: Model-based Approaches" of this volume) and attending both annual workshops, QR and DX respectively. After a joint organization of these workshops in 1998 and 1999, part of the community preferred to diversify co-organizations. A particular reason was a significant move towards cognitive science. This line of work is still active nowadays (Bredeweg and Struss 2003). For instance, Forbus (2014) returns to common sense reasoning in the spirit of the Naive Physics effort of Hayes (1979) with the goal to understand the mental models that support people's fluency in dealing with the physical world. On the other hand, Montserrat-Adell et al. (2016) rely on qualitative absolute orders of magnitude to build a set of hesitant fuzzy linguistic term sets that grasp the uncertainty existing in human reasoning. This kind of linguistic terms has been applied to a social multi-criteria evaluation real case study by Afsordegan et al. (2016). Reasoning and learning via analogy, human-machine interaction and in the field of education, interactive learning and cognitive diagnosis – diagnosis of the model proposed by the learner – are all active topics today (Forbus et al. 2002; Falkenhainer and Forbus Dedre 1989; de Koning et al. 2000; Bredeweg et al. 2013). Conceptual aspects of modeling, as a means to articulate and communicate knowledge, is also a very active topic.

The integration of qualitative analysis methods with traditional methods also defines an interesting line of work. In the fields of medicine and materials, the work by L. Ironi is a typical example that integrates qualitative methods with numerical methods and statistics to achieve automatic modeling, simulation and interpretation of experimental data (Ironi and Tentoni 2007). The analysis of systems with non-linear dynamics, possibly chaotic, also gives rise to methods integrating qualitative analysis and classical mathematical approaches (Ross et al. 2006).

3 Qualitative Spatial and Temporal Reasoning

3.1 An Overview of the Field

A simple but typical example of qualitative temporal reasoning is the following: assume we have to reason about point-like temporal entities, which we can model as points on the real line. We are only interested in *qualitative* relations between those entities, where qualitative means that we are not concerned with measuring the amount of time elapsed between two time-points. Actually, we will consider only three possible relations between two points: precedence (x precedes y), equality, and the relation that is the converse relation to precedence (x follows y).

Using those three binary relations, which we note $<$, $=$, and $>$, respectively, we define a formalism for representing and reasoning about temporal knowledge, the *Point calculus*, which has good computational properties in terms of the celebrated trade-off between expressiveness and computational complexity (Levesque and Brachman 1985). Its formulas are conjuncts of basic formulas of the form $\alpha(x, y)$, where x and y denote time-points, and α is either one of the three relations, called *basic relations*, or a disjunction of them, such as $\leq$ (the disjunction of $<$ and $=$), $\neq$ (the disjunction of $<$ and $>$), and so on. A further step consists in representing the formulas of the language using *constraint networks*, which are oriented graphs whose vertices are labelled by variables standing for points, and whose arcs are labeled by disjunctive relations.

In his 1983 seminal paper (Allen 1983), a paper dealing with reasoning about intervals, rather than points, Allen adopted the language of constraint networks. This allowed him to benefit from the link between temporal reasoning and the developing domain of constraint reasoning, based on the propagation of constraints. In particular, devising algorithms by adapting those used in the domain of *constraint satisfaction problems* (CSPs) became a natural strategy. In Allen's approach to temporal reasoning, called the *Interval calculus* or *Allen's interval calculus*, a pivotal role is played by the algebra formed by the set of disjunctive relations, called the *Interval algebra*, or *Allen's interval algebra*.

Making explicit the link with *constraint-based reasoning* and using that link to define reasoning "mechanisms" in temporal reasoning established the central *constraint-based* feature of the new domain. Nowadays, in its turn, QSTR as a body

of methods and techniques tends to emancipate itself from its limitations to space or time, and to give birth to a general domain of *qualitative constraint-based reasoning*.

In terms of research communities, the domain of *constraint-based* qualitative temporal or spatial reasoning has developed from the confluence of two research directions: for qualitative temporal reasoning, Allen's 1983 paper, whose formalism became the starting point for much of further research; and for qualitative spatial reasoning, the work of the Leeds school, whose definition of the RCC-8 formalism (Randell et al. 1992a) – also considered independently by Egenhofer (1991) under the name of the 9-intersection calculus –, a formalism for reasoning about relations between regions.

At this point, it may be in order to remark that the distinction between temporal and spatial reasoning in this context tends to be blurred or become inexistent if we consider the calculi themselves rather than what motivated them: for instance, the Point calculus can also be viewed as describing relations between points of the real line, that is, as a calculus on a 1-dimensional space. The reader is advised to keep in mind the important distinction between a formalism on the one hand, and its possible interpretations on the other hand.

On the international research activity level, a sustained activity in qualitative and spatial reasoning has taken place during the last decades. Workshops have been organized at major IA conferences such as IJCAI, AAAI, and ECAI, as well as specialized conferences such as the TIME and COSIT conferences. At the European level, the european SPACENET project (1994–1998) was a fruitful meeting point for the dissemination and promotion of the domain. At the French level, two French universities participated in the SPACENET project, and several government supported projects (GDRs) were carried out. Detailed presentations of the domain of qualitative spatial and temporal reasoning (QSTR) can be found in (Renz and Nebel 2007), (Ligozat 2013), (Chen et al. 2015).

Representing and reasoning about time was one of the motivations of Prior for developing time logics (Prior 1957) as particular kinds of modal logics. The spatial interpretation of some modal logics had been introduced in the 1940s by McKinsey and Tarsky (1944). The logical approaches have been widely developed, with a constant interaction with the constraint-based approaches (Le Ber et al. 2007).

Finally, to get a wider perspective on the topic, one has to mention various approaches using graph theoretical tools, the notion of entropy, lattices, Markov models, temporized networks, homological algebra, mathematical morphology, possibility theory and the qualitative study of shapes.

In what follows, we will successively consider constraint-based formalisms, the main problems encountered, the perspectives offered by this type of research, some alternative approaches and conclude with a brief overview of applications.

3.2 *Qualitative Calculi*

3.2.1 Allen's Interval Calculus

Allen's interval calculus considers intervals as primitive temporal entities, which can be interpreted as (closed, bounded) intervals on the real line, that is, as pairs of distinct real numbers. The qualitative relations considered between pairs of intervals are those relations that correspond to all possible orderings of the end-points. This yields a set of 13 relations (Fig. 3).

The set of basic relations has a natural structure of a lattice, which it inherits from the ordering of time points. Moreover, because intervals are pairs of points, the basic relations can be represented as regions in the plane (Fig. 4).

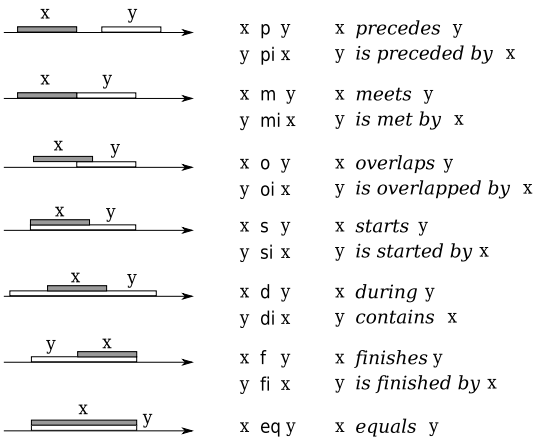


Fig. 3 The basic relations of Allen's interval calculus

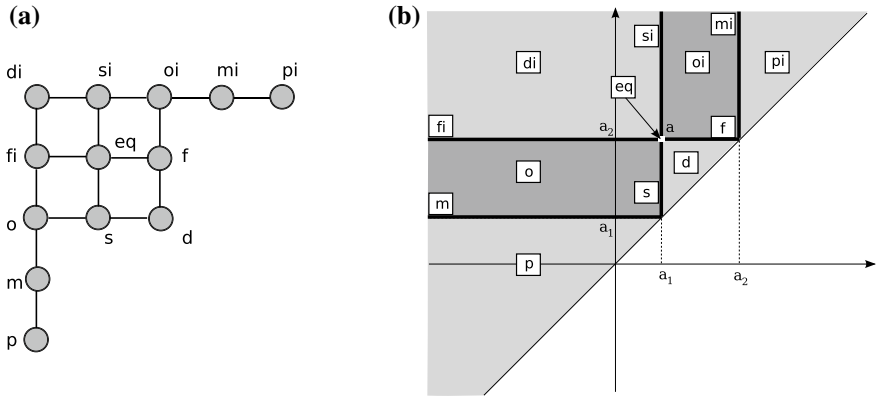


Fig. 4 Basic relations of Allen's interval calculus as a lattice **a** and as regions **b**

This set of relations has the JEPD (*jointly exhaustive and pairwise disjoint*) property, that is, it constitutes a partition of the set of all pairs of intervals: any pair of intervals belongs to one, and only one, of them. Those relations are called the *basic relations* of Allen's interval calculus.

The "formulas" of the language are conceptualized in terms of constraint networks, which are oriented graphs whose arcs are labeled by sets of basic relations interpreted as disjunctions. The nodes of the networks correspond to intervals, and the labelings to constraints between them.

Reasoning uses constraint propagation based on the existence of two operations on the (disjunctive) relations: the operation of *conversion*, that sends relation p (*precedes*) to relation pi (*is preceded by*), and similarly for other basic relations, except equality which is its own converse, and the operation of *composition* of two relations that can be described by a *composition table*; using constraint propagation, one can compute the *algebraic closure* of a given network, by repeatedly executing

$$C(i, j) \leftarrow C(i, j) \cap (C(i, k) \circ C(k, j))$$

for all triples of nodes (i, j, k) until the network is no longer changed. A network such that $C(i, j) \subseteq C(i, k) \circ C(k, j)$ for all triples (i, j, k) of nodes is called *algebraically closed* (or *path-consistent*).

3.2.2 A Review of Some Qualitative Calculi

Many qualitative calculi have been defined and studied during the past three decades. We present a quick (not exhaustive) review of some of them.

The *generalized interval calculus* (Ligozat 1991) considers temporal entities which are finite sequences of points on a line. It generalizes the Point calculus (where the sequence is reduced to one point) and Allen's interval calculus (sequences of two points). Many properties of Allen's interval calculus can be shown to be still valid for sequences of length greater than two.

The *Cardinal direction calculus* (Ligozat 1991) is basically the product of two instances of the Point calculus. Analogously, the *Rectangle calculus*, introduced by Güsgen (1989), is the product of two instances of Allen's interval calculus. This calculus, as well as its generalizations to higher dimensions, has been studied by Balbiani et al. (1998).

The Cardinal direction calculus, whose entities are points in a 2D plane, has been extended to regions in the plane (Goyal and Egenhofer 1997).

The *RCC-8* calculus has been introduced by the Leeds school (Randell et al. 1992b) as a sublanguage of the *RCC* theory, and independently, defined by Egenhofer

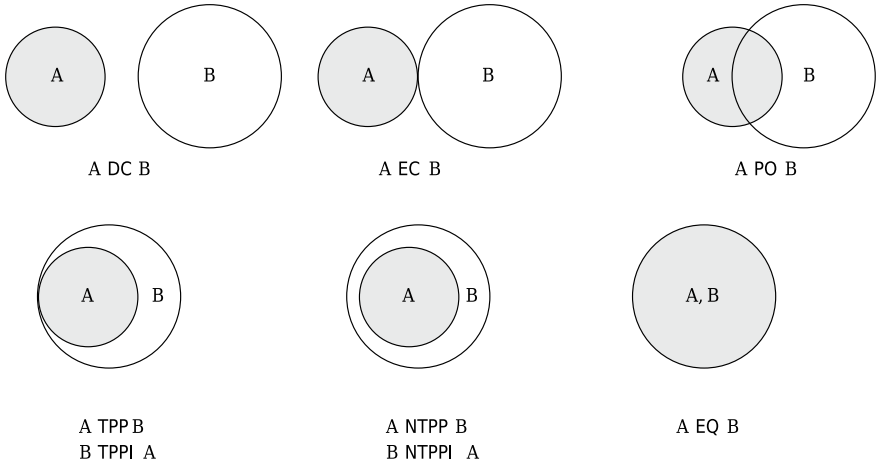


Fig. 5 The basic relations of the *RCC-8* calculus

(1989) as the *9-intersection calculus*. Its eight basic relations can be defined in a simple way for regions which are closed disks in the plane, as shown in Fig. 5.[2]

The *Cyclic interval calculus* considers entities which are arcs on an oriented circle, defined by their starting and end-points. This calculus, which is analogous to Allen's interval calculus on a circle, has been defined and studied by Balbiani and Osmani (1999, 2000).

The INDU calculus (Pujari et al. 1999) refines Allen's interval calculus by taking into account the relative durations of the intervals considered. For instance, the relation of precedence p splits into three sub-relations $\mathsf{p}^{<}$, $\mathsf{p}^{=}$, and $\mathsf{p}^{>}$ according to the fact that the first interval is shorter, has the same duration, or is longer than the second.

All calculi mentioned up to now have considered binary relations. In a plane which has no global orientation, ternary relations have to be used. The best known ternary calculus about points in a plane is Freksa's *Double-cross calculus*[3] (Freksa 1992). Ligozat (1993) has shown that this calculus is a member of a family of calculi called *qualitative triangulation calculi*, the simplest of which is the *Flip-flop calculus*.

If the entities considered are regions in the plane, ternary relations of alignment can be defined, yielding a *5-intersection calculus* (Billen and Clementini 2004).

[2] DC stands for *disconnected*, EC for *externally connected*, PO for *partial overlap*, TPP for *tangential proper part* and NTPP for *non-tangential proper part*; TPPI and NTPPI are the converses of TPP and NTPP, respectively.

[3] Those calculi divide all directions in the plane with respect to a given point of reference into a finite number of sectors with a given angle; Freksa's calculus is the case where the angles are right angles, the Flip-flop calculus where they are 180° angles.

3.3 Main Problems and Results

3.3.1 The Consistency Problem

The *consistency problem* is a central problem. It consists in answering the following question: given a (finite) constraint network, determine whether there is a finite configuration of the entities such that the constraints are satisfied. For the Point calculus, this problem can be solved in polynomial time (see chapter "Theoretical Computer Science: Computational Complexity" of volume 3), for instance by applying an algorithm of van Beek (1990). Ghallab and Alaoui (1989) give efficient techniques for solving large networks (several thousands of points). For Allen's interval calculus, the consistency problem is NP-complete (Vilain et al. 1989): the property of algebraic closure is a necessary, but not sufficient condition for consistency, as already shown in Allen's 1983 paper.

This so-called incompleteness of the algebraic closure property is true for most of the calculi we mentioned. In view of this fact, it is of interest to characterize subsets of the relations for which the problem is polynomial – the problem is then said to be *tractable* – and, when such is the case, to define suitable algorithms for deciding consistency. In particular, the question arises: in what cases does the algebraic closure property (which can be enforced in cubic time) imply consistency?

In order to characterize sub-classes of relations (subsets which are stable under intersection, converse and composition), two approaches have been developed in the literature: a *syntactic approach* (Nebel and Bürckert 1995; Koubarakis 1996, 2001; Jonsson and Bäckström 1998), and a *geometric approach* introduced by Ligozat (1994, 1996) and developed at Orsay, Toulouse and Villetaneuse (France) by Balbiani, Fariñas del Cerro, Condotta, Osmani and their students.

A central result for Allen's interval algebra is the fact that there exists a single maximal subclass of relations containing all basic relations such that the consistency problem is solvable in polynomial time. In syntactical terms, this subclass is that of *ORD-Horn relations* (Nebel and Bürckert 1995). In geometrical terms, those relations constitute the subclass of *pre-convex relations*, which can be characterized in a simple way: in the lattice representation of the set of basic relations, they are those relations corresponding to intervals of the lattice, or to intervals of the lattice from which some 1-dimensional or 0-dimensional relations have been removed. In terms of regions in the plane, they are regions satisfying some convexity conditions.

As for the RCC-8 calculus, any atomic algebraically closed network is consistent, and there exist exactly three maximal polynomial subclasses containing all atomic relations (Renz 1999).

For many qualitative spatial or temporal calculi in the literature, the consistency problem does not have such a simple solution as for Allen's interval calculus: even algebraically closed atomic networks may not be consistent; consistency may not imply global consistency (some partial solutions may not be extensible to global solutions). An important breakthrough on the syntactic front has been the characterization of *disjunctive linear relations* (DLRs), independently discovered by Koubarakis

(1996, 2001) and Jonsson and Bäckström (1998). The joint application of syntactic and geometrical methods to the INDU calculus leads to the characterization of several polynomial subclasses in the INDU algebra (Balbiani et al. 2006).

Beyond the consistency problem, other problems (Long and Li 2015; Sioutis et al. 2015b) concerning constraint networks have been considered in the litterature. In particular, much attention has been focused on the minimal labeling problem, which is the problem of determining all the basic relations of a constraint network that participate in at least one of its consistent configurations (Liu and Li 2012; Amaneddine et al. 2013).

3.3.2 Models for the Qualitative Calculi

In several cases, the relations of a qualitative spatial or temporal calculus constitute a *relation algebra* in the sense of Tarski (1941). Ligozat (1990) introduced the notion of *weak representation* of such an algebra. Intuitively, a weak representation is a weak model of the theory corresponding to the algebra, in the sense that the axioms describing the operation of composition are interpreted as implications rather than equivalences. This notion generalizes the classical notion of representation in universal algebra.

The weak representations of the algebras of the calculi based on total orderings — this covers the Point calculus, Allen's interval calculus, the Cardinal direction calculus, the generalized interval calculi, the Rectangle calculus as well as more generally the n-point and n-block calculi — have been studied by Ligozat (1991, 2001). This allowed him to show that all those calculi have the $\aleph_0$-property, that is, that up to isomorphism the corresponding theories have a unique countable model.

The models of the RCC theory, of which RCC-8 is a fragment, correspond to structures called *contact Boolean algebras* (Stell 2000). Using this characterization, all models of the calculus have been classified by Li and Ying (2003), who also defined more general calculi which have both continuous and discrete models.

3.3.3 Qualitative Constraint Calculi

What exactly is a qualitative (constraint-based) calculus? To answer this question, Ligozat and Renz (2004) related the notion of weak representation to a semantical notion of *partition scheme*: a *qualitative calculus* is defined as a non-associative algebra together with a weak representation of it satisfying some minimal conditions. The consistency of a network then means the existence, in the category of all weak representations of the algebra, of a morphism from the object corresponding to the network to the weak representation defining the calculus.

Mossakowski et al. (2006) argued that only so-called *semi-strong* weak representations, which are necessarily injective, should be considered in order to define a qualitative calculus. The notion of "qualitative (constraint) calculus" has been fur-

ther elaborated by Dylla et al. (2013), who defined *abstract partition schemes* and by Westphal et al. (2014) (Westphal 2014).

Inants (2016) has extended the framework based on weak representations and partition schemes to accomodate *heterogeneous universes* (where the entities considered are of several sorts), hence the identity is not necessarily atomic,[4] by defining *modular partition schemes*.

3.3.4 Solving the Consistency Problem

In order to solve the consistency problem for a network whose constraints are finite disjunctions of basic relations, a natural approach consists in successively enumerating all its basic subnetworks,[5] called *scenarios*. The number of scenarios obtained in this way can be reduced by enforcing the algebraic closure condition once a basic relation has been selected.

In Nebel (1996), Nebel proposes a very efficient algorithm that can be used once a class S of tractable relations is known for which the algebraic closure method is complete: the constraints of the network are decomposed into elements of S. In the case of Allen's interval relations, using the class of ORD-Horn relations results in reducing the branching factor from 13 to 5 approximately. All currently efficient algorithms are based on Nebel's approach.

All those methods can be refined using heuristics: on the one hand, for choosing which relation should be treated first, and on the other hand for choosing the component sub-relation of the current relation. A host of heuristics have been proposed — and experimentally evaluated — in the literature (van Beek and Manchak 1996).

Mainly based on the algorithms mentioned previously, several systems have been developed which implement generic tools for solving the consistency problem for networks based on the various calculi proposed in the literature (Condotta et al. 2006b; Wallgrün et al. 2006a).

Periodical Constraints

Some applications such as calendar management involve constraints that are applied recurrently over periods of time. Part of the activity research on qualitative reasoning has been devoted to representing and reasoning about this type of constraints. We have already mentioned the Cyclic interval calculus, which has 16 basic relations. Incidently, those relations have also been axiomatized in first-order predicate logic (Condotta and Ligozat 2004).

Another approach of periodicity consists in considering qualitative constraints such as those on Allen's interval algebra as constraints of their own right on indefinitely recurring periods of time. Then, a solution of the consistency problem is a

[4] This was already the case for several calculi in the literature, such as Vilain's Point-and-Interval calculus (Vilain 1982), and more generally Ligozat's $\mathbf{A}_S$ calculi, where S is a set of positive integers with more than one element (Ligozat 1991).

[5] i.e. Subnetworks for which all constraints are *basic* relations.

valuation of the variables at each instant such that for each time period the qualitative constraints are satisfied. The consistency problem of such constraint networks, for various qualitative calculi appearing in the literature, has been studied by Condotta et al. (2005).

Representations concerned with activities or events recurring a finite number of times have also been considered by Khatib (1994). The calculi proposed there allow one to explicitly specify what qualitative constraints should be satisfied between instances of recurring temporal activities. Such constraint problems can be solved using classical qualitative constraint networks, where each variable stands for an instance of an activity.

3.4 Perspectives

Among the perspectives, two active directions of research deserve to be mentioned: one is *extending and combining calculi*, and the other is *building bridges* from the domain of QSTR to other domains.

3.4.1 Extending and Combining Calculi

Extending an existing calculus to other kinds of entities is one type of extension: a case in point is the extension of the Cardinal direction calculus, a point-based calculus, to extended regions in the plane (Goyal and Egenhofer 1997).

Spatio-temporal calculi, where the intended interpretations are entities of a spatio-temporal nature, are particular cases of extensions: they include qualitative trajectory calculi (van de Weghe 2004), spatio-temporal formalisms (Muller 1998), and combinations of a spatial calculus with a temporal calculus (Gerevini and Nebel 2002).

Granularity has been a constant issue in QSTR (Hobbs 1985; Bettini et al. 2002). It is another context where the extension of existing calculi is involved. Typically, *zooming in* or *zooming out* will change the relations: for instance, two intervals separated by a short distance (relation p) will appear to meet (relation m) when zooming out, while conversely two meeting intervals could prove to be in a preceding relation when zooming in Euzenat (1996, 2001). More dramatically, a (very short) interval can be considered as a point when zooming out sufficiently (Bettini et al. 2000; Euzenat and Montanari 2005).

Cohen-Solal et al. (2015, 2017a) propose a general formalism for qualitative temporal reasoning with granularities, and show how algebraic closure can be used in this broader context to obtain new results of tractability and minimality. The problem of describing admissible sequences of configurations of spatial objects is also revisited (Cohen-Solal et al. 2017b).

Combining existing calculi describing the same entities is another possibility: the very same entities are considered from various perspectives, using distinct sets of relations. The combination can be *loose* or *tight*: in the former case, two calculi are

used independently, and procedures for exchanging information are defined; in the latter, two calculi merge into a new one (Westphal and Wöfl 2008).

Two examples of loose combinations are the combination of the *RCC-8*-calculus with the Rectangle calculus, and that of the *RCC-8*-calculus with the Cardinal direction calculus between regions (Liu et al. 2009; Cohn et al. 2014). By contrast, the INDU calculus is a typical example of tight combination.

In the case of loose combination, a general constraint propagation method, called *bi-path consistency*, has been proposed by Gerevini and Renz (2002).

Since a qualitative calculus about spatial or temporal entities corresponds to a partition of the set of pairs of elements of the corresponding universe, the set of these qualitative calculi is naturally structured as a lattice. This fact is used by Condotta et al. (2009) to study combinations of calculi.

3.4.2 Building Bridges to Other Domains

Building bridges between different domains can be beneficial in a one-sided way: translating a problem in one domain in terms of another domain permits to use methods in the latter to solve problems in the former. Or conversely. But it can also be mutually beneficial to both domains.

The first situation arises when QSTR problems are translated in terms of SAT problems, or in terms of finite CSPs (constraint satisfaction problems), allowing to use powerful methods in either the SAT or the CSP domain.

QSTR and the SAT Problem

A systematic approach to a translation of the consistency problem of a qualitative constraint network into a SAT problem consists in abstracting the semantics of the different basic relations, and then considering the QSTR problem as a combinatorial problem. Each basic relation of each qualitative constraint is represented by a propositional variable, hence to each qualitative constraint corresponds an exclusive disjunction. Another set of clauses defined from the composition table models all the infeasible qualitative configurations of three variables. Hence, all truth assignments satisfying the set of SAT clauses obtained in that way will correspond to algebraically closed scenarios.

As a consequence of this fact, such a SAT based translation is not necessary complete, unless it is used in the context of a qualitative formalism for which all algebraically closed scenarios are consistent scenarios. It should also be mentioned that the propositional clauses used in this approach are not necessarily Horn clauses. Nevertheless, this approach makes it possible to use SAT solvers (see chapter "Reasoning with Propositional Logic: from SAT Solvers to Knowledge Compilation" of volume 2) to efficiently solve the consistency problem of a qualitative constraint network. The main difficulty when using a SAT translation of the consistency problem of a qualitative constraint network arises from the number of SAT clauses thus obtained, which can be very large. Current research has devised methods for reducing

this number of clauses (Li et al. 2009b; Condotta and D'Almeida 2011; Condotta et al. 2016).

QSTR and Discrete CSPs

Another direction of research has focused on translating the consistency problem of qualitative constraint networks into discrete CSPs (see chapter "Constraint Reasoning" of volume 2). In this context, a variable V_{ij} of the CSP is associated to each constraint $C_{(i,j)}$ of the qualitative constraint network. The domain of the variable V_{ij} is defined by the set of basic relations defining the qualitative constraint $C(i, j)$. The information corresponding to the composition table of the formalism is modeled in the discrete constraint satisfaction problem by introducing a ternary constraint C_{ijk} whose scope is (V_{ij}, V_{jk}, V_{ik}) for all triples (i, j, k).

Similarly to the case of the translation into a SAT problem, the translation of QSTR problems into discrete CSPs is not necessary complete, unless it is used in the context of a qualitative formalism for which all algebraically closed scenarios are consistent. Furthermore, the translation can lead to discrete CSPs with huge sizes due to the encoding of the composition table.

Recent papers report results of experimental and theoretical comparisons between those different approaches, see for example (Westphal and Wölfl 2009).

QSTR and Ontological Reasoning

The domain of QSTR has also been related to other types of reasoning. A case in point is *ontological reasoning*, which illustrates a situation of mutual benefit: motivated by problems in ontological reasoning, and using Euzenat's connection of that domain to QSTR (Euzenat 2008), Inants (2016) extends the framework of QSTR in order to accomodate many-sorted universes — a benefit to QSTR — and uses it to overcome difficulties in the domain of qualitative ontological reasoning.

3.5 Alternative Approaches

Modal Logics and QSTR

While temporal logics have been developed over several decades, acquiring an undeniable level of maturity (Prior 1967; Bestougeff and Ligozat 1992), it has only been recently that the full potential of spatial logics has been re-evaluated (Aiello et al. 2007b). In this renewed surge of interest for the modal study of space, the Amsterdam school has played a leading part (Aiello et al. 2007a).

The linking point between the domain of modal logics and QSTR arises from the possibility to translate qualitative languages such as RCC-8 in terms of modal formulas in such a way that satisfiability is preserved.

Spatio-Temporal Logics

In order to represent situations involving simultaneously time and space, several spatio-temporal logics have been proposed (Wolter and Zakharyaschev 2000). Those

logics, based on the temporal logic LTL, allow one to reason about changes of relative positions of spatial entities through time. The relative positions of the entities are expressed using spatial variables that are related using basic relations of a qualitative calculus. For instance, using the language of RCC-8, the formula $\mathbf{F}(a\ \mathsf{NTPP}\ b)$ will express the fact that, in the future, the spatial region represented by a will be non-tangentially included into that represented by b. Temporal operators that apply to spatial variables are also considered. For instance, $\mathbf{X}a$ will represent region a at the next instant. For more detailed information on the topic of spatio-temporal logics, the reader is invited to consult the references in (Balbiani and Condotta 2002), (Sioutis et al. 2015a).

Lattices of Relations

In the specific domain of the qualitative representation of space and time, the interest of lattices is twofold: firstly, they are natural models to represent temporal or spatial algebras; secondly, they allow one to fill the gap between numerical geographical information and qualitative spatial relations.

Temporal or spatial algebras, equipped with set inclusion, allow one to generate Boolean lattices whose structure can be used for reasoning. Furthermore, for all formalisms based on total orders, basic relations possess a lattice structure that can be used to define convex and pre-convex relations.

Galois lattices – or *concept lattices* (Ganter and Wille 1999) – are specific lattice structures that are particularly useful in spatial reasoning. These structures have been used to relate spatial relations, such as those of *RCC-8*, to the output of set-theoretical operations on spatial regions (vector or raster regions). Such a lattice is described in (Napoli and Le Ber 2007) (Fig. 6). The extension R of each concept represents a disjunction of *RCC-8* basic relations. The intension C represents a conjunction of results obtained from set operations (or tests) on spatial regions. The equivalence $\bigvee_{r \in R} r(x, y) \leftrightarrow \bigwedge_{c \in C} c(x, y)$ allows one to compute the existing relations between two spatial regions x and y. The resulting Galois lattice is also useful for spatial inference (conjunction and composition of relations), but to a lesser extent than the Boolean lattice that contains it.

Finally, lattices can be used to represent geographical information itself. Indeed, they allow one to easily manipulate region decomposition and recomposition. For example, in a geographical information system, regions are often decomposed into triangles, lines and points. These spatial elements can be organised within a lattice based on a set of rules as: "two segments s_1 and s_2 share only a point denoted by $s_1 \frown s_2$". Topological relations between regions can then easily be deduced from the lattice structure. For example, the intersection of two regions is obtained as an infimum: if this infimum is a triangle then the regions overlap, if it is a segment or a point they are connected, if it is the universal minimum they are disconnected. The lattice representation also allows one to easily compute other topological notions such as neighborhood or border.

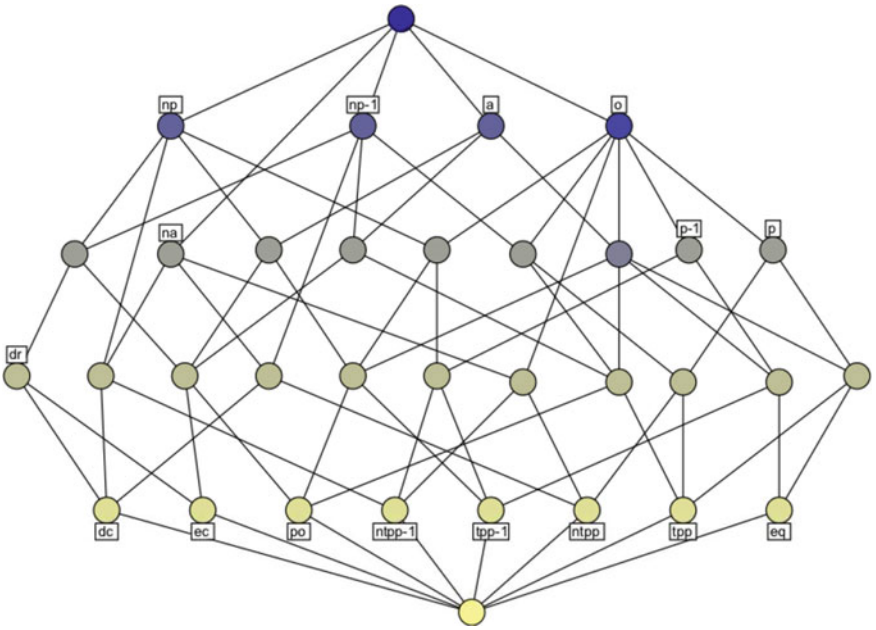


Fig. 6 The Galois lattice used in (Napoli and Le Ber 2007): each concept is defined by the properties inherited top-down (for intension) and the relations inherited bottom-up (for extension)

3.6 Applications of Qualitative Spatial and Temporal Reasoning

Three types of applications can be distinguished.

1. Transpositions to germane domains, to linguistics, or image-processing; in return these domains can influence the qualitative models that are developed in the Artificial Intelligence domain.
2. Applications to various domains such as landscape management, archaeology, etc., where qualitative models are used to formalise expert knowledge and are often related to numerical information.
3. Software tools implementing various models.

Transposition to Germane Domains

Relationships between Artificial Intelligence and linguistics are time-honored and mutually fruitful, and qualitative reasoning models of time and space are often based on research in linguistics or more generally in cognitive science. In France, these relationships have been mainly explored in Toulouse and Orsay, since the early 1990s (Bestougeff and Ligozat 1992; Vieu 1991); Muller (1998) studies the combination of spatial and temporal aspects, in order to build a model describing basic movements. Studying natural language helps to refine the notions of spatial objects and relations:

in (Aurnague et al. 1997), the authors focus on how natural language defines the location of objects and manages imprecise information: they propose a formal definition for the part-whole relation, the orientation relations, and the French spatial preposition *sur*,[6] etc. This kind of work participates in a global research domain that is also active elsewhere in Europe and in the United States (see for example, Lascarides and Asher 1991, 1993; Mark et al. 1995).

Conversely, progress in spatial and temporal reasoning models allows one to automatically perform text analysis, spatial information extraction, event detection or map production. Loustau et al. (2008) have developed a tool for the extraction of spatial information from texts in order to help the analysis of corpora of ancient travel stories, and to automatically describe routes; this tool is used to extract *named locations* and geographical concepts, the spatial relations existing between locations, as well as some syntactico-semantic relations. A geometric representation of the corresponding spatial pattern is finally generated using a geographic information system. The work presented in (Ligozat et al. 2007) focuses on the automatic graphical representation of spatio-temporal events, such as movements of troops on a battlefield, based on informations extracted from natural language texts. A typology of elementary scenes is used to this effect (Przytula-Machrouh et al. 2004). The graphical representation uses "choremes", an iconic representation developed in geography. Similar approaches are also applied in the domain of security, e.g. to detect or anticipate potentially dangerous events from textual data such as short messages or breaking news: crowd management (Ligozat et al. 2011), detection of epidemiological phenomena (Chaudet 2006), or, in a hostile country, the detection of dangerous situations (Li et al. 2009a).

Connections with the geographical information and image-processing domains also provide mutual inspiration. Those domains mainly use numerical data – vector or raster data – and the formal definition of spatial relations has to be based on set-theoretical operations: based on seminal work by Egenhofer (1989), these definitions have been adapted in France for the recognition of spatial objects and structures on satellite images (Le Ber and Napoli 2003). Poupeau and Bonin (2006) have extended this approach to 3D data: they combine geometrical and topological models to compute spatial relations between blocks, as for example the "lay on" relation. Furthermore, to bridge the semantic gap between qualitative knowledge and numerical data, various approaches have been devised, including fuzzy approaches: e.g., a fuzzy modeling of the "between" relation and of orientation relations has been proposed to analyze medical images (Bloch 1999). A review of fuzzy approaches in this context is given in (Bloch 2005). In (Atif et al. 2007), based on this type of characterisation, spatial reasoning is performed in order to detect pathological cases in brain images. Another approach consists in refining existing models, for example by specifying the EC, PO relations with different border intersection cases (points, lines, "thick" borders) (Alboody et al. 2010); this can be related to the notion of "contact" studied in qualitative spatial reasoning.

[6]This preposition roughly corresponds to the English preposition *on*.

Also pertaining to the image processing domain is Cotteret's work (Cotteret 2005) on the extraction of curvilinear elements (roads, watercourses), which mimicks the activity of a human cartographer-analyst by focusing on specific zones. Local bits of information thus obtained are then merged, based on qualitative models of proximity and orientation, allowing the reconstruction of the global space.

In (Yang et al. 2015) a framework using a tableau method is described for generating and selecting potential explanations of the given image when the background knowledge is encoded using a description that is able to handle spatial relations.

Applications to Other Domains

Qualitative models of time and space, and more generally qualitative reasoning models have been applied in various domains distinct from Artificial Intelligence and cognitive science. Some industrial or medical applications can be mentioned. Qualitative models of time are especially used for default diagnosis, for example based on constraint networks (Osmani and Lévy 2000).

Environmental assessment, urbanism or territory management, historical sciences, are obviously application domains, since they contain numerous problems relying on qualitative and weakly formalised expertise. In archaeology, e.g., temporal qualitative models have been used for document annotation, in order to automatically compare and merge several datings based on a constraint propagation mechanism (Accary-Barbier and Calabretto 2008). Another work pertaining to deep-sea archaeology, has developed a preliminary representation of the observations and knowledge about ancient ships; this representation is implemented within an ontology including qualitative spatial and temporal relations (Jeansoulin and Papini 2007).

Qualitative models of time and space have also been used to facilitate user interaction (for such users as hydrologists or ecologists) in several computer systems and various domains such as environment, prevention of natural hazards, or management of natural species; these systems usually use several geographic information sources, where the information is often numerical and has to be translated into qualitative terms (Bedel et al. 2008). Models of belief revision have been used in order to merge spatial information sources about floods (Würbel et al. 2000; Ben-Naim et al. 2004; Benferhat et al. 2010). In a closely related domain, a project has been undertaken by the French INRA[7] for modeling knowledge and reasoning about the observation and diagnosis of farming territories: the aim is to provide an aid based on an automatic monitoring to agronomists analyzing territories submitted to various constraints and aggressions (urbanisation, deforestation, agricultural pollution). In (Le Ber et al. 2003), spatial qualitative models are used to describe, compare, and classify agricultural structures at the level of the farm or the rural community. These models are also used for military terrain recognition and characterisation based on typical features (Chevriaux et al. 2005). In (de Beuvron et al. 2015) ontological and qualitative spatial reasoning are combined in order to interpret urban images.

The various applications mentioned above implement spatial and temporal models using knowledge representation languages, or object-based or logical languages

[7] French National Institute for Agronomic Research.

(Le Ber et al. 2003; Bedel et al. 2008). Currently, many approaches use description logics.

Software Tools

Implementing models of qualitative reasoning often relies on ad hoc approaches. Nevertheless, as already previously mentioned, some generic tools have been proposed to the community of researchers. The algebraic description of temporal and spatial relations allowed to develop generic tools such as QAT (Condotta et al. 2006a), in France, and SparQ (Wallgrün et al. 2006b) and GQR in Germany (Gantner et al. 2008). These tools can be used to solve constraint problems, for a given formalism specified by its composition table.

Other generic tools have been developed in the domain of knowledge representation.[8] These tools allow to perform inferences (generalisation, specialisation, composition of relations). In France, work undertaken with the object-based knowledge representation tool AROM also involved space and time representation (Miron et al. 2007).

4 Conclusion

This chapter described what we believe lies at the core of qualitative approaches for reasoning on dynamical systems on the one hand, and on space and time, on the other hand. Obviously, many connections exist between qualitative reasoning and several topics of Artificial Intelligence considered in this book. This is true in particular for modal and non-monotonic logics (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" of this volume), on techniques for solving the SAT problem (see chapter "Reasoning with Propositional Logic: from SAT Solvers to Knowledge Compilation" of volume 2), on constraint based reasoning (see chapter "Constraint Reasoning" of volume 2), on natural language processing (see chapter "Artificial Intelligence and Natural Language" of volume 3), on pattern recognition and vision (see chapter "Artificial Intelligence and Pattern Recognition, Vision, Learning" of volume 3) and on robotics (see chapter "Robotics and Artificial Intelligence" of volume 3).

References

Accary-Barbier T., Calabretto S. (2008) Building and using temporal knowledge in archaeological documentation. J. Intell. Inf. Syst. 31:147–159

Afsordegan A., Sánchez M., Agell N., Aguado J. C., Gamboa G. (2016) Absolute order-of-magnitude reasoning applied to a social multi-criteria evaluation framework. J. Exp. Theor. Artif. Intell. 28(1–2):261–274

[8] See for example the RACER tool: http://www.racer-systems.com/.

Aiello M., Pratt-Hartmann I., van Benthem J. (eds.) (2007a) Handbook of spatial logics. Springer, Netherlands

Aiello M., Pratt-Hartmann I., van Benthem J. (2007b)What is spatial logic? In [Aiello et al. 2007a], pp 1–11

Alboody A., Sedes F., Inglada J. (2010) Fuzzy intersection and difference model for topological relations. In: IFSA-EUSFLAT 2009 Proceedings, pp 1–6

Allen J. F. (1983) Maintaining knowledge about temporal intervals. Commun. ACM 26(11):832–843

Amaneddine N., Condotta J.-F., Sioutis M. (2013) Efficient approach to solve the minimal labeling problem of temporal and spatial qualitative constraints. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13), Beijing, China, 3–9 August 2013, pp 696–702

Atif J., Hudelot C., Fouquier G., Bloch I., Angelini E. (2007) From generic knowledge to specific reasoning for medical image interpretation using graph-based representations. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07), pp 224–229

Aurnague M., Vieu L., Borillo A. (1997) Représentation formelle des concepts spatiaux dans la langue. In: Denis M (ed) Langage et cognition spatiale. Masson, pp 69–102

Balbiani P., Condotta J.-F. (2002) Computational complexity of propositional linear temporal logics based on qualitative spatial or temporal reasoning. In: Proceedings of the 4th international workshop on frontiers of combining systems (FroCoS 2002). LNCS, vol 2309, pp 162–176

Balbiani P., Condotta J.-F., Fariñas del Cerro L. (1998) A model for reasoning about bidimensional temporal relations. In: Proceedings of KR-98, pp 124–130

Balbiani P., Condotta J-F., Ligozat G. (2006) On the consistency problem for the INDU calculus. J. Appl. Log. 4:119–140

Balbiani P., Osmani A. (2000) A model for reasoning about topological relations between cyclic intervals. In: Proceedings of KR-2000, Breckenridge, Colorado, pp 378–385

Barkowsky T., Knauff M., Ligozat G., Montello D. R. (eds.) (2008) Spatial cognition V: Reasoning, Action, Interaction. International Conference on Spatial Cognition 2006, Bremen, Germany, 24–28 September 2006, revised selected papers. Lecture notes in computer science, vol 4387. Springer, Berlin

Bedel O., Ferré S., Ridoux O., Quesseveur E. (2008) GEOLIS: a logical information system for geographical data. Revue Internationale de Géomatique 17(3–4):371–390

Ben-Naim J., Benferhat S., Papini O., Würbel E. (2004) An answer set programming encoding of prioritized removed sets revision: application to GIS. In: Alferes JJ, Leite JA (eds) JELIA, vol 3229. Lecture notes in computer science. Springer, Berlin, pp 604–616

Benferhat S., Ben-Naim J., Papini O., Würbel E. (2010) An answer set programming encoding of prioritized removed sets revision: application to GIS. Appl. Intell. 32(1):60–87

Bestougeff H., Ligozat G. (1992) Logical tools for temporal knowledge representation. Ellis Horwood, New York

Bettini C., Jajodia S., Wang S. X. (2000) Time granularities in databases, data mining and temporal reasoning. Springer, Berlin

Bettini C., Wang X. S., Jajodia S. (2002) Solving multi-granularity temporal constraint networks. Artif. Intell. 140:107–152

Billen R., Clementini E. (2004) A model for ternary projective relations between regions. In: Bertino E., Christodoulakis S., Plexousakis D., Christophides V., Koubarakis M., Böhm K., Ferrari E. (eds) EDBT, vol 2992. Lecture notes in computer science. Springer, Berlin, pp 310–328

Bloch I. (1999) Fuzzy relative position between objects in image processing: a morphological approach. IEEE Trans. Pattern Anal. Mach. 21(7):657–664

Bloch I. (2005) Fuzzy spatial relationships for image processing and interpretation: a review. Image Vis. Comput. 23(2):89–110

Boutilier C. (ed.) (2009) IJCAI 2009 proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, 11–17 July 2009

Bratko I., Suc D. (2003) Learning qualitative models. AI Mag. 24(4):107–119

Bredeweg B., Forbus K. (2003) Qualitative modeling in education. AI Mag. 24(4):35–46
Bredeweg B., Liem J., Beek W., Linnebank F., Gracia J., Lozano E., Wißner M., Bühling R., Salles P., Noble R et al. (2013) Dynalearn - an intelligent learning environment for learning conceptual knowledge. AI Mag. 34(4):46–65
Bredeweg B., Linnebank F., Bouwer A., Liem J. (2009) Garp3 workbench for qualitative modelling and simulation. Ecol. Inform. 4(5–6):263–281
Bredeweg B., Struss P. (2003) Current topics in qualitative reasoning. AI Mag. 24(4):13–16
Cascio F., Console L., Guagliumi M., Osella M., Panati A., Sottano S., Dupré D. (1999) Generating on-board diagnostics of dynamic automotive systems based on qualitative models [1]. AI Commun. 12(1–2):43–51
Chaudet H. (2006) Extending the event calculus for tracking epidemic spread. Artif. Intell. Med. 38(2):137–156. Special issue on Temporal Representation and Reasoning in medicine
Chen J., Cohn A. G., Liu D., Wang S., Ouyang J., Yu Q. (2015) A survey of qualitative spatial representations. Knowl. Eng. Rev. 30(1):106–136
Chevriaux Y., Saux E., Claramunt C. (2005) A landform-based approach for the representation of terrain silhouettes. In: Shahabi C., Boucelma O. (eds.) GIS. ACM, pp 260–266
Cohen-Solal Q., Bouzid M., Niveau A. (2015) An algebra of granular temporal relations for qualitative reasoning. In: Twenty-fourth International Joint Conference on Artificial Intelligence, IJCAI 2015
Cohen-Solal Q., Bouzid M., Niveau A. (2017a) Checking the consistency of combined qualitative constraint networks. In: AAAI, pp 1084–1090
Cohen-Solal Q., Bouzid M., Niveau A. (2017b) Temporal sequences of qualitative information: reasoning about the topology of constant-size moving regions. Twenty-sixth International Joint Conference on Artificial Intelligence IJCAI 2017:986–992
Cohn A., Li S., Liu W., Renz J. (2014) Reasoning about topological and cardinal direction relations between 2-dimensional spatial objects. J. Artif. Intell. Res. (JAIR) 51:493–532
Condotta J.-F., D'Almeida D. (2011) Consistency of qualitative constraint networks from tree decompositions. In: Combi C., Leucker M., Wolter F. (eds.) Proceedings of the 18th international symposium on temporal representation an reasoning (TIME'11), Lübeck, Germany, pp 149–156
Condotta J.-F., Kaci S., Schwind N. (2009) Merging qualitative constraint networks defined on different qualitative formalisms. In: Hornsby K. S., Claramunt C., Denis M., Ligozat G. (eds) COSIT. Lecture notes in computer science, vol 5756. Springer, Berlin, pp 106–123
Condotta J.-F., Ligozat G. (2004) Axiomatizing the cyclic interval calculus. In: Proceedings of KR'2004, pp 95–105
Condotta J-F., Ligozat G., Saade M. (2006a) A generic toolkit for n-ary qualitative temporal and spatial calculi. The 13th International Symposium on Temporal Representation and Reasoning (TIME'06). Budapest, Hungary, pp 78–86
Condotta J.-F., Ligozat G., Saade M., Tripakis S. (2006b) Ultimately periodic simple temporal problems (UPSTPs). In: MOI (ed.) Time. IEEE Computer Society, pp 69–77
Condotta J-F., Ligozat G., Tripakis S.(2005) Ultimately periodic qualitative constraint networks for spatial and temporal reasoning. ICTAI. IEEE Computer Society 584–588
Condotta J.-F., Nouaouri I., Sioutis M. (2016) A SAT approach for maximizing satisfiability in qualitative spatial and temporal constraint networks. In: Baral C., Delgrande J.P., Wolter F. (eds). Principles of knowledge representation and reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, 25–29 April 2016. AAAI, pp 432–442
Cotteret G. (2005). Extraction d'éléments curvilignes guidée par des mécanismes attentionnels pour des images de télédétection : approche par fusion de données. PhD thesis, Université Paris-Sud, France
Dague P. (1993a) Numeric reasoning with relative orders of magnitude. In: Proceedings of the National Conference on Artificial Intelligence, pp 541-547
Dague P. (1993b) Symbolic reasoning with relative orders of magnitude. In: Proceedings of the International Joint Conference on Artificial Intelligence, vol 13. Lawrence Erlbaum Associates Ltd, USA, p 1509

Dague P. (1995) Qualitative reasoning: a survey of techniques and applications. AI Communications 8(3/4):119–192

Dague P., Travé-Massuyès L. (2004) Raisonnement causal en physique qualitative. Intellectica. 38:247–290

Dauphin-Tanguy G et al. (2000) Les bond graphs. Hermès Science, Paris

de Beuvron F. D. B., Marc-Zwecker S., Zanni-Merk C., Le Ber F. (2015) Combining ontological and qualitative spatial reasoning: application to urban images interpretation. In: Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (IC3K (2013) CCIS, vol 454. Springer, Berlin, pp 182–198

de Jong H., Geiselmann J., Hernandez C., Page M. (2003) Genetic network analyzer: qualitative simulation of genetic regulatory networks. Bioinformatics 19(3):336–344

de Kleer J. (1977) Multiple representations of knowledge in a mechanics problem-solver. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, USA, pp 299–304

de Kleer J. (1979) Causal and teleological reasoning in circuit recognition. Massachusetts Institute of Technology, Cambridge

de Kleer J., Brown J. (1984) A qualitative physics based on confluences. Artif. Intell. 24(1–3):7–83

de Kleer J., Brown J. (1986) Theories of causal ordering. Artif. Intell. 29(1):33–61

de Koning K., Bredeweg B., Breuker J., Wielinga B. (2000) Model-based reasoning about learner behaviour. Artif. Intell. 117(2):173–229

Dylla F., Mossakowski T., Schneider T., Wolter D. (2013) Algebraic properties of qualitative spatio-temporal calculi. In: Spatial Information Theory, proceedings of COSIT-13. Springer, Berlin, pp 516–536

Egenhofer M. J. (1989) A formal definition of binary topological relationships. In: Litwin W., Schek H.-J. (eds) FODO. Lecture notes in computer science, vol 367. Springer, Berlin, pp 457–472

Egenhofer M. J. (1991) Reasoning about binary topological relations. Lecture notes in computer science 525:143–160

Euzenat J. (1996) An algebraic approach for granularity in qualitative space and time representation. In: IJCAI-95, pp 894–900

Euzenat J. (2001) Granularity in relational formalisms with application to time and space. Comput. Intell. 17(4):703–737

Euzenat J. (2008) Algebras of ontology alignment relations. Springer, Berlin

Euzenat J., Montanari A. (2005) Time granularity. Handbook of temporal reasoning in Artificial Intelligence, Chapter time granularity. Elsevier, Amsterdam, pp 59–118

Falkenhainer B., Forbus Dedre K. (1989) The structure-mapping engine: algorithm and examples. Artif. Intell. 41(1):1–63

Forbus K. (1984) Qualitative process theory. Artif. Intell. 24(1–3):85–168

Forbus K., Mostek T., Ferguson R. (2002) An analogy ontology for integrating analogical processing and first-principles reasoning. In: Proceedings of the National Conference on Artificial Intelligence, pp 878–885

Forbus K. D. (2014) Qualitative reasoning about space and motion. Mental models. Psychology, UK, pp 61–82

Freksa C. (1992) Using orientation information for qualitative spatial reasoning. In: Frank A. U., Campari I., Formentini U. (eds) Spatio-temporal reasoning. Lecture notes in computer science, vol 639. Springer, Berlin, pp 162–178

Ganter B., Wille R. (1999) Formal concept analysis. Springer, Berlin

Gantner Z., Westphal M., Wölfl S. (2008) GQR- a fast reasoner for binary qualitative constraint calculi. In: Proceedings of the AAAI'08 workshop on Spatial and Temporal Reasoning, Chicago, USA

Gerevini A., Nebel B. (2002) Qualitative spatio-temporal reasoning with RCC-8 and Allen's interval calculus: computational complexity. In: van Harmelen F. (ed.) Proceedings of ECAI 2002. IOS, pp 312–316

Gerevini A., Renz J. (2002) Combining topological and size information for spatial reasoning. Artif. Intell. 137(1–2):1–42
Ghallab M., Alaoui A. M. (1989) Managing efficiently temporal relations through indexed spanning trees. In: IJCAI, pp 1297–1303
Goyal R. K., Egenhofer M. J. (1997) The direction-relation matrix: a representation for directions relations between extended spatial objects. In: The annual assembly and the summer retreat of University Consortium for geographic information systems science, Bar Harbor, ME
Guerrin F. (1991) Qualitative reasoning about an ecological process: interpretation in hydroecology. Ecol. Model. 59(3–4):165–201
Güsgen H. (1989) Spatial reasoning based on Allen's temporal logic. Technical report TR-89-049, ICSI, Berkeley, CA
Hayes P. (1979) The naive physics manifesto. Expert systems in the microelectronic age 242–270
Hayes P. (1985) The second naive physics manifesto. In: Hobbs J., Moore R. (eds.) Formal theories of the commonsense world, pp 1-36
Hobbs J. R. (1985) Granularity. In: Proceedings of IJCAI-85, pp 432–435
Hofer B., Nica I., Wotawa F. (2017) Qualitative deviation models versus quantitative models for fault localization in spreadsheets. In: 30th International Workshop on Qualitative Reasoning (QR), IJCAI 2017, Melbourne, Australia
Inants A. (2016) Qualitative calculi with heterogeneous universes. PhD thesis, Grenoble Alpes University, France
Ironi L., Panzeri L., Plahte E. (2008) An algorithm for qualitative simulation of gene regulatory networks with steep sigmoidal response functions. Algebraic biology, pp 110–124
Ironi L., Tentoni S. (2007) Automated detection of qualitative spatio-temporal features in electrocardiac activation maps. Artif. Intell. Med. 39(2):99–111
Iwasaki Y. (1997) Real-world applications of qualitative reasoning. IEEE Expert Intell. Syst. Appl. 12(3):16–21 Special issue
Iwasaki Y., Simon H. (1986) Causality in device behavior. Artif. Intell. 29(1):3–32
Iwasaki Y., Simon H. (1994) Causality and model abstraction. Artif. Intell. 67(1):143–194
Jeansoulin R., Papini O. (2007) Underwater archaeological knowledge analysis and representation in the VENUS project: a preliminary draft. In: Georgopoulos A. (ed) XXI international CIPA symposium. The international archives of photogrammetry, remote sensing and spatial information sciences, vol XXXVI-5/C53. ICOMOS/ISPRS Committee for Documentation of Cultural Heritage, pp 394–399
Jonsson P., Bäckström C. (1998) A unifying approach to temporal constraint reasoning. Artif. Intell. 102(1):143–155
Kansou K., Bredeweg B. (2014) Hypothesis assessment with qualitative reasoning: modelling the Fontestorbes fountain. Ecol. Inform. 19:71–89
Khatib L. (1994) Reasoning with non-convex time intervals. PhD thesis, Florida Institute of Technology, Melbourne, Florida
Koubarakis M. (1996) Tractable disjunctions of linear constraints. In: Freuder, E. C. (ed.) CP. Lecture notes in computer science, vol 1118. Springer, Berlin, pp 297–307
Koubarakis M. (2001) Tractable disjunctions of linear constraints: basic results and applications to temporal reasoning. Theor. Comput. Sci. 266(1–2):311–339
Kuipers B. (1985) The limits of qualitative simulation. In: Proceedings of the 9th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, USA, pp 128–136
Kuipers B. (1986) Qualitative simulation. Artif. Intell. 29(3):289–338
Kuipers B. (1994) Qualitative reasoning: modeling and simulation with incomplete knowledge. MIT, Cambridge
Lancaster K. (1965) The theory of qualitative linear systems. Econometrica: J of the Econometric Society 33(2):395–408
Lascarides A., Asher N. (1991) Discourse relations and defeasible knowledge. In: ACL, pp 55–62
Lascarides A., Asher N. (1993) Temporal interpretation, discourse relations, and commonsense entailment. Linguistics and Philosophy 16:437–493

Le Ber F., Ligozat G., Papini O. (eds) (2007) Raisonnements sur l'espace et le temps. Hermès / Lavoisier, Paris

Le Ber F., Napoli A. (2003) Design and comparison of lattices of topological relations for spatial representation and reasoning. J. Exp. Theor. Artif. Intell. 15(3):331–371

Le Ber F., Napoli A., Metzger J-L., Lardon S. (2003) Modeling and comparing farm maps using graphs and case-based reasoning. J. Univers. Comput. Sci. 9(9):1073–1095

Levesque H., Brachman R. (1985) A fundamental tradeoff in knowledge representation and reasoning. In: Brachman R. J., Levesque H. (eds) Knowledge representation and reasoning. Morgan Kaufmann, Stanford

Li H, Muñoz-Avila H., Bransen D., Hogg C., Alonso R. (2009a) Spatial event prediction by combining value function approximation and case-based reasoning. In: McGinty L., Wilson D. (eds) ICCBR, (2009) LNAI 5650. Springer, Berlin, pp 465–478

Li J. J, Huang J., Renz J. (2009b) A divide-and-conquer approach for solving interval algebra networks. In [Boutilier 2009], pp 572–577

Li S., Ying M. (2003) Region connection calculus: its models and composition table. Artif. Intell. 145(1–2):121–146

Ligozat G. (1990) Weak representations of interval algebras. In: Proceedings of AAAI-90, pp 715–720

Ligozat G. (1991) On generalized interval calculi. In: Proceedings of AAAI-91, pp 234–240

Ligozat G. (1993) Qualitative triangulation for spatial reasoning. In: Frank A. U., Campari I. (eds) Spatial information theory (COSIT'93). LNCS, vol 716. Springer, Berlin, pp 54–68

Ligozat G. (1994) Tractable relations in temporal reasoning: pre-convex relations. In: Anger F. D., Güsgen H., Ligozat G. (eds) Proceedings of the ECAI-94 workshop on Spatial and Temporal Reasoning, Amsterdam, pp 99–108

Ligozat G. (1996) A new proof of tractability for ORD-Horn relations. In: Proceedings of AAAI-96, pp 395–401

Ligozat G. (2001) When tables tell it all. In: Montello D. R. (ed) COSIT. Lecture notes in computer science, vol 2205. Springer, Berlin, pp 60–75

Ligozat G. (2013) Qualitative spatial and temporal reasoning. Wiley, New Jersey

Ligozat G., Nowak J., Schmitt D. (2007) From language to pictorial representations. In: Vetulani Z. (ed) Proceedings of the Language and Technology Conference (L&TC'07), Poznań, Poland. Wydawnictwo Poznańskie

Ligozat G., Renz J. (2004) What is a qualitative calculus? a general framework. In: Proceedings of PRICAI'04, LNCS 3157, New Zealand, Auckland, pp 53–64

Ligozat G., Vetulani Z., Osiński J. (2011) Spatiotemporal aspects of the monitoring of complex events for public security purposes. Spat. Cogn. Comput. 11(1):103–128

Liu W., Li S. (2012) Solving minimal constraint networks in qualitative spatial and temporal reasoning. In: Principles and practice of constraint programming - 18th international conference, CP 2012, Québec City, Canada, 8–12 October 2012, Proceedings, pp 464–479

Liu W., Li S., Renz J. (2009) Combining RCC-8 with qualitative direction calculi: algorithms and complexity. In [Boutilier 2009], pp 854–859

Long Z., Li S. (2015) On distributive subalgebras of qualitative spatial and temporal calculi. In: Spatial Information Theory - 12th International Conference, COSIT 2015, Santa Fe, NM, USA, 12–16 October 2015, Proceedings, pp 354–374

Loustau P., Nodenot T., Gaio M. (2008) Spatial decision support in the pedagogical area: processing travel stories to discover itineraries hidden beneath the surface. In: The European information society – taking geoinformation science one step further, Proceedings of the 11th Agile International Conference on Geographic Information Science (AGILE 2008), LNCG, pp 359–378

Mark D., Comas D., Egenhofer M., Freudschuh S., Gould M., Nunes J. (1995) Evaluating and refining computational models of spatial relations through cross-linguistic human-subjects testing. In: Frank A. U., Kuhn W. (eds) Spatial information theory, a theoretical basis for GIS, LNCS 988. International Conference COSIT'95. Springer, Berlin

McKinsey J., Tarski A. (1944) The algebra of topology. Annals of mathematics 45:141–191

Miron A. D., Gensel J., Villanova-Oliver M., Martin H. (2007) Relations spatiales qualitatives dans les ontologies géographiques avec ONTOAST. In: SAGEO 2007, Rencontres internationales Géomatique et territoire

Montserrat-Adell J, Sánchez M., Ruiz F. J., Agell N. (2016) From qualitative absolute order-of-magnitude to the extended set of hesitant fuzzy linguistic term sets. In: 29th International Workshop on Qualitative Reasoning (QR), IJCAI 2016, New York, USA

Moore R. (1966) Interval analysis. Englewood Cliffs, New Jersey

Mossakowski T., Schröder L., Wölfl, S. (2006) A categorical perspective on qualitative constraint calculi. In: Qualitative constraint calculi: application and integration, workshop at KI 2006, proceedings, pp 28–39

Muller P. (1998) Éléments d'une théorie du mouvement pour la formalisation du raisonnement spatio-temporel de sens commun. PhD thesis, IRIT, Université Paul Sabatier, Toulouse, France

Muscettola N., Nayak P., Pell B., Williams B. (1998) Remote agent: to boldly go where no AI system has gone before. Artif Intell 103(1–2):5–47

Napoli A., Le Ber F. (2007) The Galois lattice as a hierarchical structure for topological relations. Ann. Math. Artif. Intell. 49(1–4):171–190

Ndiaye A., Della Valle G., Roussel P. (2009) Qualitative modelling of a multi-step process: the case of French breadmaking. Expert Syst. Appl. 36(2):1020–1038

Nebel B. (1996) Solving hard qualitative temporal reasoning problems: evaluating the efficiency of using the ORD-Horn class. In: Proceeding of the twelfth European Conference on Artificial Intelligence (ECAI'96)

Nebel B., Bürckert H.-J. (1995) Reasoning about temporal relations: a maximal tractable subclass of Allen's interval algebra. J ACM 42(1):43–66

Osmani A. (1999) Introduction to reasoning about cyclic intervals. In: Imam I., Kodratoff Y., El-Dessouki A., Ali M. (eds) Multiple approaches to intelligent systems, Proceedings of IEA/AIE-99. Springer LNCS, vol 1611, pp 698–706

Osmani A., Lévy F. (2000) A constraint-based approach to simulate faults in telecommunication networks. In: Loganantharaj R., Palm G. (eds) IEA/AIE. Lecture notes in computer science, vol 1821. Springer, Berlin, pp 463–473

Picardi C., Bray R., Cascio F., Console L., Dague P., Dressler O., Millet D., Rehfus B., Struss P., Vallée C. (2002) IDD: integrating diagnosis in the design of automotive systems. In: Proceedings of the European Conference on Artificial Intelligence, pp 628–632

Poupeau B., Bonin O. (2006) 3D Analysis with high-level primitives: a crystallographic approach. In: Progress in spatial data handling, proceedings of SDH'06. Springer, Berlin, pp 599–616

Prior A. (1957) Time and Modality. Clarendon, Oxford

Prior A. (1967) Past. Oxford University, Oxford, Present and Future

Przytula-Machrouh E., Ligozat G., Denis M. (2004) Vers des ontologies transmodales pour la description d'itinéraires: Le concept de scène élémentaire. Revue Internationale de Géomatique

Pujari A. K, Kumari G. V, Sattar A. (1999) INDU: an interval and duration network. In: Australian joint conference on Artificial Intelligence, pp 291–303

Raiman O. (1991) Order of magnitude reasoning. Artif. Intell. 51(1–3):11–38

Randell D., Cui Z., Cohn T. (1992a) An interval logic for space based on connection. In: Neumann B. (ed) Proceedings of ECAI-92. Wiley, New Jersey, pp 394–398

Randell D., Cui Z., Cohn T. (1992b) A spatial logic based on regions and connection. In: Neumann B. (ed) Proceedings of KR-92, CA. Morgan Kaufmann, San Mateo, pp 165–176

Renz J. (1999) Maximal tractable fragments of the region connection calculus: a complete analysis. In: Dean T. (ed) IJCAI. Morgan Kaufmann, USA, pp 448–455

Renz J., Nebel B. (2007) Qualitative spatial reasoning using constraint calculi. In [Aiello et al. 2007a], pp 161–215

Roselló L., Prats F., Agell N., Sánchez M. (2010) Measuring consensus in group decisions by means of qualitative reasoning. Int J Approx Reason 51(4):441–452

Ross N., Bradley E., Hertzberg J. (2006) Dynamics-informed data assimilation in a qualitative fluids model. In: Proceedings of the 20th International Workshop on Qualitative Reasoning

Sioutis M., Condotta J.-F., Salhi Y., Mazure B. (2015a) Generalized qualitative spatio-temporal reasoning: complexity and tableau method. In: Proceedings of the 24th International Conference automated reasoning with analytic tableaux and related methods (TABLEAUX'15), pp 54–69

Sioutis M., Li S., Condotta J.-F. (2015b) Efficiently characterizing non-redundant constraints in large real world qualitative spatial networks. In: Proceedings of the twenty-fourth International Joint Conference on Artificial Intelligence (IJCAI'15), pp 3229–3235

Stell J. (2000) Boolean connection algebras: a new approach to the region-connection calculus. Artif. Intell. 122:111–136

Struss P. (2002) Automated abstraction of numerical simulation models-theory and practical experience. In: Proceedings of the sixteenth International Workshop on Qualitative Reasoning, Sitges, Catalonia, Spain

Struss P., Price C. (2003) Model-based systems in the automotive industry. AI Mag 24(4):17

Struss P., Sterling R., Febres J., Sabir U., Keane M. M. (2014) Combining engineering and qualitative models to fault diagnosis in air handling units. In: Proceedings of the twenty-first European Conference on Artificial Intelligence. IOS, Amsterdam, pp 1185–1190

Tarski A. (1941) On the calculus of relations. J. Symb. Log 6(3):73–89

Top J., Akkermans H.(1991) Computational and physical causality. In: Proceedings of the international joint conference of Artificial Intelligence, pp 1171–1176

Travé L., Dormoy J. (1988) Qualitative calculus and applications. In: IMACS transactions on scientific computing'88, pp 53–61

Travé L., Kaszkurewicz E. (1986) Qualitative controllability and observability of linear dynamical systems. Proceedings of the IFAC/IFORS Symposium on Large Scale Systems: Theory and Applications 2:964–970

Travé-Massuyés L., Dague P. (2003) Modèles et raisonnements qualitatifs. Hermès

Travé-Massuyès L., Dormoy J. (1990) Numéro Spécial sur le Raisonnement Qualitatif. Revue d'Intelligence Artificielle 3/4

Travé-Massuyès L., Dormoy J., Guerrin F. (1997) Le raisonnement qualitatif pour les sciences de l'ingénieur (coll. Hermès, Diagnostic et Maintenance)

Travé-Massuyès L., Ironi L., Dague P. (2003) Mathematical foundations of qualitative reasoning. AI Mag 24(4):91

Travé-Massuyès L., Milne R. (1997) Gas-turbine condition monitoring using qualitative model-based diagnosis. IEEE Expert Intell Syst Appl 12(3):22–31

Travé-Massuyès L., Milne R. (2009) Application oriented qualitative reasoning. The Knowledge Engineering Review 10(02):181–204

Travé-Massuyès L., Piera N. (1989) The orders of magnitude models as qualitative algebras. In: Proceedings of the 11th International Joint Conference on Artificial Intelligence -vol 2. Morgan Kaufmann, USA, pp 1261–1266

Travé-Massuyès L, Prats F, Sánchez M., Agell N. (2005) Relative and absolute order-of-magnitude models unified. Ann Math Artif. Intell. 45(3):323–341

van Beek P. (1990) Reasoning about qualitative temporal information. In: Proceedings of AAAI-90, Boston, MA, pp 728–734

van Beek P., Manchak D. W. (1996) The design and experimental analysis of algorithms for temporal reasoning. J. Artif. Intell. Res 4:1–18

van de Weghe N. (2004) Representing and reasoning about moving objects: a qualitative approach. PhD thesis, Ghent University

Vieu L. (1991) Sémantique des relations spatiales et inférences spatio-temporelles: Une contribution à l'étude des structures formelles de l'espace en Langage Naturel. PhD thesis, Université Paul Sabatier, Toulouse, France

Vilain M., Kautz H. A., van Beek P. G. (1989) Constraint propagation algorithms for temporal reasoning: a revised report. In: Weld D, de Kleer J (eds) Readings in qualitative reasoning about physical systems. Morgan Kaufmann, USA

Vilain M. B. (1982) A system for reasoning about time. In: Proceedings of AAAI-82, pp 197–201

Wallgrün J. O., Frommberger L., Wolter D., Dylla F., Freksa C. (2006a). Qualitative spatial representation and reasoning in the sparQ-toolbox. In [Barkowsky et al. 2008], pp 39–58
Wallgrün J. O., Frommberger L., Wolter D., Dylla F., Freksa C. (2006b) Qualitative spatial representation and reasoning in the SparQ-toolbox. In [Barkowsky et al. 2008], pp 39–58
Weld D., de Kleer J. E. (1989) Readings in qualitative reasoning about physical systems. Morgan Kaufmann, San Francisco
Westphal M. (2014) Qualitative Constraint-based Reasoning: methods and applications. PhD thesis, Universitt Freiburg
Westphal M., Hué J., Wölfl S. (2014) On the scope of qualitative constraint calculi. KI 2014 Advances in Artificial Intelligence. Springer, Berlin, pp 207–218
Westphal M., Wöfl S. (2008) Bipath consistency revisited. In: Proceedings of the ECAI workshop on Spatial and Temporal Reasoning
Westphal M., Wölfl S .(2009) Qualitative CSP, finite CSP, and SAT: comparing methods for qualitative constraint-based reasoning. In [Boutilier 2009], pp 628–633
Williams B., Nayak P. (1996) A model-based approach to reactive self-configuring systems. In: Proceedings of the National Conference on Artificial Intelligence, pp 971–978
Wolter F., Zakharyaschev M. (2000) Spatio-temporal representation and reasoning based on RCC-8. In: Proceedings of the Seventh International Conference KR 2000. Morgan Kaufmann, USA, pp 3–14
Würbel E., Jeansoulin R., Papini O. (2000) Revision: an application in the framework of GIS. KR 2000:505–515
Yang Y., Atif J., Bloch I. (2015) Abductive reasoning using tableau methods for high-level image interpretation. 38th Annual German conference on AI. Dresden, Germany, pp 356–365
Yilmaz O., Say A. (2006) Causes of ineradicable spurious predictions in qualitative simulation. J. Artif. Intell. Res. 27:551–575

Reasoning with Ontologies

Meghyn Bienvenu, Michel Leclère, Marie-Laure Mugnier and Marie-Christine Rousset

Abstract This chapter considers the notion of a formal ontology, which is a conceptual vocabulary equipped with a logical semantics. Three families of knowledge representation and reasoning formalisms that put ontologies at the core of any knowledge base are presented, namely: description logics, conceptual graphs and existential rules. We present the main knowledge constructs and dialects of these families, as well as the main reasoning problems with their complexity. We highlight the relationships between these families and compare them from an expressivity viewpoint.

1 Introduction

Knowledge-based systems exploit formal representations of knowledge to solve different kinds of problems. The fundamental formalism to represent and do reasoning on knowledge is classical first-order logic. Whereas a significant amount of work in knowledge representation aimed to extend classical logic to handle more complex notions (like time, modalities, preferences, ...), most work on ontologies was devoted to simpler logical fragments and to the study of tradeoffs between the expressivity of the representation languages and the computational complexity of reasoning in these languages.

A commonly adopted definition of an *ontology* is that of an explicit specification of the conceptualisation of a domain (Gruber 1993). All ontologies include at least

M. Bienvenu (✉) · M. Leclère · M.-L. Mugnier
LIRMM-CNRS and Université de Montpellier- Inria, Montpellier, France
e-mail: meghyn@lirmm.fr

M. Leclère
e-mail: leclere@lirmm.fr

M.-L. Mugnier
e-mail: mugnier@lirmm.fr

M.-C. Rousset
LIG-CNRS and Université de Grenoble, Grenoble, France
e-mail: Marie-Christine.Rousset@imag.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_6

a conceptual vocabulary, i.e., a set of terms (in the natural language sense) used to model a domain, provided with a specification of their meaning. These terms represent *concepts* (or classes), i.e., the categories of entities in the modeled domain, as well as *relations* (or properties, roles) which may stand between entities. Concepts and relations may be further specified in different ways, depending on the expressivity of the ontological language. They are usually organized into a specialisation/generalisation hierarchy by means of axioms stating that a concept (respectively a relation) is a subconcept (respectively a subrelation) of another. Other typical ontological axioms include concept disjointness (which expresses that two concepts cannot have common instances), the domain and range of binary relations (which specifies the classes of entities that can be linked by this relation), algebraic properties of relations (for instance that a relation is symmetric or transitive), mandatory relations for instances of a class (for instance that every entity of a given class fulfils a given property), and so on.

Ontologies are widely used in data and knowledge management and they are at the core of the Semantic Web (see chapter "Semantic Web" of Volume 3). We refer the reader to the chapter on knowledge engineering (chapter "Knowledge Engineering" of this volume) for developments on building and using ontologies.

Without denying the importance of the linguistic aspects in ontologies, we focus in this chapter on formal ontologies. Therefore, an ontology will be seen as a logical theory that specifies the expected meaning of the conceptual vocabulary (Guarino 1998). More specifically, an ontology is given by a formal vocabulary (or signature) and a set of formulas built on this vocabulary, which define the acceptable models of the considered domain. Hence, any reasoning that takes into account an ontology $\mathcal{O}$ considers only the models of $\mathcal{O}$: for instance, given two pieces of knowledge G and F, deciding if G is a logical consequence of F, which we denote by $F \models G$, becomes $\mathcal{O}, F \models G$, i.e., is every model of $\mathcal{O}$ and F a model of G? Moreover, in most settings, the *unique name assumption* is made: in this case, distinct logical constants are necessarily interpreted by distinct elements of the domain of any interpretation.

In this chapter, we consider *knowledge bases* composed of two types of knowledge: on the one hand, ontological knowledge, which is general knowledge about the modelled domain, and factual knowledge, composed of facts or assertions about specific entities. Usually, a fact is a ground atom, i.e., has no variable.[1]

A parallel can be drawn between a knowledge base and a classical database (e.g., a relational database). Indeed, the database schema, which includes a vocabulary and integrity constraints, can be associated with an ontology, while data can be seen as factual knowledge. However, some important differences should be noted. In databases, data are supposed to encode a complete description of the 'world'. In other words, the *closed world assumption* is made (everything that is not asserted in the database is considered as false), as well as the related closed domain assumption (the only existing entities are those encoded in the data). By contrast, the *open world assumption* is made in knowledge bases (as well as the related open domain

[1] In the Semantic Web area, and specifically concerning the OWL language, the term ontology often includes both kinds of knowledge. Hence, it corresponds to our notion of knowledge base.

assumption); this often leads to more complex reasoning since a knowledge base encodes a possibly infinite set of all of the descriptions of the world that include the known facts and comply with the ontology. For that reason, the use of negation is often restricted, as the excluded-middle law (stating that a proposition is either true or false) leads to combinatorial explosion. The open world assumption may lead to considering existentially quantified variables in facts (and not only constants) to denote unknown individuals. Moreover, the primary aim of databases is to store and retrieve data with efficient query answering techniques, whereas knowledge bases are used to infer new knowledge that was only implicitly represented in the ontology. However, the two domains are becoming progressively closer, especially under the impulse of the Semantic Web. Indeed, there is an increasing interest in answering complex queries on large knowledge bases, on the one hand, and, on the other hand, dropping the closed world assumption in databases.

This chapter is devoted to several knowledge representation and reasoning formalisms used to build and exploit knowledge bases: description logics, graph-based representations (issued from conceptual graphs) and the more recent existential rule framework. Although description logics and graph-based representations are both rooted in semantic networks (Lehmann 1992), their development from the 80's followed different research lines, as explained in the next sections. Existential rules can be seen both as the logical counterpart of the graph-based framework and as a generalisation of Datalog, the deductive database querying language.

Several different kinds of reasoning over knowledge bases have been considered, among which we distinguish the following fundamental problems. Given a knowledge base (KB) composed of an ontology $\mathcal{O}$ and a set of facts I, we consider the following questions:

- *Knowledge base satisfiability:* determine if the KB is satisfiable (or consistent), i.e., if it has at least one model.
- *Ontological knowledge entailment:* determine if a piece of ontological knowledge o is entailed by the ontology $\mathcal{O}$, i.e., if $\mathcal{O} \models o$ holds.
- *Fact entailment:* determine if a fact is entailed by the KB, i.e., if $\mathcal{O}, I \models o$ holds.
- *Ontology-mediated query answering:* compute the answers to a query q over the KB; when q is a Boolean query (i.e., a query with a yes/no answer), the problem is whether q is entailed by the KB, i.e., whether $\mathcal{O}, I \models q$ holds. The general form of a query q is a first-order formula with possibly free variables, say $(x_1, \ldots, x_k)$. Then an answer to q in the KB is a tuple of constants $(c_1, \ldots, c_k)$ such that the Boolean query obtained from q by substituting each variable x_i by the constant c_i is entailed by the KB.

The three formalisms presented in this chapter tackle the above problems, the difference being in their expressivity and the kind of query considered. Description logics traditionally allowed for rich descriptions of ontological axioms using different kinds of constructors; standard reasoning tasks were KB satisfiability, concept subsumption (determine if a concept is a specialisation of another, which is a special case of ontological knowledge entailment) and instance checking (determine if a specific individual is an instance of a given concept, which is a special case of fact

entailment). Hence, only very specific queries were considered (single atoms without variables). The growing interest for exploiting large and complex data led the description logic community to investigate more expressive queries, however at the price of less expressive description logics, known as lightweight description logics. The queries most commonly considered so far in the context of ontology-mediated query answering are so-called *conjunctive queries*, which are the basic queries in databases: these are existentially quantified conjunctions of atoms. Conjunctive queries are natural queries in the graph-based and existential rule frameworks, but, on the other hand, these formalisms do not offer the variety of ontological axioms found in classical description logics. Some lightweight description logics, however, can be seen as special cases of the graph-based and existential rule frameworks.

The sequel of this chapter introduces each of these three formalisms and compares them from an expressivity viewpoint.

2 Description Logics

Description logics (DLs) (Baader et al. 2003, 2017) are family of knowledge representation languages corresponding to decidable[2] fragments of first-order logic using only unary and binary predicates. While the lack of higher-arity predicates may seem a strong restriction, it turns out that unary and binary predicates (classes and properties) capture a large part of modelling needs. Indeed, DLs provide the basis of the OWL Web Ontology Language (W3C 2004a), a W3C-standardized ontology language for the Semantic Web (Berners-Lee et al. 2001), and RDF (W3C 2004b), a popular format for Web data, is likewise restricted to unary and binary predicates.

A DL knowledge base (KB) has two parts: a *TBox* that contains general knowledge about the application domain, and an *ABox* that contains facts about particular individuals. The TBox can be viewed as an ontology, which provides a conceptual model for the data stored in the ABox. What distinguishes different DLs is the type of knowledge that can be expressed in the TBox.

Traditionally, the main reasoning problems considered by the DL community are: KB satisfiability, subsumption, and instance checking. Satisfiability testing is essential for identifying modelling errors, while instance checking and subsumption are used to identify TBox axioms and ABox assertions that follow from the knowledge of the KB. As the latter two tasks correspond to forms of logical entailment, they can be reduced to unsatisfiability testing for all DLs that admit full negation. Our discussion of DLs will center on these traditional reasoning tasks. However, we should point out that over the past decade, several additional reasoning tasks for DLs have been investigated, most notably, ontology-mediated conjunctive query answering, which allows for richer queries to be posed over the ABox, but which cannot be reduced to satisfiability testing and thus required the development of new algorithmic techniques (see survey Bienvenu and Ortiz 2015 and references therein). There

[2] A few undecidable DLs have been studied.

has also been quite a lot of work on reasoning support for building, debugging, and evolving ontologies, e.g., providing explanations for why a given entailment holds (Schlobach and Cornet 2003; Sebastiani and Vescovi 2009; Peñaloza and Sertkaya 2017), or extracting modules of an ontology that conform to some criterion (Grau et al. 2008; Kontchakov et al. 2010; Konev et al. 2013).

Early work on DLs in the 1980's mostly focused on building reasoning systems, and it was only later that it was discovered that some of these DLs were in fact undecidable or at the very least intractable. These initial negative results led to the introduction of simple DLs for which polynomial-time reasoning was possible, but which turned out to be too limited in their expressivity. In the late 1990's, however, new systems were developed based upon highly optimized tableaux algorithms, which demonstrated acceptable performance for expressive DLs despite their high worst-case complexity. This line of work continues to this day, with ever more sophisticated optimisations targeting ever more expressive DLs. At the same time, there has been renewed interest in lightweight DLs that provide the required scalability for applications involving very large TBoxes and/or ABoxes. Importantly, however, this new breed of low-complexity DLs provides combinations of modelling constructs that are much better suited to the needs of real-world applications than the previous generation of simple DLs.

Nowadays, there is an extensive body of results pinpointing the exact computational complexity of performing different kinds of reasoning in the whole range of DLs, allowing one to choose the optimal trade-off between expressivity and efficiency of reasoning for the application at hand. For an overview of the complexity landscape, interested readers can consult the surveys (Ortiz and Simkus 2012) and (Bienvenu and Ortiz 2015).

In this section, we introduce the basics of description logics, and then present several concrete DLs and show how varying the expressivity of the DL impacts the complexity of reasoning.

2.1 Preliminaries: DL Syntax and Semantics

In DL jargon, classes are called *concepts* and properties are called *roles*. DL knowledge bases are built starting from a set $\mathsf{N_C}$ of *atomic concepts* (unary predicates), a set $\mathsf{N_R}$ of *atomic roles* (binary predicates), and a set $\mathsf{N_I}$ of *individuals* (constants). We typically use $A, B, \ldots$ for atomic concepts, $P, Q, \ldots$ for atomic roles, and $a, b, \ldots$ for individuals. More complex concept and role expressions can be built using different constructors, with the set of available constructors depending on the particular DL (see further for more details). We will use $C, D, \ldots$ to denote (possibly complex) concepts and R, S for (possibly complex) roles.

A DL *knowledge base* $\mathcal{K}$ is a pair $\langle \mathcal{T}, \mathcal{A} \rangle$, consisting of a TBox $\mathcal{T}$ and an ABox $\mathcal{A}$. A *TBox* is a finite set of axioms expressing the relationships holding between different concepts and roles. The types of axioms allowed in the TBox depends on the choice of DL, but the most common forms of TBox axioms are *concept inclusions* ($C \sqsubseteq D$,

with C, D possibly complex concepts) and *role inclusions* ($R \sqsubseteq S$, with R, S possibly complex roles). Equivalences between concepts ($C \equiv D$) and roles ($R \equiv S$) are also common and can be seen as shorthand for inclusions in both directions (i.e., $C \equiv D$ is an abbreviation for the pair of inclusions $C \sqsubseteq D$ and $D \sqsubseteq C$).

An *ABox* is a finite set of *assertions* expressing that an individual belongs to a given concept ($C(a)$) or that a pair of individuals belongs to a role ($R(a, b)$). To simplify the presentation, we will assume in what follows that ABoxes only contain assertions involving atomic concepts and roles. This assumption can usually be made without loss of generality. For example, if we want to include $C(a)$ in the ABox, with C a general concept, it suffices to use the atomic assertion A_C (with A_C a fresh atomic concept) and add the inclusion $C \equiv A_C$ to the TBox.

The semantics of DL knowledge bases is defined in terms of (first-order) interpretations. An *interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ consists of a non-empty domain $\Delta^{\mathcal{I}}$ and an interpretation function $\cdot^{\mathcal{I}}$ that assigns a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ to every atomic concept $A \in \mathsf{N_C}$, a binary relation $P^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ to every atomic role $P \in \mathsf{N_R}$, and an element $a^{\mathcal{I}}$ to every individual $a \in \mathsf{N_I}$. It is common in DLs to adopt the *unique name assumption*, which states that distinct individuals are mapped to distinct elements of the interpretation domain.

An interpretation $\mathcal{I}$ *satisfies* a concept inclusion $C \sqsubseteq D$ (resp. role inclusion $R \sqsubseteq S$) if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ (resp. $R^{\mathcal{I}} \subseteq S^{\mathcal{I}}$). We say $\mathcal{I}$ is a *model* of a TBox $\mathcal{T}$ if it satisfies every axiom in $\mathcal{T}$. A TBox $\mathcal{T}$ logically implies an axiom α, written $\mathcal{T} \models \alpha$, if every model of $\mathcal{T}$ satisfies α. A fundamental reasoning task for TBoxes is *testing subsumption* between different concepts: given a TBox $\mathcal{T}$ and two concepts C and D, decide whether $\mathcal{T} \models C \sqsubseteq D$. An interpretation $\mathcal{I}$ *satisfies* a concept assertion $A(a)$ (resp. role assertion $P(a, b)$) if $a^{\mathcal{I}} \in A^{\mathcal{I}}$ (resp. $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in P^{\mathcal{I}}$). We call $\mathcal{I}$ a model of an ABox $\mathcal{A}$ if $\mathcal{I}$ satisfies every assertion in $\mathcal{A}$. An interpretation $\mathcal{I}$ is a model of a knowledge base $\langle \mathcal{T}, \mathcal{A} \rangle$ if it is a model of both $\mathcal{T}$ and $\mathcal{A}$. A KB $\mathcal{K}$ is *satisfiable* (or *consistent*) if it possesses at least one model. Testing satisfiability of a given KB is another standard reasoning task.

A KB *logically entails* a TBox axiom or ABox assertion α, written $\mathcal{K} \models \alpha$, if every model of $\mathcal{K}$ satisfies α. The *instance checking problem* is defined as follows: given a KB $\langle \mathcal{T}, \mathcal{A} \rangle$, a concept C, and an individual a, decide whether $\mathcal{K} \models C(a)$.

In the following sections, we give the syntax and semantics of the principal DL constructors by presenting a variety of DLs, ranging from 'simple' DLs $\mathcal{EL}$, $\mathcal{FL}_0$, and DL-Lite which offer polynomial-time reasoning (Sects. 2.2 and 2.3), to $\mathcal{ALC}$ (Sect. 2.4) which is often considered the prototypical DL, to highly expressive DLs like $\mathcal{SROIQ}$ (Sect. 2.5), which provide the logical foundations for OWL.

2.2 Lightweight Description Logics: $\mathcal{FL}_0$ and $\mathcal{EL}$

We begin by considering two DLs, $\mathcal{FL}_0$ and $\mathcal{EL}$, which are deliberately restricted in expressivity in order to allow for sound and complete polynomial-time reasoning. Both logics contain the *concept conjunction* constructor ($C_1 \sqcap C_2$), which corre-

$\mathsf{Prof} \sqsubseteq \mathsf{TeachingStaff}$	$\mathsf{TAssistant} \sqsubseteq \mathsf{TeachingStaff}$	$\mathsf{TAssistant} \sqsubseteq \mathsf{GradStudent}$
$\mathsf{TeachingStaff} \sqsubseteq \mathsf{Staff}$	$\mathsf{AdminStaff} \sqsubseteq \mathsf{Staff}$	$\mathsf{UndergradStudent} \sqsubseteq \mathsf{Student}$
$\mathsf{GradStudent} \sqsubseteq \mathsf{Student}$	$\mathsf{UndergradCourse} \sqsubseteq \mathsf{Course}$	$\mathsf{GradCourse} \sqsubseteq \mathsf{Course}$

Fig. 1 Example taxonomy of classes in the university domain

sponds to intersecting the classes represented by C_1 and C_2. Additionally, $\mathcal{FL}_0$ offers *qualified value restrictions* ($\forall R.C$), while $\mathcal{EL}$ offers *qualified existential restrictions* ($\exists R.C$), which provide suitably restricted forms of universal and existential quantification. In $\mathcal{EL}$, one can further use the top concept ($\top$), which denotes the class of all objects.

Figure 1 provides an example of a taxonomy of classes, formulated using a set of inclusions between atomic concepts. Such simple axioms form the backbone of real-world ontologies, and they are available in every DL (and in particular, in $\mathcal{FL}_0$ and $\mathcal{EL}$). The axioms in the first line of Fig. 1 stipulate that professors and teaching assistants are both kinds of teaching staff, and every teaching assistant is a graduate student. The remaining axioms state that teaching staff and admin staff are two types of staff and that students (resp. courses) can be specialized into undergraduate and graduate students (resp. courses).

In $\mathcal{FL}_0$ and $\mathcal{EL}$, we can additionally use the conjunction constructor to state that every student that is part of the teaching staff must be a graduate student:

$$\mathsf{Student} \sqcap \mathsf{TeachingStaff} \sqsubseteq \mathsf{GradStudent}$$

By making use of qualified value restrictions, we can express in $\mathcal{FL}_0$ that graduate students only take graduate courses and that a student that takes only graduate courses is a graduate student:

$$\mathsf{GradStudent} \sqsubseteq \forall \mathsf{takes}.\mathsf{GradCourse} \quad \mathsf{Student} \sqcap \forall \mathsf{takes}.\mathsf{GradCourse} \sqsubseteq \mathsf{GradStudent}$$

In $\mathcal{EL}$, we can use qualified existential restrictions to formulate the following axioms:

$$\mathsf{Student} \sqsubseteq \exists \mathsf{takes}.\mathsf{Course} \qquad \exists \mathsf{teaches}.\top \sqsubseteq \mathsf{TeachingStaff}$$
$$\exists \mathsf{teaches}.\mathsf{GradCourse} \sqsubseteq \mathsf{Prof} \qquad \mathsf{TeachingStaff} \sqsubseteq \exists \mathsf{teaches}.\mathsf{Course}$$

which state respectively that every student must take some course, that everyone who teaches something is a member of the teaching staff, that everyone who teaches a graduate course must be a professor, and that every member of teaching staff must teach some course.

The semantics of complex concepts built using the preceding constructors is defined recursively as follows (starting from the semantics of atomic concepts and roles which is directly provided by each interpretation):

Fig. 2 Translation from DLs to first-order logic (1)

DL notation	*Corresponding FOL formula*
$C_1 \sqcap C_2$	$C_1(X) \wedge C_2(X)$
$\exists R.C$	$\exists Y[R(X,Y) \wedge C(Y)]$
$\forall R.C$	$\forall Y[R(X,Y) \rightarrow C(Y)]$

- $\top^{\mathcal{I}} = \Delta^{\mathcal{I}}$
- $(C_1 \sqcap C_2)^{\mathcal{I}} = C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$
- $(\exists R.C)^{\mathcal{I}} = \{o_1 \mid \text{there exists } (o_1, o_2) \in R^{\mathcal{I}} \text{ such that } o_2 \in C^{\mathcal{I}}\}$
- $(\forall R.C)^{\mathcal{I}} = \{o_1 \mid (o_1, o_2) \in R^{\mathcal{I}} \text{ implies } o_2 \in C^{\mathcal{I}}\}$

Every DL concept can be translated into a first-order logic formula with one free variable. Figure 2 gives the first-order translation of concepts in $\mathcal{FL}_0$ and $\mathcal{EL}$, using X as the free variable. To improve readability, we commit a slight abuse of notation by using $C(X)$ to designate the translation of the concept C using free variable X. For example, if $C = A \sqcap \exists R.B$, with A and B atomic concepts, then $C(X)$ is the FOL formula $A(X) \wedge \exists Y[R(X, Y) \wedge B(Y)]$.

Every TBox axiom can be translated into a corresponding FOL sentence (that is, a formula without free variables): a concept inclusion $C \sqsubseteq D$ gives rise to the formula $\forall X(C(X) \rightarrow D(X))$ and an equivalence axiom $C \equiv D$ corresponds to the formula $\forall X(C(X) \leftrightarrow D(X))$.

The first polynomial-time reasoning procedures for lightweight DLs relied upon *structural subsumption*, in which concept expressions are first put into a normal form and then compared syntactically. This method can be used to show that subsumption between concepts w.r.t. an empty TBox is tractable in both $\mathcal{FL}_0$ and $\mathcal{EL}$. However, in most applications, one wishes to compute subsumption in the presence of a non-empty TBox. Rather interestingly, $\mathcal{FL}_0$ and $\mathcal{EL}$ exhibit dramatically different complexities for the general version of subsumption. In $\mathcal{FL}_0$, the problem becomes EXPTIME-complete (and thus provably intractable) (Baader et al. 2005) and remains coNP-hard even when restricted to TBoxes in the form of acyclic terminologies (Nebel 1990). By contrast, in $\mathcal{EL}$ (and several of its extensions), subsumption can be decided in PTIME in the presence of arbitrary TBoxes (Baader et al. 2005). This tractability result relies upon forward-chaining algorithms that construct in an iterative manner a subset of the axioms that are entailed from the TBox. A similar approach can be used to handle instance checking in $\mathcal{EL}$.

Nowadays, $\mathcal{EL}$ and its extensions have become popular ontology languages, whereas $\mathcal{FL}_0$ is no longer much in use. This is due in large part to the much more favourable computational properties of $\mathcal{EL}$, but also to the utility of the constructors provided by $\mathcal{EL}$. Indeed, while it was initially believed that value restrictions were more useful than existential restrictions, it turns out that (slight extensions of) $\mathcal{EL}$ closely match the modelling needs of many applications, particularly those in the biomedical domain. Indeed, the large-scale professional medical ontology SNOMED CT[3] (Systematized Nomenclature of Medicine, Clinical Terms),

[3] http://www.snomed.org/snomed-ct.

Fig. 3 Translation from DLs to first-order logic (2)

DL notation	*Corresponding FOL formula*
P^-	$P(Y,X)$
$\exists R$	$\exists Y R(X,Y)$
$\neg B$	$\neg B(X)$
$\neg R$	$\neg R(X,Y)$

developed by an international consortium for use in the health-care systems of several countries, is expressed in a tractable DL of the $\mathcal{EL}$ family. The importance of $\mathcal{EL}$ is further witnessed by the inclusion of the OWL 2 EL profile (W3C 2012b), based upon $\mathcal{EL}$, in the latest version of the W3C OWL standard.

2.3 DL-Lite: Another Lightweight Description Logic

The DL-Lite family of description logics (Calvanese et al. 2007) was proposed in the mid-2000's with the aim of supporting tractable reasoning while at the same time capturing the principal modelling primitives from conceptual modelling (more precisely, the entity-relationship models utilized in databases and information systems (Chen 1976) and UML[4] diagrams from software engineering). Another important motivation for introducing the DL-Lite family was to make it possible to answer more expressive queries by means of a reduction to relational databases.

In DL-Lite, complex concepts and roles can be constructed from atomic concepts and roles according to the following syntax:

$$B ::= A \mid \exists R \qquad C ::= B \mid \neg B \qquad R ::= P \mid P^- \qquad E ::= R \mid \neg R$$

where A is an atomic concept, P is an atomic role, and P^- is the *inverse* of P. Here B is called a *basic concept*, and C is a *general concept*. Likewise, we have *basic roles* R and *general roles* E. For completeness, we formally state the semantics of non-atomic concepts and roles:

- $(P^-)^{\mathcal{I}} = \{(o_2, o_1) \mid (o_1, o_2) \in P^{\mathcal{I}}\}$
- $(\exists R)^{\mathcal{I}} = \{o_1 \mid \text{there exists } o_2 \text{ and } (o_1, o_2) \in R^{\mathcal{I}}\}$
- $(\neg B)^{\mathcal{I}} = \Delta^{\mathcal{I}} \setminus B^{\mathcal{I}}$ and $(\neg R)^{\mathcal{I}} = \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \setminus R^{\mathcal{I}}$

Figure 3 gives the corresponding logical formulas. Note that when we write $R(X, Y)$, we mean $P(X, Y)$ if R is an atomic role P and $P(Y, X)$ if R is the inverse role P^-.

There are several different DL-Lite dialects, each allowing for different TBox axioms. In the core DL-Lite dialect, TBoxes are comprised of concept inclusions $B \sqsubseteq C$, where B is a basic concept and C a general concept. Observe that since only basic concepts are allowed on the left-hand side of inclusions, negation can only occur on the right-hand side.

[4] http://www.omg.org/uml.

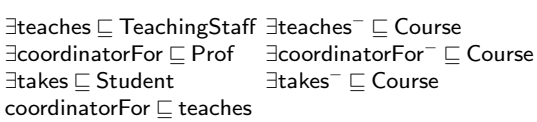


Fig. 4 Domain and range constraints and role inclusions in DL-Lite

To illustrate the expressive power of DL-Lite (core), we return to our running example of the university domain. We first remark that all of the atomic concept inclusions from Fig. 1 can be expressed in DL-Lite. Figure 4 displays DL-Lite axioms that constrain the *domain* and *range* of the roles teaches, takes, and coordinatorFor:

- if X teaches Y, then X is a member of teaching staff and Y is a course
- if X takes Y, then X is a student and Y is a course
- if X coordinatorFor Y, then X is a professor Y is a course

In DL-Lite, we can also express *disjointness constraints*, stating that two classes or properties have no elements in common, as well as *mandatory participation constraints*, requiring that elements of a certain class appear in the first or second components of a given binary relation. For example, the following axioms express that Student and AdminStaff are disjoint classes, that every professor must teach something, and that every course is taught by someone:

$$\mathsf{Student} \sqsubseteq \neg\mathsf{AdminStaff} \qquad \mathsf{Prof} \sqsubseteq \exists\mathsf{teaches} \qquad \mathsf{Course} \sqsubseteq \exists\mathsf{teaches}^-$$

We remark that if we replaced the inclusion $\mathsf{Student} \sqsubseteq \neg\mathsf{AdminStaff}$ by $\mathsf{Student} \sqsubseteq \neg\mathsf{Staff}$, then this would lead to an anomaly in the ontology. Indeed, using the atomic concept inclusions in Fig. 1, we would be able to infer that teaching assistants belong to the class Staff (from $\mathsf{TAssistant} \sqsubseteq \mathsf{TeachingStaff}$ and $\mathsf{TeachingStaff} \sqsubseteq \mathsf{Staff}$) as well as to its complement $\neg\mathsf{Staff}$ (using $\mathsf{TAssistant} \sqsubseteq \mathsf{GradStudent}$, $\mathsf{GradStudent} \sqsubseteq \mathsf{Student}$, and $\mathsf{Student} \sqsubseteq \neg\mathsf{Staff}$). This would mean that TAssistant must always be interpreted as the empty class, and thus that including even a single instance of TAssistant in the ABox would lead to an inconsistent KB.

Two other common dialects, DL-Lite$_\mathcal{R}$ and DL-Lite$_\mathcal{F}$, offer additional TBox axioms. The former allows for *role inclusions* of the form $R \sqsubseteq E$, while the latter authorizes *functionality axioms* of the form $(funct\ R)$, where R is a basic role, i.e., without negation. For example, the following two axioms are expressible in DL-Lite$_\mathcal{R}$ and DL-Lite$_\mathcal{F}$ respectively:

$$\mathsf{coordinatorFor} \sqsubseteq \mathsf{teaches} \qquad (funct\ \mathsf{coordinatorFor}^-)$$

The first axiom expresses that when X coordinatorFor Y, then X teaches Y, while the second one expresses that the role $\mathsf{coordinatorFor}^-$ is functional: if Y coordinatorFor X and Z coordinatorFor X then Y=Z.

A role inclusion $R \sqsubseteq E$ is satisfied in an interpretation $\mathcal{I}$ if $R^{\mathcal{I}} \subseteq E^{\mathcal{I}}$, and a functionality statement $(funct\ R)$ is satisfied in $\mathcal{I}$ if the binary relation $R^{\mathcal{I}}$ is a function, i.e. $(o, o_1) \in R^I$ and $(o, o_2) \in R^I$ implies that $o_1 = o_2$.

It has been shown in Calvanese et al. (2007) that in both DL-Lite$_{\mathcal{R}}$ and DL-Lite$_{\mathcal{F}}$, satisfiability, subsumption, and instance checking can all be performed in polynomial time (more precisely, these tasks are NLOGSPACE-complete). Rather surprisingly, however, if we consider the minimal DL that extends both DL-Lite$_{\mathcal{R}}$ and DL-Lite$_{\mathcal{F}}$, then satisfiability testing becomes EXPTIME-complete (Artale et al. 2009). Moreover, in both DL-Lite$_{\mathcal{R}}$ and DL-Lite$_{\mathcal{F}}$, it is possible to answer conjunctive queries in polynomial time in the size of the ABox by means of a reduction to the problem of answering first-order queries over relational databases, whereas no such reduction is possible if both role inclusions and functionality axioms are allowed (Calvanese et al. 2007). (Ontology-mediated conjunctive query answering and the technique of first-order query rewriting will be discussed in more detail in Sects. 3 and 4.)

The importance of the DL-Lite family of DLs is witnessed by the inclusion of the OWL 2 QL profile (W3C 2012b), based upon DL-Lite$_{\mathcal{R}}$, in the OWL 2 standard, which is specifically designed to be the ontology language of choice for applications involving querying of large amounts of data.

2.4 $\mathcal{ALC}$: The Prototypical Description Logic

The description logic $\mathcal{ALC}$ can be seen as the result of adding (full) *concept negation* to $\mathcal{EL}$. In $\mathcal{ALC}$, it is possible to construct the disjunction (or union) of two concepts $C_1 \sqcup C_2$ (which is just shorthand for $\neg(\neg C_1 \sqcap \neg C_2)$), qualified value restrictions (since $\forall R.C$ is equivalent to $\neg(\exists R.\neg C)$), and the bottom concept $\bot$ (corresponding to $\neg\top$, which is always interpreted as the empty set).

$\mathcal{ALC}$ is often considered to be the prototypical DL because it is a fragment of a natural first-order logic (allowing the standard Boolean operators plus restricted forms of universal and existential quantification) and because $\mathcal{ALC}$ concepts correspond precisely to the formulas expressible in the basic multi-modal logic K_n (Blackburn et al. 2006).

Returning to our university example, we first note since $\mathcal{ALC}$ extends both $\mathcal{FL}_0$ and $\mathcal{EL}$, all axioms from Sect. 2.2 can be expressed in $\mathcal{ALC}$. Additionally, the new constructors available in $\mathcal{ALC}$ allow us to express *disjointness* constraints, as in DL-Lite, and *covering* constraints, e.g., that every course is either an undergraduate course or a graduate course:

$$\mathsf{Student} \sqsubseteq \neg\mathsf{Prof} \qquad \mathsf{Course} \sqsubseteq \mathsf{UndergradCourse} \sqcup \mathsf{GradCourse}$$

For completeness, we formally specify the semantics of the new constructors:

- $\bot^{\mathcal{I}} = \emptyset$
- $\neg C^{\mathcal{I}} = \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
- $(C_1 \sqcup C_2)^{\mathcal{I}} = C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$

The problem of testing satisfiability of $\mathcal{ALC}$ KBs has been shown to be EXPTIME-complete (Schild 1991). The same result holds for subsumption testing and instance checking, which can be straightforwardly reduced to (un)satisfiability:

- $\mathcal{T} \models C \sqsubseteq D$ iff the KB $\langle \mathcal{T} \cup \{A_C \equiv C, A_{\neg D} \equiv \neg D\}, \{A_C(a), A_{\neg D}(a)\}\rangle$ is unsatisfiable
- $\mathcal{K} \models C(a)$ iff the KB $\mathcal{K} \cup \{A_{\neg C} \equiv \neg C\}, \{A_{\neg C}(a)\}$ is unsatisfiable.

To determine whether a given $\mathcal{ALC}$ KB is satisfiable, one can use *tableaux algorithms*, which work by exploring in an exhaustive manner all ways of constructing a model of the KB. If a (compact representation of a) model is found, the KB is satisfiable, and if all attempts fail, then one can conclude that the KB is unsatisfiable.

2.5 From $\mathcal{ALC}$ to $\mathcal{SHIQ}$ to $\mathcal{SROIQ}$: Highly Expressive DLs

The description logic $\mathcal{ALC}$ is the starting point for defining other (highly) expressive DLs. For example, the DL $\mathcal{SHIQ}$ (Horrocks et al. 1999) extends $\mathcal{ALC}$ with *inverse roles* (P^-, as in DL-Lite), (qualified) *cardinality restrictions* ($\geq n\ R.C$, $\leq n\ R.C$), *role inclusions* ($R \sqsubseteq R'$), and *transitive roles* (using transitivity axioms of the form $(Trans\ R)$), where R, R' can be either plain or inverse roles.

In our university example, we could use cardinality restrictions to express that every professor must teach at least 2 courses, and students that take at most 3 courses are part-time students:

$$\mathsf{Prof} \sqsubseteq \geq 2\mathsf{teaches.Course} \quad \mathsf{Student} \sqcap \leq 3\mathsf{takes.Course} \sqsubseteq \mathsf{PartTimeStudent}$$

The even more expressive $\mathcal{SROIQ}$ (Horrocks et al. 2006), which provides the logical underpinnings of OWL 2 (the latest version of OWL standard) (W3C 2012a), extends $\mathcal{SHIQ}$ with *nominals* ($\{a\}$), the universal role (u), and more *complex role axioms* of the forms $R \circ S \sqsubseteq R$ and $S \circ R \sqsubseteq R$ (where $\circ$ denotes role composition). It further allows for roles to be declared as reflexive, irreflexive, or antisymmetric, and for pairs of role to be declared disjoint.

The semantics of the new constructors is as follows:

- $(\geq\ nP.C)^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} \mid \sharp\{e \mid (d,e) \in P^{\mathcal{I}} \text{ and } e \in C^{\mathcal{I}}\} \geq n\}$
- $(\leq\ nP)^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} \mid \sharp\{e \mid (d,e) \in P^{\mathcal{I}} \text{ and } e \in C^{\mathcal{I}}\} \leq n\}$
- $\{a\}^{\mathcal{I}} = \{a^{\mathcal{I}}\}$
- $u^{\mathcal{I}} = \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$
- $(R \circ S)^{\mathcal{I}} = \{(d_1,d_3) \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \mid$ there exists $d_2 \in \Delta^{\mathcal{I}}$ such that $(d_1,d_2) \in R^{\mathcal{I}}$ and $(d_2,d_3) \in S^{\mathcal{I}}\}$

An axiom of the form $(Trans\ R)$ is satisfied in $\mathcal{I}$ if $R^{\mathcal{I}}$ is a transitive relation, i.e., $(d_1,d_2) \in R^{\mathcal{I}}$ and $(d_2,d_3) \in R^{\mathcal{I}}$ implies $(d_1,d_3) \in R^{\mathcal{I}}$. TBox axioms declaring roles to be reflexive, irreflexive, or antisymmetric are handled analogously.

In the DL $\mathcal{SHIQ}$, the standard reasoning tasks (satisfiability, subsumption, and instance checking) are all EXPTIME-complete, and thus of the same complexity as in $\mathcal{ALC}$. For the DL $\mathcal{SHOIQ}$ obtained by adding nominals to $\mathcal{SHIQ}$, the complexity rises to NEXPTIME-complete, and for $\mathcal{SRIQ}$, which extends $\mathcal{SHIQ}$ with complex role inclusions and additional types of axioms concerning roles, the problem becomes 2EXPTIME-complete. If we move all the way up to $\mathcal{SROIQ}$, then reasoning becomes even more difficult (2NEXPTIME-complete).

The preceding complexity results show that automated reasoning with (highly) expressive DLs like $\mathcal{ALC}$, $\mathcal{SHIQ}$, and $\mathcal{SROIQ}$ may require (doubly) exponential time in the worst case. However, in practice, modern DL reasoners,[5] mainly employing highly optimized tableaux algorithms, demonstrate acceptable performance for reasonably-sized ontologies. The reason is that the type of ontological constraints that are needed to model even complex real-world applications do not give rise to the pathological combinations of constructors that are required for establishing the negative complexity results.

Finally, several proposals have been made to overcome a limitation of description logics, which is the tree-shaped structure of terminological descriptions, in particular by combining them with Datalog rules. These proposals impose restrictions on the interaction between DL axioms and datalog rules, as early work has shown undecidability of standard reasoning if no restriction was made (Levy and Rousset 1998). The existential rules presented in Sect. 4 can be seen as another way of overcoming the tree-shaped structure limitation of description logics.

3 Conceptual Graphs

Conceptual graphs (Sowa 1976, 1984) are mainly rooted in semantic networks, natural language processing and Peirce's existential graphs, a diagrammatical system of logic alternative to predicate logic. They have been studied along different directions. One research line consists in developing conceptual graphs as a graphical interface to first-order logic. Another research line follows the existential graph approach: conceptual graphs are then seen as diagrams, rather than graphs in the graph-theoretic meaning, and inferences are based on diagrammatic operations that do not aim to be automated (see, in particular, Dau 2003). A third research line, which is the one presented in this chapter, develops conceptual graphs as a knowledge representation and reasoning formalism, equipped with its own reasoning mechanisms. This formalism is both graph- and logic-based: on the one hand, the basic objects are labelled graphs and reasoning is based on graph operations, with graph homomorphism at the core; on the other hand, these objects have a logical semantics and reasoning mechanisms are sound and complete with respect to this semantics. This approach to conceptual graphs is similar to the description logic approach in the sense that it defines and studies a family of formalisms that offer different tradeoffs between expressiv-

[5] See http://owl.cs.manchester.ac.uk/tools/list-of-reasoners/ for an up-to-date list of DL reasoners.

ity and complexity of reasoning. However, we will see in Sect. 3.5 that the logical fragments studied in both families are quite different. The interested reader will find an in-depth presentation of theoretical and algorithmic results on conceptual graphs in (Chein and Mugnier 2009). All aspects presented here are implemented in the software tools CoGUI[6] and CoGITaNT.[7]

3.1 The Kernel: Basic Conceptual Graphs

A basic conceptual graph (BG) defines entities and relationships among these entities. Hence, it is a bipartite graph: one class of nodes, called *concept* nodes, represents entities, and the other class, called *relation* nodes, represent relationships among these entities. Nodes are labelled according to a vocabulary, which contains a set of concept types and a set of relation symbols, with both sets being partially ordered by specialisation. This vocabulary can be seen as a lightweight ontology, which can be further enriched by rules and constraints in more complex conceptual graph fragments.

3.1.1 Syntax

A *vocabulary*, also called a *support*, is a triple (T_C, T_R, I) where:

- T_C is a finite set of *concept types*, partially ordered by $\leq$, and provided with a greatest element, denoted by $\top$;
- T_R is a finite set of *relation symbols* (or simply relations) of any arity, partially ordered by $\leq$, such that only relations with the same arity are comparable;
- I is a possibly infinite set of elements called *individual markers*; furthermore, the symbol $*$ denotes the *generic marker*, with $* \notin I$. The set of all markers $I \cup \{*\}$ is partially ordered by $\leq$ as follows: for all $m \in I$, $m \leq *$, and elements in I are pairwise incomparable.
- T_C, T_R and I are pairwise disjoint sets.

The partial orders on T_C and T_R encode a specialisation relation, i.e., $t' \leq t$ means that "t' is a specialisation of t". Figure 5 pictures a set of concept types, which correspond to the set of DL inclusions from Fig. 1 except for the part in italics. Each individual marker refers to a specific and distinct entity (i.e., the unique name assumption is made) and the generic marker refers to an unspecified entity.

A *basic conceptual graph* (BG) $G = (C_G, R_G, E_G, l_G)$ defined over a support $S = (T_C, T_R, I)$ is a finite, labelled, undirected and bipartite multigraph (i.e., there may be several edges between two nodes), where C_G is the set of concept nodes, R_G

[6] http://www.lirmm.fr/cogui

[7] http://cogitant.sourceforge.net

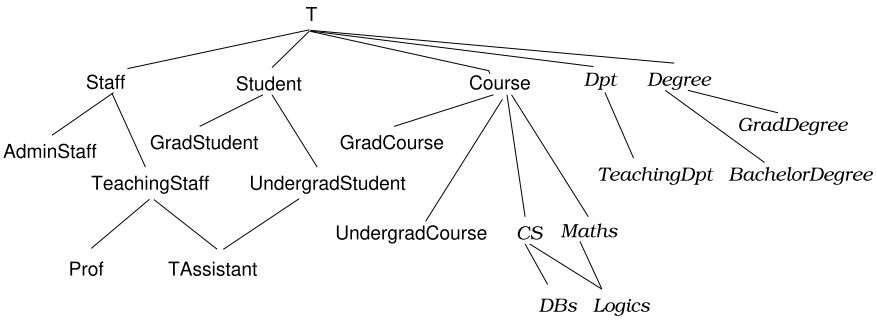
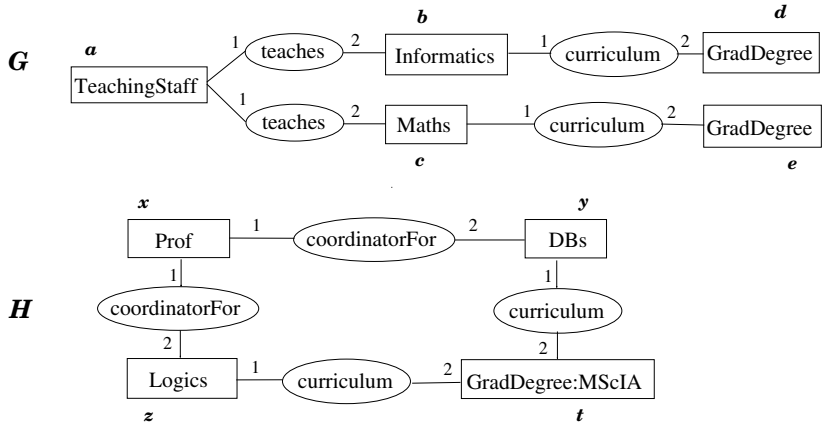


Fig. 5 A set of concept types

is the set of relation nodes, E_G is the multiset of edges, and l_G is a labelling function of the nodes and edges that satisfies the following conditions:

- each concept node c is labelled by a concept type t_c and a marker m such that $(t_c, m) \in T_C \times (I \cup \{*\})$; if $m = *$, c is called *generic*, otherwise it is *individual*;
- each relation node r is labelled by a relation $t_r \in T_R$ and the number of edges incident to r is equal to the arity of t_r; these edges are labelled from 1 to the arity of t_r; we denote by $(c_1 \ldots c_k)$ the list of arguments of r, where c_j denotes the extremity of the jth edge incident to r.

Note that a BG is not necessarily connected. By convention, concept nodes are pictured as rectangles and relation nodes as ovals. For instance, the BG H pictured in Fig. 6 may represent the following knowledge "there is a professor coordinator for a course on databases and a course on logics in the graduate degree MSc IA". A BG can also be seen as a hypergraph, with the relations being encoded by hyperedges.



The generic marker has been omitted in concept node labels

Fig. 6 Basic Conceptual Graphs

Then the graph view of a BG corresponds to the incidence bipartite graph of this hypergraph.

The notion of a support can be extended to allow for multi-instantiation. Then a concept node is labelled by a set of concept types, called a *conjunctive type*, instead of a single concept type. For instance, a course can be both a course in mathematics and a course in computer science, which is denoted by the conjunctive type $\{Maths, CS\}$. The set of concept types T_C is then defined in intension by a partially ordered set of primitive types, and its elements are all the conjunctive types that can be built with from primitive types. The partial order on conjunctive types is the natural extension of the order on primitives types: given conjunctive types t_1 and t_2, $t_2 \leq t_1$ if for all primitive type t_{1_i} in t_1, there is a primitive type t_{2_j} in t_2 with $t_{2_j} \leq t_{1_i}$. For instance, $\{Logics\} \leq \{CS, Maths\}$ (note that $\{CS, Maths\} \not\leq \{Logics\}$).

3.1.2 Semantics

This basic formalism is provided with a semantics in first-order logic by a translation denoted by Φ. Concept types are translated into unary predicates, relation symbols into predicates with the same arity and individual markers into constants (for simplicity, the same names are used for the elements of the support and their translation). To a support S is assigned a set of formulas $\Phi(S)$ that translates the partial orders on T_C and T_R, i.e., if $t_2 \leq t_1$, one has the formula $\forall x_1 \dots x_k \ (t_2(x_1 \dots x_k) \rightarrow t_1(x_1 \dots x_k))$, where k is the arity of predicates t_1 and t_2.

A BG is translated into a formula $\Phi(G)$ built as follows: to each concept node is assigned a term, which is a new variable if its marker is generic, otherwise the constant assigned to its individual marker; to each relation (resp. concept) node is assigned an atom $t(e_1, \dots, e_k)$ where t is the predicate assigned to its label (resp. concept type) and $(e_1, \dots, e_k)$ is the list of terms assigned to its arguments (resp. the term assigned to the concept node); $\Phi(G)$ is then the existential closure of the conjunction of these atoms. For instance, for H in Fig. 6: $\Phi(H) = \exists x \exists y \exists z (Prof(x) \wedge DBs(y) \wedge Logics(z) \wedge coordinatorFor(x, y) \wedge coordinatorFor(x, z) \wedge curriculum(y, MScIA) \wedge curriculum(z, MScIA) \wedge GradDegree(MScIA))$.

The BG fragment is equivalent to the existential, positive and conjunctive fragment of first-order logic (without functional symbols except for constants). Indeed, a polynomial translation of the support and BGs, which preserves logical entailment, allows one to obtain a "flat" support S' for which $\Phi(S') = \emptyset$.

The fundamental notion for reasoning on BGs is a *homomorphism* (often called "projection" in the conceptual graph community). A homomorphism from a BG G to a BG H is a mapping from $C_G \cup R_G$ to $C_H \cup R_H$ that preserves the node bipartition and the edges, and may specialize node labels, i.e.,

- for all relation $r \in R_G$ with arguments $(c_1 \dots c_k)$, $h(r)$ has arguments $(h(c_1) \dots h(c_k))$ (equivalently: for all edge rc in G, there is an edge $h(r)h(c)$ with the same label in H) ;

- for all node $x \in C_G \cup R_G, l_H(h(x)) \leq l_G(x)$ (for concept nodes one considers the product order on $T_C \times (I \cup \{\star\})$, i.e., $(t, m) \leq (t', m')$ if $t \leq t'$ and $m \leq m'$).

Consider the graphs in Fig. 6 and assume that *coordinatorFor* $\leq$ *teaches* is the only comparable pair of distinct relations: there are two homomorphisms from G to H. The first one maps concept nodes in this way: $a \mapsto x$, $b \mapsto y$, $c \mapsto z$, $d \mapsto t$, $e \mapsto t$; each relation node *teaches* is mapped to a relation node *coordinatorFor* and each relation node *curriculum* is mapped to a node with the same label. Note that both entities of type *GradDegree* are mapped to a single entity, which is identified as "the MSc AI". The second homomorphism maps concept nodes in this way: $a \mapsto x$, $b \mapsto z$, $c \mapsto z$, $d \mapsto t$, $e \mapsto t$. This homomorphism uses the fact that *Logics* is a specialisation of both *Maths* and *CS*, which allows one to map b and c to z.

BG-homomorphism induces a preorder on BGs, called the "specialisation / generalisation" relation: in the following, we note $H \leq G$ (H is a specialisation of G) if there is a homomorphism from G to H. This relation is sound and complete with respect to logical entailment on the formulas assigned to the BGs (also using the formulas assigned to the support), i.e., for all BGs G and H on a support S, $H \leq G$ if and only if $\Phi(S), \Phi(H) \models \Phi(G)$. Completeness is up to a normality condition for H: an individual marker has to occur at most once in H (in other words, two distinct nodes cannot refer to the same identified entity). The following problem, called **BG-Homomorphism** is thus the fundamental problem on BGs: given two BGs G and H, is there a homomorphism from G to H? This problem is NP-complete in general, but belongs to PTime when the source graph (i.e., G) is an acyclic graph (or an acyclic hypergraph, this latter notion being more general than the former), and more generally if it has a bounded treewidth (or hypertreewidth).

Homomorphism being a fundamental notion in the study of relational structures, it is not surprising that BG-Homomorphism is strongly equivalent to other fundamental problems in AI and databases, which allows one to import algorithmic techniques from one domain to another. The logical translation of a BG is the same as a (Boolean) conjunctive query (CQ) in databases. The problems of evaluating a conjunctive query (e.g., given a CQ q and a relational database instance D, does D contain an answer to q?) or determining if a conjunctive query is contained in another (given two CQs q_1 and q_2, is the set of answers to q_1 included in the set of answers to q_2 for any database instance?) are essentially the same as BG-Homomorphism. The same remark holds for the basic constraint satisfaction problem (CSP): given a constraint network (in which constraints are given in extension), does this network have a solution? (see chapter "Constraint Reasoning" of Volume 2).

Finally, let us consider the ontology-mediated query answering problem in the conceptual graph framework, where the knowledge base is composed of a support (the ontology) and of BGs (the facts), and the query is itself a BG. Checking if the query is entailed by the KB is NP-complete in combined complexity (since it amounts to a BG-homomorphism test) and polynomial in data complexity.[8]

[8]For query answering problems, the distinction between combined and data complexities is often made: data complexity is the complexity with respect to the size of the data (here the fact base),

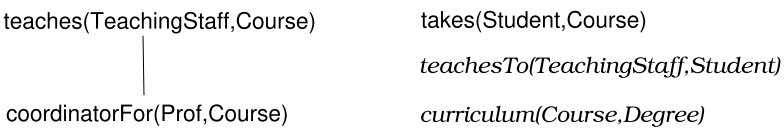


Fig. 7 Relations and their signatures

3.2 Simple Extensions of the Support

Two simple extensions of the support are often considered, namely relation signatures and concept type incompatibility. A relation signature specifies the maximal type of each of its arguments. Formally, one adds to the support a mapping σ that assigns to each relation r with arity k a signature $\sigma(r) \in (T_C)^k$. Let $\sigma_i(r)$ denote the ith element of $\sigma(r)$; the formula assigned to $\sigma(r)$ is:

$$\forall x_1 \dots x_k (r(x_1 \dots x_k) \rightarrow \sigma_1(r)(x_1) \wedge \dots \wedge \sigma_k(r)(x_k))$$

Figure 7 shows a partially ordered set of relations with their signature that corresponds to the set of DL inclusions from Fig. 4, except for the part in italics.

Relation signatures must be covariant with respect to the partial orders on concept types and relations: for all relations r_1 et r_2 with arity k, if $r_1 \leq r_2$ then $\sigma(r_1) \leq \sigma(r_2)$, i.e., for all i, $\sigma_i(r_1) \leq \sigma_i(r_2)$. This covariance condition translates the fact that when a relation is specialized into another, the maximal type of each argument can be specialized as well, but not generalized. For instance, if the relation *teaches* links an entity of type *TeachingStaff* to an entity of type *Course*, its specialisation into the relation *coordinatorFor* may enforce that the first argument is of type *Prof*, which is a specialisation of *TeachingStaff*.

The support added with relation signatures can be seen as a generalisation of the ontological part of RDFS (i.e., the schema) with relations of any arity (see Baget et al. 2010 for translations between RDFS and basic conceptual graphs).

When multi-instantiation is allowed, i.e., when conjunctive concept types are considered, it is useful to express incompatibility between concept types. This can be achieved by stating that some conjunctive types are forbidden. The logical formula assigned to a *banned type* $\{t_1, t_2\}$ is the following:

$$\forall x \neg (t_1(x) \wedge t_2(x))$$

The set of banned types is said to be compatible with the set of primitive types if no primitive type is a specialization of a banned type. For instance, the banned type {*Student, Staff*} (which corresponds to the DL inclusion $Student \sqsubseteq \neg Staff$) is not compatible with $Student \leq Staff$, *a fortiori* with $GradStudent \leq Staff$. A BG

while combined complexity considers all components of the problem (here, the knowledge base and the query).

complies with the set of banned types if no node is labelled by a concept type that specializes a banned type. Note that the logical translation $\Phi(S)$ of a support S added with banned types is always consistent. However, for a BG F on S, $\Phi(S) \cup \Phi(F)$ may be inconsistent.

3.3 Conceptual Graph Rules

Rules of the form " if *premise* then *conclusion*" are an essential knowledge construct in AI. They represent implicit knowledge that can be made explicit by applying them to factual knowledge. A *basic graph rule* is a pair $R = (P(c_{1_1} \dots c_{1_k}), C(c_{2_1} \dots c_{2_k}))$, where $k \geq 0$, P and C are BGs, and the c_{1_i} (respectively c_{2_i}) are distinct generic concept nodes from P (respectively C) called the *frontier nodes* of the rule. In Fig. 8 the bijection between the frontier nodes of the premise and of the conclusion is depicted by dotted lines; the blue nodes form the conclusion of the rule. This rule represents the following knowledge: "if a student X takes a course Y then there is a teaching staff member Z who teaches Y and teaches to X".

The logical translation of a BG-rule $R = (P(c_{1_1} \dots c_{1_k}), C(c_{2_1} \dots c_{2_k}))$ is the formula $\Phi(R) = \forall x_1 \dots x_k \ (\Phi'(P) \rightarrow \Phi'(C))$, in which the same variable x_i is assigned to frontier nodes c_{1_i} and c_{2_i}, and $\Phi'(P)$ (resp. $\Phi'(C)$) is obtained from $\Phi(P)$ (resp. $\Phi(C)$) by leaving variables $x_1 \dots x_k$ free. Equivalently, all the variables in the premise of the rule can be universally quantified, in which case their scope is the whole formula. The logical translation of a rule is thus exactly an *existential rule*, as defined in the next section. For instance, the logical translation of the rule R from Fig. 8 is $\Phi(R) = \forall x \forall y((Student(x) \wedge Course(y) \wedge takes(x, y)) \rightarrow \exists z(TeachingStaff(z) \wedge teaches(z, y) \wedge teachesTo(z, x)))$.

BG-rules are provided with forward and backward chaining mechanisms that proceed directly on their graphical form. A BG-rule R is applicable to a BG F if there is a homomorphism h from its premise to F; applying R to F according to

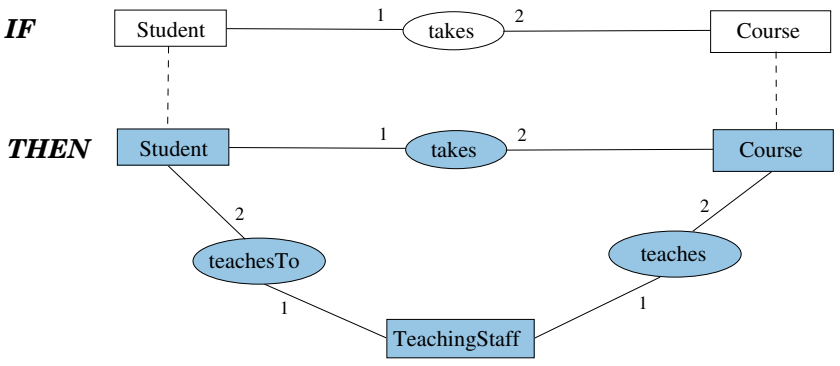


Fig. 8 Conceptual Graph Rule

h consists of adding C to F, then merging each frontier node c_{2_i} from C with the node $h(c_{1_i})$ from F.[9] Rule application is the basis of a sound and complete forward chaining mechanism, i.e., given a knowledge base $\mathcal{K} = (S, F, \mathcal{R})$, where S is the support, F is the fact base (remember that a BG needs not to be connected) and $\mathcal{R}$ is the set of rules, and a BG Q (which can be seen as a query), $\Phi(\mathcal{K}) \models \Phi(Q)$ if and only if there is a sequence of applications of rules in $\mathcal{R}$ leading from F to a BG F' such that $F' \leq Q$.

The backward chaining mechanism relies on a specific unification operation (between two subgraphs, respectively of a rule conclusion and of the current BG query), which exploits the complex structure of rule conclusions induced by non-frontier concept nodes (see the existential variables in existential rules). Hence, instead of processing a goal atom by atom as backward chaining *a la* Prolog would do, entire subgraphs are unified at once. This mechanism is also sound and complete.

Note that the partial orders on concept types and relations can be encoded by BG-rules. Indeed, $t_1 \leq t_2$ is logically translated into the logical rule $\forall x_1 \ldots x_k (t_1(x_1 \ldots x_k) \rightarrow t_2(x_1 \ldots x_k))$, where k is the arity of the associated predicates. However, the fact that the partial orders are intrinsically taken into account in BG-homomorphism (which allows one to compare concept types or relations in constant time, or almost constant time, depending on the order encoding) leads to better algorithmic efficiency.

BG-rules are able to simulate the behavior of a Turing machine, hence they provide a model of computation. Therefore, the associated entailment problems are undecidable. However, many decidable cases obtained by syntactic restrictions on rules, or sets of rules, have been defined, mostly in the framework of existential rules (see Sect. 4.3).

3.4 Conceptual Graph Constraints

A BG-constraint has the same shape as a BG-rule. It can be positive or negative, depending on whether it expresses an obligation or a prohibition. A positive constraint (P, C) expresses knowledge of the form "whenever P is true, C must also be true". It is satisfied by a BG if every homomorphism from P to F can be extended to a homomorphism from C to F (i.e., given h the considered homomorphism from P to F, there is a homomorphism h' from C to F such that $h'(c_{2_i}) = h(c_{1_i})$ for all frontier nodes). A negative constraint (P, C) expresses knowledge of the form "whenever P is true, C must not be true". It is satisfied by a BG if no homomorphism from P to F can be extended to a homomorphism from C to F. A negative constraint can

[9] If c_{1_i} and c_{2_i} have the same concept type, the obtained node is labelled by the same label as $h(c_{1_i})$; if the type of c_{2_i} is strictly more specific than the type of c_{1_i}, it may happen that the labels of $h(c_{1_i})$ and c_{2_i} are incompatible (with respect to banned types), which points to an inconsistency in the knowledge base; otherwise, the label of the obtained node is the greatest lower bound of both labels: the obtained type is the conjunction of the types of $h(c_{1_i})$ and c_{2_i} and the obtained marker is the smallest of both markers.

also be represented as a BG, let C^-, obtained by merging P and C on their frontier nodes (with each c_{1_i} being merged with c_{2_i}); then C^- is satisfied by F if there is no homomorphism from C^- to F.

For instance, the constraint that a student cannot belong to the administrative staff can be expressed by the formula $\forall x(Student(x) \rightarrow \neg AdminStaff(x))$, which corresponds to the form (P, C), or by the equivalent formula $\neg\exists x(Student(x) \wedge AdminStaff(x))$, which amounts to forbid a BG. Note that the extensions to the support introduced in Sect. 3.2 can be encoded by constraints, namely relation signatures by positive constraints and banned types by negative constraints. Other frequent forms of constraints in ontologies are cardinality constraints: positive constraints allow one to express the condition "at least 1" (like "every professor must teach at least one undergraduate course") and negative constraints the condition "at most 0" (like "no teaching assistant can be coordinator for a course").

Negative constraints can actually be seen as particular positive constraints (with C restricted to a concept node with banned type). Positive constraints strictly generalize negative constraints, in the sense that the associated consistency problems are not in the same complexity class: the problem of determining whether a given BG satisfies a given constraint is co-NP-complete if the constraint is negative, and Π_P^2-complete otherwise.

Finally, equality is represented in conceptual graphs by so-called "co-reference links" which pairwise connect concept nodes that refer to the same entity. While co-reference links do not increase the expressivity of BG (though their use may be interesting for visualisation purposes), they do increase the expressivity of BG-rules, allowing in particular to express functional dependencies.

3.5 *Relationships with Description Logics*

Description logics and conceptual graphs are both rooted in semantic networks. They both remedy two criticisms on these common ancestors, i.e., the lack of distinction between factual and ontological knowledge, and the lack of formal semantics. Due to these common properties, their relationships have often been questioned.

Provided that relations are restricted to binary relations, a support can be seen as a simple TBox composed of atomic concept and atomic role inclusions. Relation signatures then correspond to the notions of domain and range, and banned concept types to class disjointness constraints. On the other hand, an ABox can be seen as a particular BG without generic concept nodes.

With the aim of characterizing the intersection of BGs (on a simple support) and DLs, two equivalent fragments were identified in Baader et al. (1999b). On the CG side, we obtain rooted BG trees with binary relations. On the DL side, we obtain the DL $\mathcal{ELIRO}_1$ (a DL specially tailored for the comparison), in which the constructors are $\exists R.C$ (existential restriction), $C \sqcap D$ (concept intersection), R^- (role inverse), $R \sqcap R'$ (role intersection) and $\{i\}$ (unary one-of, where i is an individual, which allows one to integrate specific individuals in concept expressions). It is to be

noticed that this comparison with conceptual graphs was one of the sources of the $\mathcal{EL}$ family, in which homomorphism is a central notion (Baader et al. 1999a).

In this intersection, both formalisms lose some natural features: on the CG side, relations of any arity and unrestricted structure, in particular cycles on generic concept nodes, while, on the DL side, the variety of constructors.

Other results support the claim that both formalisms are quite "orthogonal". On the one hand, it is known that even the most expressive DLs cannot express the whole existential positive conjunctive fragment of first-order logic (Borgida 1996). On the other hand, BG-homomorphism cannot handle negation in a logically complete way, even when restricted to atomic negation on primitive concept types.

More relationships between DLs and CGs can be found if we turn our attention to richer fragments of conceptual graphs including some classes of BG-rules and negative BG-constraints on the one hand, and to the ontology-mediated query answering problem on the other hand. Indeed, description logics historically focused on reasoning about the ontology (i.e., the TBox). The instance checking problem can only be seen as a very specific querying problem, which asks if a given individual belongs to a given concept. To handle conjunctive queries, new description logics were considered more recently (see Sects. 2.2 and 2.3), such as the DL-Lite family, specifically designed to query data, the $\mathcal{EL}$ family, and more generally Horn description logics. These DLs can be seen as specific fragments of the existential rule framework (see the next section), which in turn can be seen as the logical translation of the conceptual graph framework described in this section.

4 Existential Rules

As already mentioned, the increasing volume of complex and heterogeneous data has spurred an intense research effort on the issue of ontology-mediated query answering in recent years. This work has deeply modified the description logic field and led to the emergence of new dialects and algorithmic techniques (Sect. 2.3). Meanwhile, the framework of existential rules has been developed to address this issue. The existential rule framework has a double origin: on the one hand it corresponds to the logical translation of the conceptual graph fragment (BGs, rules and negative constraints) presented in the previous section (Baget et al. 2011a), on the other hand it has been proposed as an extension to Datalog, the language of deductive databases, under the name Datalog± (Calì et al. 2009).

In the relational database field, Datalog was originally designed to provide first-order queries (or equivalently, core SQL queries) with recursivity (Abiteboul et al. 1995). In its plain version (i.e., without negation nor disjunction), a Datalog query can be seen as a set of rules, which are closed formulas of the form $\forall x_1 \ldots x_n\ (B \rightarrow H)$, where B and H, respectively called the body and the head of the rule (according to the logic programming terminology), are conjunctions of atoms; moreover, these rules satisfy the constraint of being "range-restricted", i.e., all the variables that occur in the head of a rule must also occur in its body. Hence, a plain Datalog rule is

logically translated into a Horn clause without function symbols. These rules could be used as a means of encoding implicit background knowledge. However, they lack a property considered crucial for representing ontological knowledge, which is the ability to reason on open domains. Indeed, when the open-world assumption is made, it cannot be assumed that the only existing entities are those encoded in the data. Hence, one should be able to infer knowledge on unknown individuals, which may (or may not) be equal to entities from the data. These considerations motivated the extension of Datalog rules with existentially quantified variables in rule heads.

4.1 The Existential Rule Framework

Formally, an existential rule is of the form $R = \forall \mathbf{x} \forall \mathbf{y}\, (B[\mathbf{x}, \mathbf{y}] \rightarrow \exists \mathbf{z}\, H[\mathbf{y}, \mathbf{z}])$, where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are sets of variables, and B, H are conjunctions of atoms, also denoted by $body(R)$ and $head(R)$. The *frontier* of R is the set of variables shared between the body and the head of R, i.e., $\mathbf{y}$. The *existential variables* in R are the existentially quantified variables, i.e., $\mathbf{z}$.

We now consider knowledge bases of the form $\mathcal{K} = (F, \mathcal{R})$, where F is a fact base[10] and $\mathcal{R}$ is a set of (pure) existential rules.

The logical translation of the BG-rules seen in the preceding section yields existential rules. For instance, the formula assigned to the BG-rule R from Fig. 8 is $\Phi(R) = \forall x \forall y((Student(x) \wedge Course(y) \wedge takes(x, y)) \rightarrow \exists z(TeachingStaff(z) \wedge teaches(z, y) \wedge teachesTo(z, x)))$, where the frontier is $\{x, y\}$ (note that here all the variables from the body are frontier variables) and the only existential variable is z. Any conceptual graph KB of the form $\mathcal{K} = (S, F, \mathcal{R})$ can be translated into a logically equivalent existential rule KB of the form $\mathcal{K}' = (F', \mathcal{R}')$, and reciprocally. In the following, we omit quantifiers in rules as there is no ambiguity.

Beside these "pure" existential rules, two other kinds of rules are generally considered in the framework: *negative constraints*, which are existential rules with a head restricted to $\bot$, and *equality rules*, which are existential rules with a head restricted to an equality of the form $e_1 = e_2$, where the e_i are variables from the body or constants. These rules also correspond to constructs in the conceptual graph framework, namely negative BG-constraints and BG-rules with a conclusion restricted to two co-referent concept nodes.

Existential rules and classical description logics like $\mathcal{ALC}$ are incomparable with respect to expressivity. For instance, the $\mathcal{ALC}$ inclusions $\exists$coordinatorFor.Course $\sqsubseteq$ Prof or the $\mathcal{SHIQ}$ transitivity axiom $(Trans\ \ P)$ can be expressed by existential rules, but not the $\mathcal{ALC}$ inclusion Course $\sqsubseteq$

[10] A fact is usually defined as a ground atom. However, in the existential rule setting, a more general notion of a fact can be considered, where a fact is an existentially closed conjunction of atoms, which is in line with the view of a fact as a rule with an empty body. This generalized notion allows one to encode unknown values in a natural way.

$\mathsf{UndergradCourse} \sqcup \mathsf{GradCourse}$ which would require a disjunctive head, and the existential rule R from the previous example cannot be expressed in a description logic.

On the other hand, existential rules are strictly more expressive than so-called Horn description logics, which can be seen as DLs whose logical translation yield existential rules (in other words, the skolemisation of their logical translation yields Horn clauses with possibly functional symbols). The lightweight description logics $\mathcal{EL}$ and the DL-Lite dialects seen in Sect. 2 are examples of Horn description logics. Existential rules can be seen as overcoming two limitations of (Horn) description logics: first, predicates of any arity are allowed; second, there is no restriction on the atoms composing the body and the head of a rule, which allows one to describe complex relationships between entities (see e.g., the above rule R), whereas description logics are essentially limited to "acyclic" structures.

4.2 Relationships with Database Theory

An important connection with relational database theory has to be pointed out. Indeed, existential rules have the same logical form as Tuple-Generating Dependencies (TGDs), a high-level class of database constraints that generalize many constraints of practical database systems (and correspond to the CG positive constraints from the preceding section). Negative constraints are also considered in databases and equality rules (with equality between two variables) have the same logical form as the database Equality Generating Dependencies (EGDs), which generalize constraints on keys (see e.g., Abiteboul et al. 1995). Note that, despite their syntactic correspondence, database constructs and rules have different roles: TGDs/EGDs act as constraints to check the consistency of a database instance, whereas rules act as ontological knowledge to generate new data. However, in the database setting, it is possible to repair constraint violations with respect to TGDs/EGDs by applying them in a forward chaining manner as if they were rules. This process, known as the *chase*, is considered as one of the fundamental tools in database theory. The similarities between the studied objects explain that many theoretical results obtained in one domain are actually of interest to the other. In particular, it has long been shown that the entailment of an atom from a set of TGDs (hence a set of pure existential rules) and a database instance is an undecidable problem when no restriction is made.

An existential rule R can be applied to a fact base F if there is a homomorphism h from $body(R)$ to F, i.e., a substitution h of the variables in $body(R)$ by terms in F such that $h(body(R)) \subseteq F$ (both seen as sets of atoms). Applying R to F according to h consists in adding $h(head(R))$ to F, where $h(head(R))$ is obtained from H by substituting each frontier variable x by $h(x)$ and safely renaming existential variables by fresh existential variables. The saturation of the fact base consists in iteratively applying rules on it until no rule application is possible. This process may of course not terminate since entailment is undecidable. Several forward chaining (or chase) variants have been defined, which differ in how they deal with the possible

redundancies introduced by existential variables. It is well known that the (possibly infinite) saturation obtained by any of these chase variants forms a *universal model* of the knowledge base, i.e., a model that can be mapped by homomorphism to any other model of the KB. Hence, a universal model acts as a representative of all models of the KB, sufficient to decide conjunctive query entailment from the KB.

4.3 Decidability Results

Interest in the existential rule framework gave rise to fruitful work on finding classes of existential rules for which (conjunctive) query answering is decidable. A wide range of rule classes offering various expressivity/tractability tradeoffs is now known (see e.g., Mugnier 2011; Gottlob et al. 2012; Thomazo 2013; Mugnier and Thomazo 2014 for syntheses). Most of these classes can be understood according to abstract properties that underlie decidability:

1. The set of rules $\mathcal{R}$ ensures that any KB $\mathcal{K} = (F, \mathcal{R})$ has a finite universal model. In other words, some chase variant is guaranteed to halt on any fact base. Hence, for any (Boolean) CQ q, $\mathcal{K} \models q$ if and only if $F^* \models q$, where F^* is the saturation of F. Such sets of rules are called *finite expansion sets (fes)* (Baget et al. 2011a).
2. The set of rules $\mathcal{R}$ ensures that any (Boolean) CQ q can be rewritten using the rules into a (finite) union of conjunctive queries Q such that for any KB $\mathcal{K} = (F, \mathcal{R})$ holds that $\mathcal{K} \models q$ if and only if $F \models Q$. Such sets of rules are called UCQ-rewritable or *finite unification sets (fus)* (Baget et al. 2011a). More general forms of rewritings can be considered, such as first-order queries, which may produce a more succinct rewriting, or Datalog queries, which may provide a finite rewriting when there is no finite rewriting as a first-order query (see Gottlob and Schwentick 2012; Bienvenu et al. 2018 among others). It is known that UCQ-rewritability and first-order rewritability are actually equivalent properties (e.g., Gottlob et al. 2014).
3. The existence of a finite universal model may not be guaranteed, but the set of rules $\mathcal{R}$ ensures that the saturation of any KB $\mathcal{K} = (F, \mathcal{R})$, seen as a graph, has a bounded treewidth. This allows for finite encodings of infinite saturations. Such sets of rules are called *(greedy) bounded-treewidth sets ((g)bts)* (Baget et al. 2011a, b; Thomazo 2013).

The two first families of rules clearly enable one to come back to a classical database query answering problem: either the knowledge that can be entailed by the rules is encoded in the facts, or the relevant part of the rules is encoded in the query. In the third case, querying the finite encoding is more involved. Deciding whether a given set of rules satisfies one of these three abstract properties is undecidable, however each abstract property admits some "concrete" cases defined by recognizable syntactic criteria. Generalisations or combinations of these properties have been defined, however their presentation is outside the scope of this chapter.

Table 1 Fundamental classes of existential rules with polynomial-time data complexity

Rule class	Data complexity
Datalog	PTime-c (Dantsin et al. 2001)
Weakly-acyclic	PTime-c (Dantsin et al. 2001) (LB) (Fagin et al. 2005) (UB)
aGRD	$AC_0(1)$
Linear	AC_0 (Calì et al. 2009) (1)
Sticky	AC_0 (Calì et al. 2010) (1)
Guarded	PTime-c (Calì et al. 2009)
Frontier-guarded	PTime-c (Baget et al. 2011b)
Frontier-1	PTime-c (Baget et al. 2011b)

Table 1 presents the main currently known concrete rule classes for which ontology-mediated conjunctive query answering has polynomial time data complexity. We chose to present the simplest classes, in order to highlight the fundamental ideas, even if most of these classes admit generalisations that often keep the same data complexity. The existence of a finite universal model (*fes* property) is ensured by some acyclicity conditions that prevent infinite creation of new variables during the chase. Such classes include range-restricted rules (i.e., Datalog rules), *weakly-acyclic* rules, and *aGRD* rules. These two last classes are both defined by an acyclicity condition on a directed graph, which encodes variables sharing between positions in predicates in the first case, and dependencies between rules in the second case. In the first graph, called *position (dependency) graph* (Fagin et al. 2005), the nodes represent all positions in predicates occurring in rules, i.e., the node *(p,i)* represents the position i in some predicate p. Then, for each rule R and each variable x in $body(R)$ occurring in position (p, i), edges with origin (p, i) are built as follows: if x is a frontier variable, there is an edge from (p, i) to each position of x in $head(R)$; furthermore, for each existential variable y in $head(R)$ occurring in position (q, j), there is a special edge from (p, i) to (q, j). A set of rules is said to be *weakly acyclic* if its position graph has no circuit passing through a special edge. Intuitively, this condition ensures that the introduction of an existential variable in a given position can never lead to create another existential variable in the same position, hence an infinite number of existential variables.

For example, let $R_1 = h(x) \rightarrow p(x, y)$ and $R_2 = p(u, v), q(v) \rightarrow h(v)$. The position graph of $\{R_1, R_2\}$ contains a special edge from $(h, 1)$ to $(p, 2)$ due to R_1 and an edge from $(p, 2)$ to $(h, 1)$ due to R_2. Hence, $\{R_1, R_2\}$ is not weakly-acyclic.

Range-restricted rules are a special case of weakly-acyclic rules since they do not have existential variables at all.

The second graph is called *graph of rule dependencies* (GRD) (Baget et al. 2011a; Grau et al. 2013). Intuitively, a rule R_j *depends* on a rule R_i if there is a fact base such that an application of R_i on this fact base leads to a new application of R_j. This abstract condition can be effectively computed by a specific unifier between the head of R_i and the body of R_j. The GRD of a set of rules $\mathcal{R}$ has a set of nodes in bijection with $\mathcal{R}$ and edges (R_i, R_j) whenever the rule R_j depends on the rule R_i. A set

of rules is *aGRD* if its GRD has no circuit. In the above example, R_1 depends on R_2 but not the contrary (indeed, one can check that an application of R_1 can never lead to trigger an application of R_2: it produces an atom of the form $p(x, y)$, where y is a new existential variable, but it does not produce the atom $q(y)$, which on the other hand cannot exist in the fact base since y is new, hence no new application of R_2 is made possible); hence, $\{R_1, R_2\}$ is aGRD. Weak-acyclicity and aGRD are in fact incomparable properties, but they admit common generalisations (Grau et al. 2013; Rocher 2016).

The *fus* property is ensured by conditions that allow one to bound the maximal size of a non-redundant CQ generated during the rewriting. Concrete *fus* classes include in particular *linear* rules and *sticky* rules (these two classes being incomparable). A linear rule has a body restricted to a single atom. The stickiness of a set of rules is defined by a marking procedure of the variables occurring in rules; then the set of rules is said to be sticky if no marked variable in a rule body occurs in two different atoms; intuitively, this ensures that a variable generated during the rewriting process occurs in at most one atom (Calì et al. 2010; Thomazo 2013). The decidability of ontology-mediated query answering for sets of rules with the *bts* property comes from an indirect argument (following a result by Courcelle), which does not directly provide a suitable algorithm. However, the expressive subclass known as *gbts* allows one to greedily build a tree decomposition of the (possibly infinite) saturated fact base, such that this tree decomposition has a bounded width. Concrete rule classes in the *gbts* family are also known as the guarded family, inspired by the guarded fragment of first-order logic. We list here the main members of this family (Calì et al. 2008; Baget et al. 2011a). A rule is *guarded* if an atom of its body (called a guard) contains all the variables that occur in its body. Note that a linear rule is by definition guarded, hence it is not only *fus* but also *gbts*. A rule is *frontier-one* if it has only one frontier variable. A rule is *frontier-guarded* if an atom of its body guards all the variables of its frontier (hence, this class generalizes both guarded and frontier-one rules).

Most Horn description logics belong to the *gbts* family, except those including transitivity, and more generally composition, of binary relations. Indeed, transitivity destroys the tree-like structure of the saturation. For instance, $\mathcal{EL}$ and $\mathcal{ELHI}$ are frontier-guarded, while DL-Lite$_\mathcal{R}$ is linear (hence, also *fus*).

5 Conclusion

Reasoning with ontologies is becoming central in many data-centric applications for which ontologies are a way to integrate heterogeneous data by providing a common conceptual vocabulary. In this setting, it is crucial to deeply understand the impact of the ontological constructs on the complexity of the main reasoning problems. This chapter provides the required formal background and results to help data practitioners to choose the knowledge representation formalism with the best expressivity / complexity tradeoff regarding their application needs.

Ontology-mediated query answering is a vibrant area at the crossroads of several domains, namely data management, knowledge representation and reasoning, and the Semantic Web. Undoubtedly, many issues remain to be solved before the widespread adoption of the framework in practice. We will mention some of the challenges currently addressed in the area, without any claim to be exhaustive. Up to recently, most work were limited to conjunctive queries, or slight extensions of them, while the ability to process more complex queries is required. The combination of conjunctive queries and navigational queries has begun to be investigated (e.g., Stefanoni et al. 2014; Bienvenu et al. 2015; Baget et al. 2017). New algorithmic techniques are being developed to meet the challenge of scalability beyond simple ontological languages (e.g., approaches that combine materialization of inferences and query rewriting Lutz et al. 2013; Feier et al. 2015). The integration of heterogeneous data under the form of a (possibly virtual) fact base relies on so-called mappings from these data to facts over the ontological vocabulary (Poggi et al. 2008): while mappings are a classical notion in data integration, their introduction poses new challenges in the presence of an ontology (e.g., Bienvenu and Rosati 2016; Botoeva et al. 2016). Representing and reasoning with temporal and spatial data, as well as information about their reliability and provenance, are of uttermost importance in most data-centric applications and have only recently begun to be explored in the context of ontology-mediated query answering (Artale et al. 2015; Borgwardt et al. 2015; Bereta and Koubarakis 2016; Brandt et al. 2017). Last but not least, practically robust query answering has to be tolerant to data inconsistencies, which are likely to occur in large datasets especially when the data issues from multiple data sources (e.g., Lembo et al. 2015; Lukasiewicz et al. 2015; Bienvenu and Bourgaux 2016).

References

Abiteboul S, Hull R, Vianu V (1995) Foundations of databases. Addison-Wesley, Boston

Artale A, Calvanese D, Kontchakov R, Zakharyaschev M (2009) The DL-Lite family and relations. J Artif Intell Res (JAIR) 36:1–69

Artale A, Kontchakov R, Kovtunova A, Ryzhikov V, Wolter F, Zakharyaschev M (2015) First-order rewritability of temporal ontology-mediated queries. In: Yang Q, Wooldridge M (eds) Proceedings of the 24th international joint conference on artificial intelligence (IJCAI). AAAI Press, California, pp 2706–2712

Baader F, Brandt S, Lutz C (2005) Pushing the EL envelope. In: Kaelbling LP, Saffiotti A (eds) Proceedings of the 19th international joint conference on artificial intelligence (IJCAI), pp 364–369

Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider PF (eds) (2003) The description logic handbook: theory, implementation, and applications. Cambridge University Press, Cambridge

Baader F, Horrocks I, Lutz C, Sattler U (2017) An introduction to description logic. Cambridge University Press, Cambridge

Baader F, Küsters R, Molitor R (1999a) Computing least common subsumers in description logics with existential restrictions. In: Proceedings of the 16th international joint conference on artificial intelligence (IJCAI), pp 96–103

Baader F, Molitor R, Tobies S (1999b) Tractable and decidable fragments of conceptual graphs. In: Proceedings of the 7th international conference on conceptual structures (ICCS). LNAI, vol 1640, Springer, Berlin, pp 480–493

Baget J, Bienvenu M, Mugnier M, Thomazo M (2017) Answering conjunctive regular path queries over guarded existential rules. In: Proceedings of the 26th international joint conference on artificial intelligence (IJCAI), pp 793–799

Baget J-F, Croitoru M, Gutierrez A, Leclère M, Mugnier M-L (2010) Translations between RDF(S) and conceptual graphs. In: Proceedings of the 18th international conference on conceptual structures (ICCS), pp 28–41

Baget J-F, Leclère M, Mugnier M-L, Salvat E (2011a) On rules with existential variables: walking the decidability line. Artif Intell (AIJ) 175(9–10):1620–1654

Baget J-F, Mugnier M-L, Rudolph S, Thomazo M (2011b) Walking the complexity lines for generalized guarded existential rules. In: Proceedings of the 22nd international conference on artificial intelligence (IJCAI), pp 712–717

Bereta K, Koubarakis M (2016) Ontop of geospatial databases. In: Groth PT, Simperl E, Gray AJG, Sabou M, Krötzsch M, Lécué F, Flöck F, Gil Y (eds) Proceedings of the 15th international semantic web conference (ISWC), pp 37–52

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 279

Bienvenu M, Bourgaux C (2016) Inconsistency-tolerant querying of description logic knowledge bases. In: Lecture notes of the 12th international reasoning web summer school. LNCS, vol 9885. Springer, Berlin, pp 156–202

Bienvenu M, Kikot S, Kontchakov R, Podolskii VV, Zakharyaschev M (2018) Ontology-mediated queries: combined complexity and succinctness of rewritings via circuit complexity. J ACM 65(5):28:1–28:51

Bienvenu M, Ortiz M (2015) Ontology-mediated query answering with data-tractable description logics. Lecture notes of the 11th international reasoning web summer school. LNCS, vol 9203. Springer, Berlin, pp 218–307

Bienvenu M, Ortiz M, Simkus M (2015) Regular path queries in lightweight description logics: complexity and algorithms. J Artif Intell Res (JAIR) 53:315–374

Bienvenu M, Rosati R (2016) Query-based comparison of mappings in ontology-based data access. In: Proceedings of the 15th international conference on the principles of knowledge representation and reasoning (KR), pp 197–206

Blackburn P, Benthem JV, Wolter F (2006) Handbook of modal logic. Springer, Berlin

Borgida A (1996) On the relative expressiveness of description logics and predicate logics. Artif intell (AIJ) 82:353–367

Borgwardt S, Lippmann M, Thost V (2015) Temporalizing rewritable query languages over knowledge bases. J Web Sem 33:50–70

Botoeva E, Calvanese D, Santarelli V, Savo DF, Solimando A, Xiao G (2016) Beyond OWL 2 QL in OBDA: rewritings and approximations. In: Proceedings of the 30th AAAI conference on artificial intelligence, pp 921–928

Brandt S, Kalayci EG, Kontchakov R, Ryzhikov V, Xiao G, Zakharyaschev M (2017) Ontology-based data access with a horn fragment of metric temporal logic. In: Singh SP, Markovitch S (eds) Proceedings of the 31st AAAI conference on artificial intelligence. AAAI Press, California, pp 1070–1076

Calì A, Gottlob G, Kifer M (2008) Taming the infinite chase: query answering under expressive relational constraints. In: Proceedings of the 11th international conference on principles of knowledge representation and reasoning (KR), pp 70–80

Calì A, Gottlob G, Lukasiewicz T (2009) A general datalog-based framework for tractable query answering over ontologies. In: Proceedings of the 28th international conference on principles of database systems (PODS), pp 77–86

Calì A, Gottlob G, Pieris A (2010) Advanced processing for ontological queries. PVLDB 3(1):554–565

Calvanese D, Giacomo GD, Lembo D, Lenzerini M, Rosati R (2007) Tractable reasoning and efficient query answering in description logics: the DL-Lite family. J Autom Reas (JAR) 39(3):385–429

Chein M, Mugnier M-L (2009) Graph-based knowledge representation and reasoning-computational foundations of conceptual graphs. Advanced information and knowledge processing, Springer, Berlin

Chen P (1976) The entity-relationship model: Toward a unified view of data. ACM Trans Database Syst (TODS) 1(1):9–36

Dantsin E, Eiter T, Gottlob G, Voronkov A (2001) Complexity and expressive power of logic programming. ACM Comput Surv 33(3):374–425

Dau F (2003) The logic system of concept graphs with negation and its relationship to predicate logic. Lecture notes in computer science, vol 2892. Springer, Berlin

Fagin R, Kolaitis PG, Miller RJ, Popa L (2005) Data exchange: semantics and query answering. Theor Comput Sci 336(1):89–124

Feier C, Carral D, Stefanoni G, Grau BC, Horrocks I (2015) The combined approach to query answering beyond the OWL 2 profiles. In: Proceedings of the 24th international joint conference on artificial intelligence (IJCAI), pp 2971–2977

Gottlob G, Kikot S, Kontchakov R, Podolskii VV, Schwentick T, Zakharyaschev M (2014) The price of query rewriting in ontology-based data access. Artif Intell (AIJ) 213:42–59

Gottlob G, Orsi G, Pieris A, Simkus M (2012) Datalog and its extensions for semantic web databases. In: Lecture notes of the international reasoning web summer school. LNCS, vol 7487. Springer, Berlin, pp 54–77

Gottlob G, Schwentick T (2012) Rewriting ontological queries into small nonrecursive datalog programs. In: Proceedings of the 13th international conference on the principles of knowledge representation and reasoning (KR)

Grau BC, Horrocks I, Kazakov Y, Sattler U (2008) Modular reuse of ontologies: theory and practice. J Artif Intell Res (JAIR) 31:273–318

Grau BC, Horrocks I, Krötzsch M, Kupke C, Magka D, Motik B, Wang Z (2013) Acyclicity notions for existential rules and their application to query answering in ontologies. J Artif Intell Res (JAIR) 47:741–808

Gruber R (1993) A translation approach to portable ontology specifications. Knowl Acquis 5(2):199–220

Guarino N (1998) Formal ontology and information systems. In: Guarino N (ed) Formal ontology and information systems. IOS Press, Amsterdam, pp 3–15

Horrocks I, Kutz O, Sattler (2006) The even more irresistible SROIQ. In: Proceedings of the 10th international conference on principles of knowledge representation and reasoning (KR), pp 57–67

Horrocks I, Sattler U, Tobies S (1999) Practical reasoning for expressive description logics. In: Proceedings of the 6th international conference on logic programming and automated reasoning (LPAR), pp 161–180

Konev B, Lutz C, Walther D, Wolter F (2013) Model-theoretic inseparability and modularity of description logic ontologies. Artif Intell (AIJ) 203:66–103

Kontchakov R, Wolter F, Zakharyaschev M (2010) Logic-based ontology comparison and module extraction, with an application to DL-Lite. Artif Intell (AIJ) 174(15):1093–1141

Lehmann F (1992) Semantic networks in artificial intelligence. Elsevier Science Inc., New York

Lembo D, Lenzerini M, Rosati R, Ruzzi M, Savo DF (2015) Inconsistency-tolerant query answering in ontology-based data access. J Web Sem 33:3–29

Levy A, Rousset M-C (1998) Combining Horn rules and description logics in CARIN. Artif Intell (AIJ) 101

Lukasiewicz T, Martinez MV, Pieris A, Simari GI (2015) From classical to consistent query answering under existential rules. In: Proceedings of the 29th AAAI conference on artificial intelligence, pp 1546–1552

Lutz C, Seylan I, Toman D, Wolter F (2013) The combined approach to OBDA: taming role hierarchies using filters. In: Proceedings of the 12th international semantic web conference (ISWC), pp 314–330
Mugnier M, Thomazo M (2014) An introduction to ontology-based query answering with existential rules. In: Lecture notes of the 10th international reasoning web summer school. LNCS, vol 8714. Springer, Berlin, pp 245–278
Mugnier M-L (2011) Ontological query answering with existential rules. In: Proceedings of the 5th international conference on web reasoning and rule systems (RR), pp 2–23
Nebel B (1990) Terminological reasoning is inherently intractable. Artif Intell (AIJ) 43(2):235–249
Ortiz M, Simkus M (2012) Reasoning and query answering in description logics. In: Lecture notes of the 8th international reasoning web summer school, LNCS, vol 7487. Springer, Berlin, pp 1–53
Peñaloza R, Sertkaya B (2017) Understanding the complexity of axiom pinpointing in lightweight description logics. Artif Intell (AIJ) 250:80–104
Poggi A, Lembo D, Calvanese D, De Giacomo G, Lenzerini M, Rosati R (2008) Linking data to ontologies. J Data Semant 10:133–173
Rocher S (2016) Querying existential rule knowledge bases: decidability and complexity. Université de Montpellier PhD thesis
Schild K (1991) A correspondence theory for terminological logics: preliminary report. In: Proceedings of the 12th international joint conference on artificial intelligence (IJCAI)
Schlobach S, Cornet R (2003) Non-standard reasoning services for the debugging of description logic terminologies. In: Proceedings of the 18th international joint conference on artificial intelligence (IJCAI), pp 355–362
Sebastiani R, Vescovi M (2009) Axiom pinpointing in lightweight description logics via horn-sat encoding and conflict analysis. In: Proceedings of the 22nd international conference on automated deduction (CADE). LNCS, vol 5663. Springer, Berlin, pp 84–99
Sowa JF (1976) Conceptual graphs. IBM J Res Devel
Sowa JF (1984) Conceptual structures: information processing in mind and machine. Addison-Wesley, Boston
Stefanoni G, Motik B, Krötzsch M, Rudolph S (2014) The complexity of answering conjunctive and navigational queries over OWL 2 EL knowledge bases. J Artif Intell Res (JAIR) 51:645–705
Thomazo M (2013) Conjunctive query answering under existential rules - decidability, complexity, and algorithms. PhD thesis, Montpellier 2 University, France
W3C (2004a) OWL web ontology language. http://www.w3.org/2004/OWL/
W3C (2004b) RDF vocabulary description language 1.0: RDF schema. http://www.w3.org/TR/rdf-schema/
W3C (2012a) OWL 2 web ontology language. https://www.w3.org/TR/owl-syntax/
W3C (2012b) OWL 2 web ontology language profiles. https://www.w3.org/TR/owl2-profiles/

Compact Representation of Preferences

Souhila Kaci, Jérôme Lang and Patrice Perny

Abstract This chapter presents the main families of representation languages for preferences on combinatorial domains (composed by several attributes or variables with discrete value domains). In the first part of the chapter, we present the problem in its full generality. A large part of these languages are said to be *graphical*, because they work by expressing elementary preferences in a local way, using structural independence properties that are represented under the form of a graph. In the second (respectively, third) part of the chapter we review graphical languages for expressing *ordinal* (respectively, *cardinal*) preferences. Another class of preference representation languages makes use of *(propositional) logic*; they will be reviewed in the fourth part of the chapter, together with proper 'preference logics'.

1 Introduction

The specification of a decision making problem includes the expression of the preferences of an agent, or of several agents, on the set of available alternatives. This is for instance the case in planning, where an autonomous agent acts for the user who programmed it. This is also the case in individual or collective decision aid, where an autonomous agent has to help a user or a group of users to make a decision; examples of such decision aid problems are recommender systems, product configuration etc. In each of these examples, specifying a goal, as for instance in classical planning, is often insufficiently expressive, since it does not allow to choose a suboptimal decision, but yet satisfactory, when the objective is not reachable.

S. Kaci (✉)
LIRMM, Université de Montpellier, Montpellier, France
e-mail: souhila.kaci@lirmm.fr

J. Lang
LAMSADE-CNRS, Université Paris-Dauphine, PSL Research University, Paris, France
e-mail: lang@lamsade.dauphine.fr

P. Perny
LIP6, Sorbonne Université, Paris, France
e-mail: patrice.perny@lip6.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_7

Preference modelling consists in studying different classes of mathematical structures for preference and characterizing them axiomatically. These models can be numerical (preferences are then modelled by utility functions mapping alternatives into numbers), qualitative (the numerical scale being replaced by an ordered qualitative scale) or ordinal (preferences are then binary relations, generally orders or weak orders).

This being said, the choice of a mathematical model for modelling the preferences of an agent does not say how they are *represented*, or more precisely, how they are *specified* in a satisfactory way. Of course, a possibility is to write them *explicitly* by listing all possible alternatives with their utility value (in the case of numerical preferences) or by listing them in their order of preference (in the case of ordinal preferences).

Clearly, this explicit mode of representation is feasible in practice only if the number of possible alternatives is small enough with respect to the computational resources available and the time we allow the user to interact with the system. This assumption is often unrealistic, in particular when the set of alternatives has a combinatorial structure, that is, when each alternative consists of the assignment of a value to each of a set of decision variables: in this case, the set of alternatives is the Cartesian product of the value domains for these variables, and its size increases exponentially with the number of variables. We give two illustrative examples below.

In the first example, an agent is asked to express her preferences about meals consisting of a first dish, a main dish, a dessert and a wine, with six possibilities for each: this makes 6^4 alternatives. This would not be a problem if the preferences between the values of a variable were independent of the value taken by other variables: in such a case it would be enough to ask the agent to specify her preferences independently on each of the value domains, and the joint preference over the set of all alternatives would be determined for instance via an aggregation function. In our example, expressing a utility function over the 6^4 alternatives would come down to specifying four utility functions on six alternatives each. But this becomes much more complicated of the agent wishes to express *dependencies* between variables, such as "I prefer white wine if the first dish or the main dish is made out of fish, and none of them is meat, red wine if one of the dishes is made out of meat, and I am indifferent between white wine and red wine in all other cases".

Our second example is a hiring problem: a committee has to recruit k new assistant professors out of n applicants. The space of alternatives cannot be identified to the set of applicants: it is instead the set of all subsets of k applicants among n, and thus it has a combinatorial structure. A committee member may express her preferences in an explicit way only if the dependencies between applicants can be ignored, which means that she cannot express correlated preferences between applicants, such as: "My favorite candidate is A, the next one is B, and lastly, C; however, A and B work in the same field, while C works in a totally different field, therefore I prefer to hire A and C together, and even B and C together, rather than A and B together."

For such problems, the size of the space of alternatives and the impossibility to decompose the description of the agent's preferences into smaller descriptions bearing each on an isolated variable makes it practically infeasible to ask the agent

to give her utility function or her preference relation in an explicit way under the form of a table or a list. Therefore, allowing the expression of a utility function or a preference relation over such a set of alternatives requires first to design a *language* for expressing preferences in a *concise* (or *succinct*) way. These languages, which we call *succinct preference representation languages*, should not only be succinct but should ideally be as expressive as possible, that is, to allow representing a set of utility functions or preference relations as large as possible.

The key problem in succinct preference representation is the expression of *preferential dependencies* between variables, as in the two examples above. In general, a tradeoff must be made between the expressivity of the language and its succinctness, which can translate into specific assumptions on the nature of the preferential dependencies we want to be able to express. The succinctness of the language then comes from the exploitation of these preferential independencies between variables.

Upstream the problem of representing preferences is the problem of *eliciting* them, that is, to interact with the user so as to acquire enough information on her preferences to suggest her a satisfactory (or, in some cases, optimal) alternative. The design of an elicitation protocol depends on the chosen preference representation language, and generally exploits specific assumptions on the structure of the user's preferences so as to reduce the amount of information to elicit and the cognitive effort required to communicate them; moreover, the difficulty of the elicitation process sometimes requires a trade-off between expressivity and communication complexity.[1] Besides, to make elicitation easier, it is important that the chosen language be cognitively relevant, i.e., close enough to intuition; ideally, the specification of preferences in a representation language should be easily translated from the way the agent expresses them in natural language.

Lastly, these languages must be equipped with *algorithms* that should be as efficient as possible, so as to allow the automation of the comparison of alternatives, of the ranking of several alternatives, and of the search for an optimal alternative.

Such preference representation languages have been particularly well-studied in the Artificial Intelligence literature, and more specifically within the research communities "*Knowledge Representation and Reasoning*" and "*Uncertainty in Artificial Intelligence*", that gather in, respectively, biennal and annual conferences, as well in the biennal specialised conference *Algorithmic Decision Theory* and the almost annual specialised workshop *International Multidisciplinary Workshop on Preference Handling*.

A large part of these languages are said to be "graphical", because they consist in expressing elementary preferences in a local way (that is, on variables or subsets of variables), by exploiting structural preferential independence relations under a graphical form, as do for instance Bayesian networks for the representation of joint probability distributions.

[1]The communication complexity of an individual or collective decision problem is the minimal amount of information to be communicated by the agents so that the outcome of the decision problem be completely determined.

After listing in more detail, in Sect. 2, the features of preference representation languages, each of the subsequent sections will be dedicated to a particular class of languages. In Sects. 3 and 4 we will survey graphical representation languages for, respectively, ordinal and cardinal preferences. In Sect. 5 we will survey logical preference representation languages, which we will briefly connect with preference logics.

2 Compact Preference Representation Languages

In this section, we first give the general definition of a preference representation language, then we formally describe the criteria for evaluating them.

In the rest of this chapter, we consider a set of feasible alternatives $\mathscr{X}$. A *utility function* over $\mathscr{X}$ is a function $u : \mathscr{X} \to \mathbb{R}$. A *preference relation* $\succsim$ over $\mathscr{X}$ is partial weak order, that is, a reflexive, transitive relation (not necessarily complete nor antisymmetric). The *strict preference* induced by $\succsim$ is the strict order $\succ$ defined by: $x \succ x'$ if and only if $x \succsim x'$ and not $(x' \succsim x)$. The *indifference relation* induced by $\succsim$ is the equivalence relation $\sim$ defined by $x \sim x'$ if and only if $x \succsim x'$ and $x' \succsim x$. If u is a utility function then the preference relation $\succsim_u$ induced by u is defined by $x \succsim_u x'$ if and only if $u(x) \geq u(x')$. We will use the terminology "preference structure" for designating, whichever is the case, a utility function (also called cardinal preference structure) or a preference relation (also called ordinal preference structure).

A *preference representation language* is a pair $\mathscr{R} = \langle L, \mathscr{I} \rangle$, where L is a formal language, and $\mathscr{I}$ a function mapping each $\Phi \in L$ to a preference relation $\succsim_\Phi$ over $\mathscr{X}$ or a utility function u_Φ over $\mathscr{X}$, depending on the ordinal or cardinal nature of the language L. For example, propositional logic can be seen as a compact preference representation language: L is the set of all propositional formulas built on a finite set of propositional symbols PS, $\mathscr{X}$ is the set of all truth assignments on PS, i.e., $\mathscr{X} = 2^{PS}$, and $\mathscr{I}(\varphi)$ is the function u_φ defined by: for all $x \in 2^{PS}$, $u_\varphi(x) = +1$ if $x \models \varphi$ and 0 if $x \models \neg\varphi$; or, if one prefers an ordinal output, $\mathscr{I}(\varphi)$ is the preference relation $\succsim_\varphi$ defined by: for all $x, y \in 2^{PS}$, $x \succsim y$ if and only if $x \models \varphi$ or $y \models \neg\varphi$.

The criteria according to which the different languages can be evaluated are their expressivity, their succinctness power, their cognitive relevance, and the complexity of the associated computational tasks.

The *expressivity* of a language $\langle L, \mathscr{I} \rangle$ is the set of all preference structures that can be expressed in L, that is, $\mathscr{I}(L)$. For example, the set of all preference relations expressible by propositional logic is the set of all dichotomous preference relations, that is, the set of all relations $\succsim$ such that $\mathscr{X}$ can be partitioned into $\mathscr{X}^+$ and $\mathscr{X}^-$, with $x \succsim x'$ if and only if $x \in \mathscr{X}^+$ or $x' \in \mathscr{X}^-$: $\mathscr{X}^+$ represents the set of all "good" alternatives and $\mathscr{X}^-$ the set of "bad" ones. A language $\langle L_1, \mathscr{I}_1 \rangle$ is at least as expressive as a language $\langle L_2, \mathscr{I}_2 \rangle$ if $\mathscr{I}_1(L_1) \supseteq \mathscr{I}_2(L_2)$.

The *succinctness power* of a language is a relative notion: a language $\langle L_1, \mathscr{I}_1 \rangle$ is at least as succinct as a language $\langle L_2, \mathscr{I}_2 \rangle$ in, informally, every preference structure that can be expressed in L_2 can also be expressed in L_1 *without significative*

superpolynomial increase of the representation size; or, formally, if there exists a function $f : L_2 \to L_1$ such that (a) $\mathscr{I}_2 = \mathscr{I}_1 \circ f$ and (b) there exists a polynomial p such that for all $\Phi \in L_2$, $|f(\Phi)| \leq p(|\Phi|)$. Obviously, if $\langle L_1, \mathscr{I}_1 \rangle$ is at least as succinct as $\langle L_2, \mathscr{I}_2 \rangle$ then $\langle L_1, \mathscr{I}_1 \rangle$ is at least as expressive as $\langle L_2, \mathscr{I}_2 \rangle$. For examples of comparison between languages from the point of view of expressivity and succinctness, see for instance Coste-Marquis (2004) and Uckelman et al. (2009).

The computational difficulty of a language L consists in determining the computational complexity, and designing efficient algorithms, for the following tasks:

- COMPARISON: given two alternatives x and x', determine whether $x \succsim x'$;
- OPTIMALITY: given an alternative x, determine whether x is nondominated, that is, if there does not exist an alternative x' such that $x' \succ x$;
- OPTIMISATION: find a non-dominated alternative, either in the full set of alternatives, or in a subset of feasible alternatives defined by a feasibility constraint.

3 Graphical Languages and Ordinal Preferences: CP-Nets, Variants and Extensions

3.1 Preferential Independence

Let $\mathscr{V} = \{X_1, \ldots, X_n\}$ be a set of *variables*, or *attributes*, associated with finite *value domains* $D_1, \ldots, D_n$. A variable X_i is *binary* if D_i has two elements, which by convention we note x_i and $\overline{x}_i$. The set of available alternatives is, by default, $\mathscr{X} = D_{\mathscr{V}} = D_1 \times \cdots \times D_n$; sometimes, it will be a subset of $D_1 \times \cdots \times D_n$ defined by feasibility constraints. If $\mathscr{W} \subseteq \mathscr{V}$, we let $D_{\mathscr{W}} = \times_{X_i \in \mathscr{W}} D_i$. Elements of $\mathscr{X}$ will generally be denoted using vectorial notation $\mathbf{x}$. For all disjoint subsets disjoints $\mathscr{U}$ and $\mathscr{W}$ of $\mathscr{V}$, the concatenation of the assignments $\mathbf{u} \in \mathscr{U}$ and $\mathbf{w} \in \mathscr{W}$, denoted $\mathbf{uw}$, is the $(\mathscr{U} \cup \mathscr{W})$-assignment, which assigns to the variables of $\mathscr{U}$ (resp. $\mathscr{W}$) the value assigned by $\mathbf{u}$ (resp. $\mathbf{w}$). If $\mathbf{x} \in \mathscr{X}$ and $\mathscr{U} \subseteq \mathscr{V}$, we note $\mathbf{x}^{\downarrow \mathscr{U}}$ the projection of $\mathbf{x}$ on the variables of $\mathscr{U}$.

Conditional Preference Networks, for short *CP-nets* (Boutilier et al. 2004a), are a graphical language for the representation of preferences based on the notion of *preferential independence* (Keeney and Raiffa 1976). Let $\{\mathscr{U}, \mathscr{V}, \mathscr{W}\}$ be a partition of the set of the variables $\mathscr{V}$, and $\succ$ a strict preference relation. $\mathscr{U}$ is *preferentially independent of* $\mathscr{V}$ *given* $\mathscr{W}$ w.r.t. $\succ$ if for all $\mathbf{u}_1, \mathbf{u}_2 \in D_{\mathscr{U}}$, $\mathbf{v}_1, \mathbf{v}_2 \in D_{\mathscr{V}}$ and $\mathbf{w} \in D_{\mathscr{W}}$, we have $\mathbf{u}_1\mathbf{v}_1\mathbf{w} \succ \mathbf{u}_2\mathbf{v}_1\mathbf{w}$ if and only if $\mathbf{u}_1\mathbf{v}_2\mathbf{w} \succ \mathbf{u}_2\mathbf{v}_2\mathbf{w}$.[2] Unlike probabilistic independence, preferential independence is a *directed* notion: X can be preferentially independent of Y given Z without Y being preferentially independent of X given Z. If, for every variable $X_i \in V$, X_i is preferentially independent of $V \setminus \{X_i\}$, then the preference relation $\succ$ is said to be *weakly separable*.

[2]This notion can be analogously be defined for weak orders $\succsim$ exactly in the same way.

For example, let $\mathcal{V} = \{A, B, C\}$ with $D_A = \{a, \overline{a}\}$, $D_B = \{b, \overline{b}\}$, $D_C = \{c, \overline{c}\}$, and the preference relation $\succ$ defined par

$$abc \succ \overline{a}bc \succ ab\overline{c} \succ a\overline{b}\overline{c} \succ a\overline{b}c \succ \overline{a}\overline{b}\overline{c} \succ \overline{a}\overline{b}c \succ \overline{a}b\overline{c}.$$

With respect to $\succ$, A is preferentially independent of $\{B, C\}$, C is preferentially independent of A given B, but depends on B given A, and B depends both on A and C. An example of a weakly separable preference relation is

$$abc \succ \overline{a}bc \succ ab\overline{c} \succ ab\overline{c} \succ \overline{a}b\overline{c} \succ a\overline{b}c \succ \overline{a}\overline{b}c \succ a\overline{b}\overline{c} \succ \overline{a}\overline{b}\overline{c}.$$

Here, $A = a$ is preferred to $A = \overline{a}$ whatever the fixed values of B and C, and similarly for B and C.

3.2 CP-Nets

A CP-net (Boutilier et al. 2004a) is composed of a directed graph representing the dependences between variables and of a set of conditional preference tables expressing, for each variable, the local preferences on the values of its domain given every combination of values of its parents.

Formally, a *CP-net* over a set of variables $\mathcal{V} = \{X_1, \ldots, X_n\}$ is a pair $\mathcal{N} = \langle G, P \rangle$ where G is a directed graph on $\mathcal{V}$ and P is a set of conditional preference tables $CPT(X_i)$ for each $X_i \in \mathcal{V}$. For each variable X_i, $Par(X_i)$ denotes the set of the parents of X_i in G, and we let $NonPar(X_i) = V \setminus (\{X_i\} \cup Par(X_i))$. The edges of G express preferential dependencies: each variable is preferentially independent of its non-parents in G given its parents. Each conditional preference table associates a linear order[3] on D_i with each instantiation $\mathbf{u}$ of $Par(X_i)$, denoted $\mathbf{u} :>$; the meaning of $\mathbf{u} : x_i^j \succ x_i^k$ is that for each instantiation $\mathbf{z}$ of $NonPar(X_i)$, we have $\mathbf{u}x_i^j\mathbf{z} \succ \mathbf{u}x_i^k\mathbf{z}$. In more readable terms: when $U = \mathbf{u}$, $X = x^j$ is preferred to $X = x^i$, everything else being equal (ceteris paribus).

Example 1 A user is looking for a plane ticket. Let there be three variables: T (time of the flight), with possible values d (day) and n (night); S (stop), with possible values s (yes) and $\overline{s}$ (no); and C (airline), with possible values c_1 and c_2. The user has the following preferences:

- she prefers a day flight to a night flight, unconditionally;
- for a day flight she prefers to have a stop, but for a night flight she prefers not to;
- for a day flight with a stop she prefers airline c_1 because she will be able to spend a few hours in an airport she likes; in all other cases she prefers c_2.

[3] It is also possible to define CP-nets with indifferences—see Boutilier et al. (2004a), which does not change much to the definitions nor to the results. For the sake of concision, we will omit this possibility.

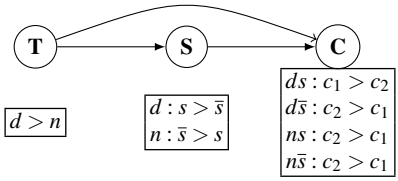


Fig. 1 A CP-net $\mathscr{N}$ with acyclic dependencies

The preferences of the user are expressed by the CP-net $\mathscr{N}$ whose set of the variables is $\mathscr{V} = \{T, S, C\}$, the set of the alternatives is $D_T \times D_S \times D_C = \{d, n\} \times \{s, \overline{s}\} \times \{c_1, c_2\}$, and the conditional preference tables are represented on Fig. 1.

3.3 Semantics of CP-Nets

The semantics of a CP-net is defined as follows. A strict preference relation $\succ$ satisfies $\mathscr{N}$ if for every variable X_i, for all values $x_i, x_i' \in D_i$, all assignments $\mathbf{u}$ of $Par(X_i)$, and every assignment $\mathbf{z}$ of $NonPar(X_i)$, we have $\mathbf{u}x_i\mathbf{z} \succ \mathbf{u}x_i'\mathbf{z}$ if and only if $CPT(X_i)$ contains the entry $\mathbf{u} : x_i > x_i'$. A CP-net is *satisfiable* if there exists a preference relation that satisfies it. For any *satisfiable* CP-net $\mathscr{N}$, $\succ_{\mathscr{N}}$ is defined as the smallest preference relation that satisfies $\mathscr{N}$, or equivalently, as the transitive closure of $\{\mathbf{u}x_i\mathbf{z} \succ \mathbf{u}x_i'\mathbf{z} \mid i = 1, \ldots, n; x_i, x_i' \in D_i; \mathbf{u} \in Par(X_i); \mathbf{z} \in NonPar(X_i); CPT(X_i)$ contains $\mathbf{u} : x_i > x_i'\}$.

Example 1, Continued

- $Par(T) = \emptyset$ and $NonPar(T) = \{S, C\}$; the table associated with T indicates that $T = d$ is preferred to $T = n$ ceteris paribus, that is, for each fixed pair of values for S and C; this represents the following four pairs in the preference relation $\succ_{\mathscr{N}}$:

$$\{dsc_1 \succ_{\mathscr{N}} nsc_1, dsc_2 \succ_{\mathscr{N}} nsc_2, d\overline{s}c_1 \succ_{\mathscr{N}} n\overline{s}c_1, d\overline{s}c_2 \succ_{\mathscr{N}} n\overline{s}c_2\}.$$

- $Par(S) = \{T\}$ and $NonPar(S) = \{C\}$; the table associated with S indicates that when $T = d$, $S = s$ is preferred to $S = \overline{s}$, and when $T = n$, $S = \overline{s}$ is preferred to $S = s$; this represents the following four pairs in $\succ_{\mathscr{N}}$:

$$\{dsc_1 \succ_{\mathscr{N}} d\overline{s}c_1, dsc_2 \succ_{\mathscr{N}} d\overline{s}c_2, n\overline{s}c_1 \succ_{\mathscr{N}} nsc_1, n\overline{s}c_2 \succ_{\mathscr{N}} nsc_2\}.$$

- $Par(T) = \{S, C\}$ and $NonPar(T) = \emptyset$; the table associated with C represents the following four pairs in $\succ_{\mathscr{N}}$:

$$\{dsc_1 \succ_{\mathscr{N}} dsc_2, d\overline{s}c_2 \succ_{\mathscr{N}} d\overline{s}c_1, nsc_2 \succ_{\mathscr{N}} nsc_1, n\overline{s}c_2 \succ_{\mathscr{N}} n\overline{s}c_1\}.$$

The induced preference relation $\succ_{\mathscr{N}}$ is represented on Fig. 2 (the edges obtained by transitivity are omitted).

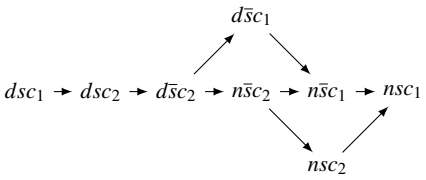


Fig. 2 Preference relation $\succ_{\mathscr{N}}$ induced by the CP-net $\mathscr{N}$

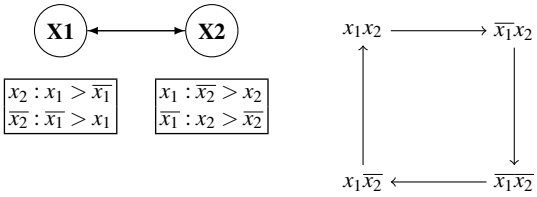


Fig. 3 An unsatisfiable CP-net

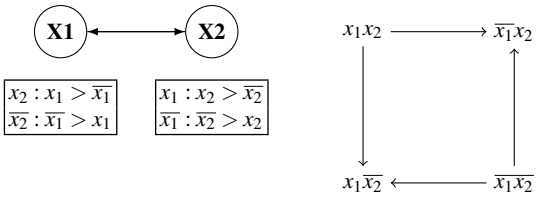


Fig. 4 A satisfiable CP-net with cyclic dependencies

A particularity of Example 1 is that the dependency graph of $\mathscr{G}$ is *acyclic*. Numerous works on CP-nets assume this, which makes many things simpler, because under this assumption, the CP-net is guaranteed to be satisfiable, and the associated queries, consisting in comparing two alternatives or in searching for a non-dominated alternative, are computable in polynomial time (Boutilier et al. 2004a).

When the dependency graph G is cyclic, the CP-net may be unsatisfiable, as we can see on the following example (Fig. 3).

Besides, a CP-net whose dependency graph contains cycles can sometimes be satisfiable, as the example on Fig. 4 shows.

The preference relation $\succ_{\mathscr{N}}$ induced by a CP-net $\mathscr{N}$ is generally not complete. The complete preference relations extending $\succ_{\mathscr{N}}$ can be seen as the possible models of the user's preferences, and an assertion on her preferences satisfied in each of these models can be seen as a consequence of the CP-net (Boutilier et al. 2004a). Thus we also define a notion of *consequence* in a CP-net: $\mathscr{N} \models \mathbf{x} \succ \mathbf{x}'$ if $\mathbf{x} \succ \mathbf{x}'$ is verified in each complete preference relation $\succ$ extending $\succ_{\mathscr{N}}$. Finally, for every preference relation $\succ$ there exists a satisfiable CP-net $\mathscr{N}$ (whose dependency graph can possibly contain cycles) such that $\succ$ extends $\succ_{\mathscr{N}}$. These remarks allow for a better understanding of the meaning of CP-nets. For the sake of clarity, in the rest of this paragraph we assume that all the variables X_i are binary. We first define the *hypercube* associated with $D_1 \times \cdots \times D_n$ as the set of pairs of alternatives that differs only on the value of one variable (such a pair will be called *pair of adjacent alternatives*). A *directed hypercube* associated with D is a function that for each

edge of the hypercube, specifies a direction (that is, specifies which of the two adjacent alternatives is preferred to the other). When an agent expresses a CP-net, she expresses only *a part* of her preference relation, that corresponds to *the projection of her preference relation on the hypercube associated with* $\mathscr{X}$. Thus, expressing a CP-net often implies a loss of information. For example, in the example of Fig. 4, the agent, by expressing the CP-net that corresponds to her preferences, has not been able to express her preference between x_1x_2 and $\overline{x_1}\overline{x_2}$, nor her preference between $x_1\overline{x_2}$ and $\overline{x_1}x_2$. There are *four* preference relations compatible with the expressed CP-net:

- $x_1x_2 \succ \overline{x_1}\overline{x_2} \succ x_1\overline{x_2} \succ \overline{x_1}x_2$;
- $x_1x_2 \succ \overline{x_1}\overline{x_2} \succ \overline{x_1}x_2 \succ x_1\overline{x_2}$;
- $\overline{x_1}\overline{x_2} \succ x_1x_2 \succ x_1\overline{x_2} \succ \overline{x_1}x_2$;
- $\overline{x_1}\overline{x_2} \succ x_1x_2 \succ \overline{x_1}x_2 \succ x_1\overline{x_2}$.

From these observations, let us now discuss the expressivity of CP-nets. There are two different ways of doing so. If one sticks to the formal definition of a compact preference representation languages, as defined in Sect. 2, then the function Ind is defined by $Ind(\mathscr{N}) = \succ_{\mathscr{N}}$: the field of expressivity of the CP-nets is therefore reduced to directed hypercubes. But this does not correspond to the practical use of CP-nets: whatever the application domain, there is no reason for assuming that the agent is only able to compare pairs of adjacent alternatives; then, the language of CP-nets allows only for an agent to express *a part* of her preference relation (that is, its projection on the hypercube), but does not require any restriction on the possible preferences of the agent: indeed, as we said above, *for every preference relation* $\succ$ *there exists a satisfiable CP-net* $\mathscr{N}$ *such that* $\succ$ *extends* $\succ_{\mathscr{N}}$. In some sense, CP-nets are *fully expressive* (because they do not impose any restriction on the user's preferences) *but not fully informative* (because they lead to a loss of information).

3.4 CP-Nets: Comparison and Optimisation

One of the main objectives of a preference representation language is to help answering various requests of the decider, such as the comparison of alternatives and the search for an optimal alternative. CP-nets are not only an intuitively satisfactory language for eliciting the preferences of a user, but they also allow (in many cases) to solve other tasks relatively easily.

3.4.1 Comparison

When the CP-net $\mathscr{N}$ *is satisfiable*, the induced preference relation $\succ_{\mathscr{N}}$ can be characterised equivalently in terms of *flipping sequences*. A descending flipping sequence is a sequence $\mathbf{x}_1, \ldots, \mathbf{x}_k$, where for each $i = j, \ldots, k-1$, (a) $\mathbf{x}_j$ and $\mathbf{x}_{j+1}$ differ on

one single variable X_i, and (b) $CPT(X_i)$ contains $\mathbf{u} : x_i > x_i'$, where $U = Par(X_i)$ and $\mathbf{u} = \mathbf{x}_i^{\downarrow U} = \mathbf{x}_{i+1}^{\downarrow U}$. We then have the following property (Boutilier et al. 2004a): for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$, $\mathbf{x} \succ_{\mathscr{N}} \mathbf{y}$ if and only if there exists a descending flipping sequence from $\mathbf{x}$ to $\mathbf{y}$.

Thus, on Example 1, there are three descending flipping sequences from dsc_2 to nsc_1:

$$dsc_2, d\bar{s}c_2, n\bar{s}c_2, n\bar{s}c_1, nsc_1$$
$$dsc_2, d\bar{s}c_2, d\bar{s}c_1, n\bar{s}c_1, nsc_1$$
$$dsc_2, d\bar{s}c_2, n\bar{s}c_2, nsc_2, nsc_1$$

This property shows that in practice, one can solve the dominance problem in CP-nets by looking for flipping sequences. One can also notice a strong structural proximity between the search of flipping sequences and STRIPS planning restricted to operators that with effects on a single variable (Boutilier et al. 2004a, b; Goldsmith et al. 2008).

The complexity of the problem of comparing alternatives depends on the structure of the dependency graph and of the nature (binary or not) of the variables: the problem is polynomial when the variables are binary and G is a hyper-tree (Boutilier et al. 2004a), NP-complete if the variables are binary and G verifies the property that the number of paths between two variables is bounded by a polynomial in the size of the CP-net (Boutilier et al. 2004a), and PSPACE-complete without any assumption on G, and this even if the variables are all binary (Goldsmith et al. 2008).

3.4.2 Optimisation

When the dependency graph of the CP-net is acyclic, there exists a unique dominating alternative (and a fortiori a unique non-dominated alternative), and this alternative can be determined in polynomial time by the *forward sweep* procedure, consisting in considering the variables in an order compatible with G (without loss of generality, $X_1 > \cdots > X_n$) and in choosing, for each variable X_i, the preferred value of X_i for the values of $X_1, \ldots, X_{i-1}$ already chosen. For example, with the CP-net of Fig. 3, the forward sweep procedure runs as follows:

- *step 1*: the preferred value of T (unconditionally) is d; this leads to the assignment $T := d$.
- *step 2*: the preferred value of S given $T = d$ is s; this leads to the assignment $S := s$.
- *step 3*: the preferred value of C given $T = d$ and $S = s$ is c_1; this leads to the assignment $C := c_1$, and one finally obtains the alternative dsc_1.

The *forward sweep* algorithm does not work anymore in the general case where G contains cycles; the problem of the existence of a non-dominated alternative is in this case NP-complete, and the search for a non-dominated alternative can be translated into a model finding problem in propositional logic (in the binary variable

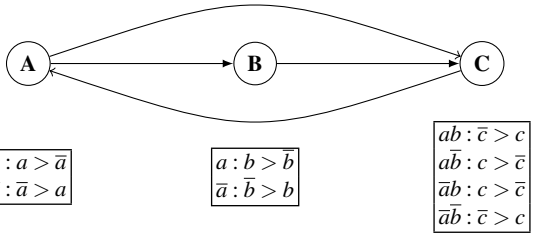


Fig. 5 An unsatisfiable CP-net can have nondominated alternatives

case) or of the search of a solution in a CSP in the general case (Brafman and Dimopoulos 2004). In the binary case, each entry $u : x \succ \overline{x}$ (respectively $u : \overline{x} \succ x$) of each table is translated into the clause $u \rightarrow x$ (respectively $u \rightarrow \neg x$). Thus, in Example 2, the clauses corresponding to the entries are $b \rightarrow a$, $\neg b \rightarrow \neg a$, $a \rightarrow b$, $\neg a \rightarrow \neg b$; their conjunction is the formula $\Phi_{\mathscr{N}} = (b \rightarrow a) \wedge (\neg b \rightarrow \neg a) \wedge (a \rightarrow b) \wedge (\neg a \rightarrow \neg b)$, which is equivalent to $a \leftrightarrow b$. The set of the models of $\Phi_{\mathscr{N}}$ is $\{ab, \overline{a}\overline{b}\}$: these are the non-dominated alternatives for $\mathscr{N}$.

Note that $\Phi_{\mathscr{N}}$ can be satisfiable even when $\mathscr{N}$ is unsatisfiable, as it can be seen on the following example (Fig. 5):

Example 2 $\succ_{\mathscr{N}}$ has a cycle: $abc \succ \overline{a}bc \succ \overline{a}b\overline{c} \succ ab\overline{c} \succ abc$. $\Phi_{\mathscr{N}} \equiv (a \leftrightarrow c) \wedge (b \leftrightarrow a) \wedge (c \leftrightarrow (a \leftrightarrow b)) \equiv \neg a \wedge \neg b \wedge c$; Φ_N is satisfiable and its unique model is $\{\overline{a}\overline{b}c\}$, which means that $\overline{a}\overline{b}c$ is undominated.

This shows how we can perform optimisation tasks from of an unsatisfiable CP-net.

The following table gives the complexity of the main queries, w.r.t. the structure of the dependency graph of the CP-net, when the variables are binary:

	G hypertree	G acyclic	Any G
Optimisation	P	P	NP-hard
Comparison	P	NP-hard (in NP?)	PSPACE-complete
Optimality	P	P	P
Satisfiability	trivial	trivial	PSPACE-complete

The random generation of CP-nets following a uniform distribution is addressed in Allen et al. (2016).

3.5 Constrained CP-Nets

In many concrete problems, not all assignments of $\mathscr{X}$ correspond to feasible alternatives. A *constrained CP-net* consists in a CP-net $\mathscr{N}$ and a set of constraints Γ restricting the feasible alternatives. Constrained optimisation is particularly relevant for example for configuration problems (Domshlak et al. 2001).

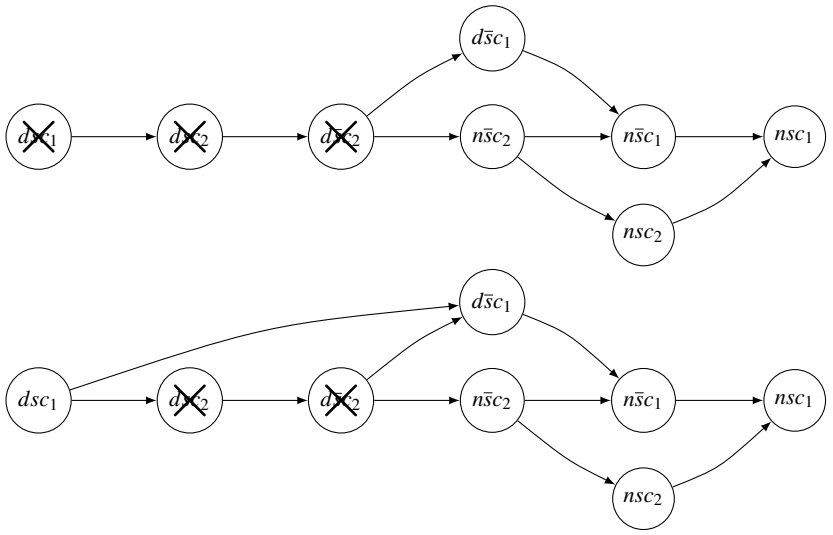


Fig. 6 Constrained CP-nets: two examples

The constraints can be expressed in a compact representation language, typically in the language of the constraint satisfaction problems (CSP), or, in the case of binary variables, of propositional logic. Every alternative satisfying Γ is said to be *feasible*. The goal is to find an alternative $\mathbf{x}$ both feasible and undominated, that is, such that there is no feasible alternative $\mathbf{x}'$ such that $\mathbf{x}' \succ_{\mathscr{N}} \mathbf{x}$ (Boutilier et al. 2004b).

A different way of defining the optimal solutions in a constrained CP-net is suggested in Domshlak et al. (2006): $\mathbf{x}$ dominates $\mathbf{x}'$ if there exists a sequence of elementary flips from $\mathbf{x}$ to $\mathbf{x}'$ *that passes only through feasible alternatives*, and again, one looks for undominated alternatives, or equivalently, of feasible alternatives $\mathbf{x}$ such that there exists no elementary flip from another feasible alternative to $\mathbf{x}$.

Example 3 Consider again Example 1, and let us add the constraint that it is not possible to have a day flight with a stop: $T = d \Rightarrow S = \bar{s}$, and the constraint that airline c_2 has only night flights: $C = c_2 \Rightarrow T = n$. Alternative dsc_1, which was the optimal alternative of $\succ_{\mathscr{N}}$, is now unfeasible. The new undominated alternatives, w.r.t. the two above definitions, are $d\bar{s}c_1$ and $n\bar{s}c_2$. Suppose now that we have the constraint $C = c_2 \Rightarrow T = n$. According to Domshlak et al. (2006), dsc_1 and $n\bar{s}c_2$ are undominated, whereas only dsc_1 is non-dominated according to Boutilier et al. (2004b). The two examples are represented on Fig. 6.

3.6 Extensions and Variants of CP-Nets

CP-nets allow to represent preferences between different values of a variable, conditionally on the values of its parents, but they do not allow to express *importance*

relations between variables, nor *explicit preferences between tuples of values of several variables*. Several extensions of CP-nets have been defined so as to cope with this lack of expressivity.

TCP-nets (Brafman et al. 2006) enrich CP-nets by allowing the expression of *relative importance* relations between variables, conditionally to the values of other variables. A TCP-net contains (1) some preference statements (exactly as in CP-nets); (2) unconditional importance statements of the form $A \rhd B$ (A is more important than B); and (3) conditional importance statements of the form $A = a : B \rhd C$ (if $A = a$ then B is more important than C).

Example 4 Consider the TCP-net on the set of the three binary variables $\{A, B, C\}$, containing the conditional importance statements

$$\begin{array}{l} a : B \rhd C \\ \overline{a} : C \rhd B \end{array}$$

and the conditional preference statements

$$a > \overline{a} \qquad \begin{array}{l} a : b > \overline{b} \\ \overline{a} : \overline{b} > b \end{array} \qquad c > \overline{c}$$

The preference relation induced by this TCP-net is depicted on Fig. 7.

This relative importance notion $\rhd$ comes with a gain of expressivity. It allows, for instance, to express the preference $(a, \overline{b}, c) \succ (\overline{a}, b, c)$ by resolving the conflict between $a \succ \overline{a}$ and $b \succ \overline{b}$ by the fact that $A \rhd B$. However, if one considers a third value b' in the domain of B such that $\overline{b} > b'$, the decision maker might feel that (a, b', c) is preferred to $(\overline{a}, b, c)$, considering that falling from b to b' on B is too important for being compensated by an amelioration of $\overline{a}$ in a on A. In order to express this second preference in the language of the TCP-nets we would need to add $B \rhd A$, which contradicts the importance inequality $A \rhd B$ above. This difficulty comes from the fact that one does not take into account the values of the relevant attributes for expressing a notion of relative importance between attributes and that one does not really allow compensations. We will see further that with quantitative models, it is easier to represent these phenomena, since preference intensities can be expressed; for a discussion on this topic, see Gonzales et al. (2008). Obviously, this leads to an additional elicitation effort.

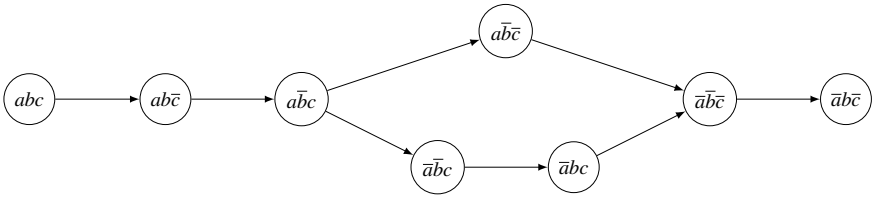


Fig. 7 A TCP-net and its associated preference relation

CP-theories (Wilson 2004) are even more general: they allow the expression of preferences on the values of a variable, conditionally on the values of its parent variables, given that some of the remaining variables can vary when the preferential statement is interpreted, as for instance

$$\text{if } A = a \text{ then } B = b \succ B = \overline{b} \text{ whatever the value of } C\text{, ceteris paribus (the values of } D\text{, etc. being fixed)}$$

This statement validates for instance the comparison $abcd \succ a\overline{b}\overline{c}d$, but does not validate the comparison $abcd \succ a\overline{b}c\overline{d}$.

The language considered in Wilson (2009) is even more general: the preferential statements allow not only to compare values of single variables but also tuples of values of several variables.

Conditionally lexicographic preferences allow a user to express importance between variables depending on the values assigned to more important variables; they can be represented compactly using *lexicographic preference trees* (Wilson 2009, 2014, 2017; Booth et al. 2010; Liu and Truszczynski 2015), which are closely related to TCP-nets and CP-theories.

Conditional Importance Networks (CI-nets) (Bouveret et al. 2009) allow to express preferences of the form

$$\text{if } A = a \text{ and } B = \overline{b} \text{ then } \{C, D, E\}\text{, together, are more important than } \{F, G\} \text{ together, ceteris paribus.}$$

They are particularly well suited to the expression of preferences between sets of objects in fair division problems with ordinal criteria.

Probabilistic CP-nets (Bigot et al. 2013; Cornelio 2013) aim at expressing compactly a probability distribution over preference relations; uncertainty may come either from the ill-defined context of the comparison between alternatives, or by the fact that the represented preferences are a collective synthesis of individual preferences of a population of agents.

3.7 Elicitation and Learning

One major interest of CP-nets is that when the dependency graph is simple enough, their elicitation is relatively easy: il suffices to ask the user to report her preferences on each of the variables conditionally on the values of its parents (provided that the dependency graph has been learned beforehand, or that it is obvious). The number of queries needed to elicit a CP-net is studied in the framework of learning with queries (Koriche and Zanuttini 2009; Alanazi et al. 2016), whereas the passive learning of CP-nets (from observed comparisons between alternatives) has been studied in Dimopoulos et al. (2009), Lang and Mengin (2009), Liu et al. (2014). See Chevaleyre et al. (2010) for a survey.

3.8 Applications

Constrained optimisation is particularly relevant for configuration problems (see for instance Domshlak et al. 2001 for an application of CP-nets to the personalized configuration of web pages content). An other form of optimisation under constraints can come from the fact that an alternative is feasible if and only if there exists a plan that allows to realize it; in Brafman and Chernyavsky (2005), preferences between states are specified using a TCP-net, and one looks for a plan that results in an optimal alternative, that is, a state α such that no other state reachable from the initial state dominates α. CP-nets have been applied to expressing preferences between documents in information retrieval (Boubekeur et al. 2006). An extension of CP-nets has been applied to recommender systems (Trabelsi et al. 2010).

Beyond individual decision making, CP-nets are particularly well suited to collective decision making on combinatorial domains (see Lang and Xia 2016 for a survey) and to the compact description of players' preferences in noncooperative game theory (Bonzon et al. 2009). An other approach relating CP-nets to games is studied in Apt et al. (2005), where CP-nets are seen as games in normal form and *vice versa*: each player corresponds to a variable of the CP-net, whose domain is the set of the actions available to the player.

4 Graphical Languages and Cardinal Representations of Preferences: Utility Networks

Ordinal graphical models such as CP-nets and TCP-nets provide compact languages to describe ceteris paribus preferences including conditional judgements (the preferences related to group of variables may depend on the values taken by other variables). However, these models do not provide the usual advantages of numerical models based on a utility function u defined over $\mathscr{X}$, such that $x \succsim y \Leftrightarrow u(x) \geq u(y)$ for all $x, y \in \mathscr{X}$. A utility function can easily represent any weak order on a finite set of alternatives; moreover it makes it possible to compare any pair of alternatives by simply computing their respective utilities, and to reduce the search of the preferred alternatives to a utility maximization problem. Last but not least, when the preference information is sufficiently rich, utility functions allow cardinal information to be expressed under the form of *preference intensities* corresponding to utility differences $u(x) - u(y)$, thus providing more information than simple preference orderings. It can be useful to express strength of preferences, but also to better discriminate between groups of alternatives, and to make better decisions under uncertainty. In order to combine the advantages of graphical models like CP-nets with those of utility functions, several graphical languages involving utility functions have been proposed to compactly represent preferences over a set of multiattribute alternatives. For the sake of simplicity, we first introduce additively decomposable utility functions and then the associated graphical models.

4.1 Additively Decomposable Utilities

In order to characterize the utility function of an agent, we need to know the value $u(x)$ of each element x in the set of alternatives $\mathscr{X}$, but this may be sometimes difficult, especially when this set has a combinatorial structure and is defined implicitly. In this case, storing the numbers $u(x),\ x \in \mathscr{X}$ for multiple users would require a prohibitive memory cost. Fortunately, individual preferences over multiattribute objects often have an underlying structure due to independence between groups of attributes, allowing the decomposition of the utility function under a more compact form and the simplification of the elicitation process. The simplest example of such a preference decomposition over a product set $\mathscr{X} = D_1 \times \cdots \times D_n$ is given by the additive utility of the form $u(x) = \sum_{i=1}^{n} u_i(x_i)$ for all $x = (x_1, \ldots, x_n) \in \mathscr{X}$. In this model, we only need to know the marginal utilities $u_i(x_i)$ for all $x_i \in X_i$ to characterize the utility function. However, such a decomposition is not always appropriate because it rules out any possibility of interaction between attributes. When the agent's preferences are more complex, a more sophisticated model is necessary, as illustrated in the following Example:

Example 5 Let $\mathscr{X}$ be a set of menus (X_1, X_2, X_3) where $D_1 = \{\text{meat } (m), \text{ fish } (f)\}$ represents the choice of the main dish, $D_2 = \{\text{red wine } (r), \text{ white wine } (w)\}$ represents the choice of the wine and $D_3 = \{\text{cake } (c), \text{ ice cream } (i)\}$ represents the choice of the dessert.

First Case. Let us suppose that an agent explains her preferences in natural language as follows:

- I always prefer a menu with meat to a menu with fish.
- With meat, I prefer red wine to white wine. This is also the case for fish.
- I prefer the cake to the ice cream, everything else being equal.

These preferences are based on ceteris paribus judgements. The preferences over various possible instances of a given variable X_i characterizing a menu do not depend on the values taken by other variables; such preferences can be elicited independently on each variable. In this simple case, the preferences can be represented by an additive utility $u(x) = u_1(x_1) + u_2(x_2) + u_3(x_3)$ characterized by the following marginal utilities: $u_1(m) = 4; u_1(f) = 0; u_2(r) = 2; u_2(w) = 0; u_3(c) = 1; u_3(i) = 0$. Hence the utilities of the 2^3 possible menus $x^{(i)}$ are:

$$\begin{array}{lll} u(x^{(1)}) = u(m,r,c) = 7; & u(x^{(2)}) = u(m,r,i) = 6; & u(x^{(3)}) = u(m,w,c) = 5; \\ u(x^{(4)}) = u(m,w,i) = 4; & u(x^{(5)}) = u(f,r,c) = 3; & u(x^{(6)}) = u(f,r,i) = 2; \\ u(x^{(7)}) = u(f,w,c) = 1; & u(x^{(8)}) = u(f,w,i) = 0; & \end{array}$$

which leads to the following preferences:

$$x^{(1)} \succ x^{(2)} \succ x^{(3)} \succ x^{(4)} \succ x^{(5)} \succ x^{(6)} \succ x^{(7)} \succ x^{(8)}$$

Second Case. Let us suppose that another agent has the following preferences: $x^{(1)} \succ x^{(2)} \succ x^{(3)} \succ x^{(4)} \succ x^{(7)} \succ x^{(8)} \succ x^{(5)} \succ x^{(6)}$. Note that such preferences are perfectly rational and could be described as follows: (i) a menu with meat is preferred to any menu with fish; (ii) then, the second most important goal is to match the main dish with the wine (red wine with meat, white wine with fish); and (iii) the cake is preferred to the ice cream, everything else being equal.

Although rational, such preferences are not representable by an additive utility function because $x^{(1)} \succ x^{(3)} \Rightarrow u_2(r) > u_2(w)$ but $x^{(7)} \succ x^{(5)} \Rightarrow u_2(w) > u_2(r)$ thus yielding a contradiction. However, it is possible to resort to a less decomposed representation in order to model such preferences. For example we may use $u(x) = u_{1,2}(x_1, x_2) + u_3(x_3)$ with $u_{1,2}(m, r) = 6$, $u_{1,2}(m, w) = 4$, $u_{1,2}(f, w) = 2$, $u_{1,2}(f, r) = 0, u_3(c) = 1, u_3(i) = 0$. This indeed represents the agent's preferences. Note here that the choice of the wine depends on the choice of the main dish but not on the dessert. This explains why we do not need any factor linking wine and dessert.

Third Case. Let us suppose that the preferences of a third agent are: $x^{(2)} \succ x^{(1)} \succ x^{(4)} \succ x^{(3)} \succ x^{(7)} \succ x^{(8)} \succ x^{(5)} \succ x^{(6)}$. Such preferences are similar to those introduced in the second case with a slight sophistication concerning the dessert. The agent prefers the cake to the ice cream when the main dish is fish but this is the opposite when the main dish is meat. In that case, one can see that the previous decomposition does not fit anymore due to the new interaction between attributes X_1 and X_3. Nevertheless, these preferences can be represented by an additively decomposable utility of the form: $u(x) = u_{1,2}(x_1, x_2) + u_{1,3}(x_1, x_3)$, by setting:

$$u_{1,2}(m, r) = 6;\ u_{1,2}(f, w) = 2;\ u_{1,2}(m, w) = 4;\ u_{1,2}(f, r) = 0;$$
$$u_{1,3}(m, c) = 0;\ u_{1,3}(m, i) = 1;\ u_{1,3}(f, c) = 1;\ \ u_{1,3}(f, i) = 0.$$

One could object here that the latter representation is not more compact than the extensive representation (8 utility values must be stored in both cases) but this is due to the small size of the domain of variables X_i. Assuming for instance that the cardinality of X_i is m for every i, the above utility decomposition requires to store $2m^2$ values instead of m^3, which saves some memory space as soon as $m > 2$; of course the space saving becomes more important as m increases.

Such a decomposition of the utility function as the sum of overlapping factors is named GAI (Fishburn 1970; Bacchus and Grove 1995), where GAI refers to the Generalized Additive Independence axiom satisfied by any preference represented by such a decomposition. GAI utilities include additive and multilinear decompositions as special cases, but they are more flexible since they allow some interactions between attributes without making any a priori assumption on the form of these interactions. More precisely, GAI decompositions can formally be introduced as follows:

Definition 1 (*GAI-decomposable utility*) Let $\mathscr{X} = \times_{i=1}^{n} D_i$. Let $C_1, \ldots, C_k$ be k subsets of $N = \{1, \ldots, n\}$ such that $N = \bigcup_{i=1}^{k} C_i$. For all i, let $D_{C_i} = \times_{j \in C_i} D_j$; in other words D_{C_i} is the the product set of attributes domains associated to the variables

of C_i. The utility function $u(\cdot)$ representing $\succsim$ is GAI-decomposable with respect to subsets D_{C_i} if and only if there exists k functions $u_i : D_{C_i} \mapsto \mathbb{R}$, $i = 1, \ldots, k$, such that:

$$u(x) = \sum_{i=1}^{k} u_i(x_{C_i}), \quad \forall x = (x_1, \ldots, x_n) \in \mathscr{X},$$

where x_{C_i} is the n-tuple formed by x_j, $j \in C_i$.

4.2 Graphical Models Associated with a Decomposable Utility Function

Graphical representations of GAI-decomposable utility functions are named *Utility Networks*. Different variants of utility networks have been proposed for the compact representation of GAI-utilities and we introduce them below.

4.2.1 UCP-Nets

A UCP-net is an extension of a CP-net allowing a compact encoding of a GAI utility function representing ceteris paribus preferences (Boutilier et al. 2001). Like CP-nets, UCP-nets are based on directed dependency graphs, but preferences are measured by utilities. The conditional preference tables of CP-nets are here replaced by local utility tables. Considering the third case of Example 5 mentioned above, we can represent the dependence structure between variables by a CP-net containing the edge $X_1 \rightarrow X_2$ to model the fact that the choice of the wine (X_1) depends on the main dish (X_2), and on the other hand, X_3 is disconnected from the rest of the graph to express that the choice of the dessert is independent of the other variables characterizing the menu. A convenient GAI decomposition for this graph is: $u(X_1, X_2, X_3) = v_1(X_1) + v_{12}(X_1, X_2) + v_3(X_3)$ with $v_1(m) = 4$, $v_2(f) = 0$, $v_{12}(m, r) = v_{12}(f, w) = 2$ and $v_3(c) = 1$ and $v_{12}(i) = 0$ which is represented by the following UCP-net:

Remark that the function u used in the UCP-net corresponds exactly to the utility function introduced in the second case of Example 5 under the form $u(X_1, X_2, X_3) = u_{12}(X_1, X_2) + u_3(X_3)$. The correspondence easily appears by setting $u_{12}(X_1, X_2) = v_1(X_1) + v_{12}(X_1, X_2)$ and $u_3(X_3) = v_3(X_3)$. Swapping fish for meat on attribute X_1 saves 4 points, a decisive advantage that cannot be compensated by another swap on variable X_2 or X_3. Moreover, the order induced by the utility function over menus refines the partial order induced by the underlying CP-net, allowing some incomparabilities to be ruled out. Beyond this example, one can generally define a UCP-net as follows:

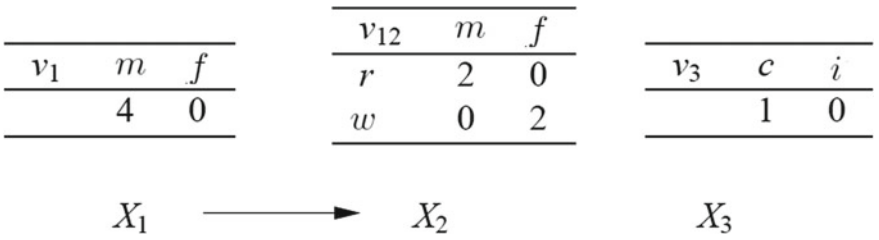


Fig. 8 An example of UCP-net

Definition 2 Let $u(X_1, \ldots, X_n)$ be a utility function representing $\succsim$ the preference of the Decision Maker. A UCP-net for u (or UCP network) is characterized by a directed acyclic graph G over variables $X_1, \ldots, X_n$ and an additive decomposition of $u(X_1, \ldots, X_n)$ into factors $u_i(X_i|P(X_i))$ representing the utility of X_i given its parents $P(X_i)$ in the graph, in such a way that:

- $u(X_1, \ldots, X_n) = \sum_{i=1}^{n} u_i(X_i|P(X_i))$;
- G is the directed graph associated with $\succsim$ in the sense of Sects. 3.1 and 3.2: w.r.t. $\succsim$, every variable X_i is independent of the other variables in G, conditionally to its parents (see Sects. 3.1 and 3.2)

$$\text{for all } x_1, x_2 \in D_i, \forall y \in D_{P(X_i)}, \text{ for all } z_1, z_2 \in D_{N\setminus\{i\cup P(X_i)\}},$$
$$\text{we have } x_1 y z_1 \succsim x_2 y z_1 \text{ if and only if } x_1 y z_2 \succsim x_2 y z_2.$$

In the example of UCP-net represented on Fig. 8, the utility decomposition defined by $u(X_1, X_2, X_3) = v_1(X_1) + v_{12}(X_1, X_2) + v_3(X_3)$ matches the definition since X_1 and X_3 have no parent and the factor $v_{12}(X_1, X_2)$ provides the utility of X_2 given X_1, playing the role of $u_2(X_2|X_1)$. The decomposability property of the utility function required in this definition makes u a GAI decomposable function, compatible with the underlying CP-net. The structure imposed by the underlying CP-net is constraining but it has the advantage of simplifying the elicitation process, especially when the CP-net graph is acyclic (it is sufficient to start the elicitation process with variables that do not have any parent in the graph, and then to continue with their descendants ordered according to the dependency graph of the CP-net). Let us remark however that some GAI decomposable utility functions cannot be represented by a UCP-net. We present below another graphical representation that fits to any GAI-decomposable utility function.

4.2.2 GAI Networks

GAI decompositions can be represented by non-directed graphical structures named *GAI networks* (Gonzales and Perny 2004) or *GAI-nets*. Such structures are similar to junction graphs used for Bayesian Networks see Jensen and Graven-Nielsen (2007) and chapter "Languages for Probabilistic Modeling over Structured and Relational

Domains" of volume 2. Roughly speaking, this is a graph composed of one or several trees whose nodes correspond to the factors of the GAI decomposition, where an edges connecting a pair of nodes corresponds to the presence of a factor having at least one variable in common. Typically, in the third case of Example 5, a convenient GAI network would be a graph with two nodes corresponding to two factors $u_{1,2}(x_1, x_2)$ with variables $\{X_1, X_2\}$, and $u_{1,3}(x_1, x_3)$ with variables $\{X_1, X_3\}$. These two nodes are connected by an edge labelled by variable X_1 linking the two factors. More generally, a GAI-network is defined as follows:

Definition 3 (*GAI-network*) Let $X = \times_{i=1}^{n} X_i$. Let $C_1, \ldots, C_k$ be k subsets of $N = \{1, \ldots, n\}$ such that $N = \bigcup_{i=1}^{k} C_i$. Let us assume that $\succsim$ is representable by a GAI utility $u(x) = \sum_{i=1}^{k} u_i(x_{C_i})\ \forall x \in X$. A GAI network representing $u(\cdot)$ is a non-directed graph $\mathscr{G} = (\mathscr{C}, \mathscr{E})$ satisfying the following properties:

1. $\mathscr{C} = \{X_{C_1}, \ldots, X_{C_k}\}$;
2. if $(X_{C_i}, X_{C_j}) \in \mathscr{E}$ then $C_i \cap C_j \neq \emptyset$.
3. for all X_{C_i}, X_{C_j} such that $C_i \cap C_j = T_{ij} \neq \emptyset$, there exists a path $\mathscr{G}$ linking X_{C_i} and X_{C_j} such that all its nodes include all the indices in T_{ij} (Running intersection property).

The nodes of $\mathscr{C}$ are called *cliques*. Every edge $(X_{C_i}, X_{C_j}) \in \mathscr{E}$ is labelled by $X_{T_{ij}} = X_{C_i \cap C_j}$ and is called a *separator*.

The cliques are represented by ellipses and the separators by rectangles. Here, we only consider acyclic GAI networks. As recalled and illustrated in Gonzales and Perny (2004), this is not restrictive since general GAI networks can always be recompiled into acyclic GAI networks grouping some factors to eliminate cycles. In a GAI network, any edge connecting two cliques reflects a non-empty intersection between the sets of attributes present in the two cliques. The intersection being a commutative operation, it is convenient to define a GAI net as a non-directed graph. This differs from UCP nets where dependencies between factors are conditional and justify the use of directed graphs. Let us give an example of GAI network derived from Gonzales and Perny (2005).

Example 6 If we consider the following utility function defined using 7 attributes:

$$u(A, B, C, D, E, F, G) = u_1(A, B) + u_2(C, E) + u_3(B, C, D) + u_4(B, D, F) + u_5(B, G)$$

then, as shown in Fig. 9, the cliques are: AB, CE, BCD, BDF et BG. By Property 2 of Definition 3, the set of edges of a GAI network can be determined using algorithms preserving the running intersection property (see Cowell et al. 1999).

This running intersection property is very useful because it allows an easy identification of conditional independencies between variables by looking at the separators. In the above example (Fig. 9), the separators are the groups of variables appearing in squares. If the variables of a separator are instantiated, one necessarily divides the GAI network (which is a tree) into several connected components which become preferentially independent (conditionally to the instantiation). This can be used to

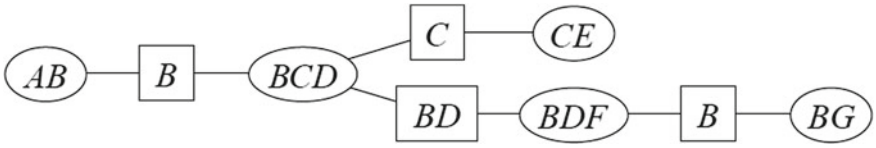


Fig. 9 A GAI tree

elicit some utility tables without taking the rest of the graph into account. This can also be useful to perform optimization; one can indeed optimize over some variables of the networks without taking the other variables into account. Let us give an illustration using the example of Fig. 9. If separator B is instanciated into b, on can see that $u(A, b, C, D, E, F, G)$ can be decomposed into two independent factors, namely $u_1(A, b)$ and $u_2(C, E) + u_3(b, C, D) + u_4(b, D, F) + u_5(b, G)$, with no common variable. Given that $B = b$ one can therefore elicit the preferences over A without taking care of the other variables. Similarly, in optimization, one can optimize the value of A conditionally to each possible value of B without taking care of the other variables. These principles which are in the core of GAI networks are largely used in elicitation and optimization algorithms (Gonzales and Perny 2004, 2005; Braziunas and Boutilier 2005). In particular, when one wants to determine the tuple of maximal utility, one can resort to a non-serial dynamic programming (Bistarelli et al. 1999) based on a variable elimination sequence (Koller and Friedman 2009); this type of algorithm is exponential in the treewidth of the GAI tree, defined as the size of the largest clique of the graph. Knowing that agents are rarely able to express interactions involving more than two or three variables in practical cases, the utility factors of a GAI decomposition are generally relatively small, which allows fast optimizations.

For the sake of illustration, we give below an example of optimization performed using the GAI network of Fig. 9 where $A = \{a^0, a^1, a^2\}$, $B = \{b^0, b^1\}$, $C = \{c^0, c^1\}$, $D = \{d^0, d^1\}$, $E = \{e^0, e^1, e^2\}$, $F = \{f^0, f^1\}$, $G = \{g^0, g^1\}$, with the utility tables given in Fig. 10. Determining the optimal tuple amounts to solving the following problem: $\max_{a,b,c,d,e,f,g} u_1(a, b) + u_2(c, e) + u_3(b, c, d) + u_4(b, d, f) + u_5(b, g)$. The following properties can be used to efficiently solve this problem.

1. computing $\max\ u(X_1, \ldots, X_n)$ over the set of variables $X_1, \ldots, X_n$ can be decomposed into $\max_{X_1} \max_{X_2} \ldots \max_{X_n} u(X_1, \ldots, X_n)$ and the order of variables in the sequence has no importance;
2. if $u(X_1, \ldots, X_n)$ can be decomposed into $f() + g()$ where $f()$ does not depend on variable X_i, then $\max_{X_i}[f() + g()] = f() + \max_{X_i} g()$;
3. in a GAI network, the running intersection property guarantees that a variable present in an outer clique X_C and not in the neighbor clique of X_C will be absent of any other clique of the GAI network.

In order to determine an optimal tuple, Properties 2 and 3 suggest a strategy consisting first in maximizing over variables appearing only in outer cliques; the results are then transmitted to the neighbor cliques and the outer cliques are removed. This process is iterated from outer cliques to inner cliques until all cliques are removed.

Fig. 10 Utility tables for $u(\cdot)$

$u_1(a,b)$	b^0	b^1
a^0	8	2
a^1	4	3
a^2	1	7

$u_2(c,e)$	e^0	e^1	e^2
c^0	6	3	5
c^1	3	4	0

$u_3(b,c,d)$	b^0		b^1	
	d^0	d^1	d^0	d^1
c^0	0	2	7	1
c^1	5	1	2	4

$u_4(b,d,f)$	b^0		b^1	
	f^0	f^1	f^0	f^1
d^0	4	2	5	8
d^1	3	8	9	0

$u_5(b,g)$	g^0	g^1
b^0	0	9
b^1	6	4

Let us illustrate this iterative optimization process on the running example. The optimization problem:

$$\max_{b,c,d} [u_3(b,c,d) + \max_f [u_4(b,d,f) + \max_g u_5(b,g)] + [\max_e u_2(c,e)] + [\max_a u_1(a,b)]] \quad (1)$$

is solved using the following operations:

1. in the clique AB, compute $u_1^*(b) = \max_{a \in A} u_1(a,b)$ for all $b \in B$;
2. in the clique CE, compute $u_2^*(c) = \max_{e \in E} u_2(c,e)$ for all $c \in C$;
3. in the clique BG, compute $u_5^*(b) = \max_{g \in G} u_5(b,g)$ for all $b \in B$;
4. in the clique BDF, substitute $u_4(b,d,f)$ by $u_4(b,d,f) + u_5^*(b)$ for all tuples $(b,d,f) \in B \times D \times F$. Then, compute $u_4^*(b,d) = \max_{f \in F} u_4(b,d,f)$ for all tuples $(b,d) \in B \times D$;
5. in the clique BCD, substitute $u_3(b,c,d)$ by $u_3(b,c,d) + u_1^*(b) + u_2^*(c) + u_4^*(b,d)$ for all tuples $(b,c,d) \in B \times C \times D$. Then, compute $\max_{b,c,d} u_3(b,c,d)$, the maximal utility (we obtain 34).

Figure 11 shows the content of u_i^* and u_i after substitution. At the end of step 5 we obtain the maximal utility value (here 34), defined by Eq. (1). At the end of the collect phase, the optimal value of u over X is known. In order to determine an optimal tuple corresponding to this value (i.e., an optimal solution) we resort to an instantiation and diffusion phase that consists in retropropagating the arguments achieving the maximum at any step, in the reverse order of variables as the one used for the collect phase. Thus, at the last step of the collect phase, one can see that utility 34 for u_3 corresponds to the tuple (b^1, c^0, d^0), which entails that, in the optimal tuple, we have $B = b^1$, $C = c^0$, $D = d^0$. At step 4, $u_4^*(b^1, d^0)$ corresponds to $u_4(b^1, d^0, f^1) = 14$ which implies that $F = f^1$. Then, at step 5, one can see that $u_5^*(b^1) = 6$ corresponds to $u_5(b^1, g^0)$ and therefore $G = g^0$, which completes the characterization of the optimal tuple. We finally obtain $(a^2, b^1, c^0, d^0, e^0, f^1, g^0)$ as an optimal solution (Fig. 12).

The algorithm used for the optimisation of GAI functions is similar to the algorithm used for computing the most plausible explanation in a Bayesian network (Nilsson 1998). This algorithm is founded on the principle of non-serial dynamic programming (Bertele and Brioschi 1972); it is also used to solve valued constraint

	b^0	b^1
$u_1^*(b)$	8	7

	c^0	c^1
$u_2^*(c)$	6	4

	b^0	b^1
$u_5^*(b)$	9	6

$u_4(b,d,f)$	b^0		b^1	
	f^0	f^1	f^0	f^1
d^0	13	11	11	14
d^1	12	17	15	6

$u_4^*(b,d)$	b^0	b^1
d^0	13	14
d^1	17	15

$u_3(b,c,d)$	b^0		b^1	
	d^0	d^1	d^0	d^1
c^0	27	33	34	29
c^1	30	30	27	30

Fig. 11 Content of u_i^* and u_i after the substitutions

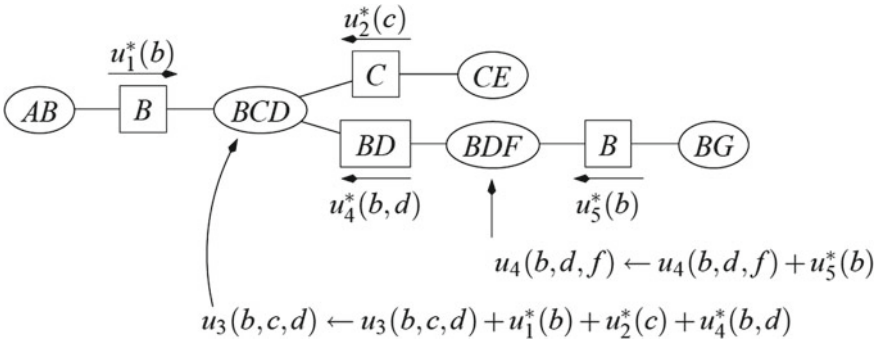


Fig. 12 Steps 1 to 5 to obtain the optimal utility

satisfaction problems (Bistarelli et al. 1999). This is not surprising because there is a clear similarity between GAI networks, junction trees used in Bayesian networks and hypergraphs resulting from the triangulation of a cost constraint network. Whether we want to optimize a multiattribute utility function, a joint probability law or the overall cost induced by a constraint network, in all cases we use a decomposition of the objective function into factors whose scopes include only a subset of variables, and it is this decomposition which is efficiently exploited in graphical models to perform the optimization efficiently. One can however highlight some specificities of GAI networks compared to the two other types of networks:

- the number of variables in utility networks is generally reasonably small (the variables represent here the descriptive attributes of the alternatives). Moreover, the factors used in the decomposition often include only a limited number of attributes; interactions between attributes do not involve many attributes due to cognitive limitation of agents. This is an important difference with constraint networks that may have to deal with global constraints whose scope includes all variables. In Bayesian networks, there also may exist large sets of interdependent variables, making the cliques larger than in GAI networks.
- the problems addressed in utility networks sometimes differ from those considered in other types of networks. For example, preference elicitation is a different exercise than learning a Bayesian network or a constraint network, even if there are some bridges between the two problems. Moreover, utility networks have multia-

gent or multicriteria extensions which lead to consider complex utility functions that are no longer additively decomposable. In such cases, the variable elimination process does not apply directly and requires the implementation of more complex optimization algorithms (Gonzales et al. 2008, 2011).

As mentioned above, GAI networks are non-directed graphs that seem natural to handle preferences represented by a GAI decomposable utility function. However, one may wonder if it would not be more informative to resort to a directed graph, as for UCP nets, to specify which group of variables depends on the other group in an interaction and to simplify the elicitation process. In order to benefit both from the advantages of GAI networks and those of CP-nets, one could introduce another type of directed graph, closer to a Bayesian network than to a junction tree. For example, when the preferences over the instances of X_2 depend on the instance of X_1, we may decompose the GAI factor $u_{12}(X_1, X_2)$ into the sum $v_1(X_1) + v_{12}(X_1|X_2)$ and thus introduce an edge $X_1 \rightarrow X_2$ to represent this dependency. The resulting directed graph is similar to a Bayesian network excepted that utility tables replace probability tables and are aggregated using a sum instead of a product. This idea has been proposed by Brafman and Engel (2009). Structurally, the resulting networks are strongly related to GAI networks since one can pass from these directed graphs to GAI networks as we pass from Bayesian nets to junction trees. Finally, an interesting variant of GAI networks, called CUI networks, has been investigated by Engel and Wellman (2008) with the idea of relaxing GAI independence into a weaker independence condition.

5 Logical Languages

We now give a briefer overview of the main classes of preference representation languages based on logic. Some of these languages are not logics on their own but rather make use of logic. Preferences are expressed in these languages by means of logical formulas expressing goals, associated with ordinal or cardinal labels expressing their importance. We give an overview of these languages in the first part of this section. In the second part we give a brief overview of *preference logics*. From now on, the set of variables $\mathscr{V}$ is a set of propositional symbols, and the set of alternatives $\mathscr{X}$ is the set of possible worlds associated with $\mathscr{V}$, namely $2^{\mathscr{V}}$. As in the rest of the chapter, alternatives are denoted by $\mathbf{x}$, $\mathbf{y}$ etc.

5.1 Logics, Priorities and Weights

5.1.1 Weights

Numerical preferences over a combinatorial domain composed of binary variables can be represented by means of *weighted propositional logic formulas*. Existing logics differ from each other on the interpretation of weights associated with formulas.

A set of weighted formulas is a set of pairs

$$G = \{\langle \varphi_1, a_1 \rangle, \ldots, \langle \varphi_n, a_n \rangle\},$$

where φ_i is a propositional logic formula and a_i is a real number representing how much the satisfaction of φ_i contributes to the utility of the agent. In general, the utility function u_G is defined as the aggregation of weights associated with non-satisfied formulas: for every $\mathbf{x} \in \mathscr{X} = 2^{\mathscr{V}}$,

$$u_G(\mathbf{x}) = -F(\{a_i \mid \mathbf{x} \models \neg\varphi_i\})$$

where F is a function from $\mathbb{R}^+ \times \mathbb{R}^+$ to $\mathbb{R}^+$, nondecreasing, commutative and associative.[4]

A first choice consists in interpreting the weights as penalties, namely the price to pay if the formula is not satisfied (Haddawy and Hanks 1992; Dupin de Saint-Cyr et al. 1994; Pinkas 1995; Lafage and Lang 2000). The more important the goal expressed by a formula, the higher the associated weight. Therefore, given a weighted formula (φ, a), an alternative $\mathbf{x}$ has a penalty 0 if it satisfies φ and a penalty a if it falsifies φ. The weights associated to formulas are aggregated additively: given a set of weighted formulas $\Delta = \{\langle \varphi_i, a_i \rangle | i = 1, \ldots, n\}$, the penalty associated to an alternative is the sum of its penalties w.r.t. each formula of Δ:

$$\forall \mathbf{x} \in \mathscr{X}, p_\Delta(\mathbf{x}) = \sum \{a_i | \langle \varphi_i, a_i \rangle \in \Delta, \mathbf{x} \not\models \varphi_i\}.$$

The penalty of an alternative is its disutility, that is, $u_\Delta(\mathbf{x}) = -p_\Delta(\mathbf{x})$.

Example 7 (*Example* 1 *continued*) Suppose that the user's preferences are represented by means of the following penalty base:

$$\Delta = \{\langle d, 100 \rangle, \langle d \rightarrow s, 30 \rangle, \langle n \rightarrow \neg s, 30 \rangle, \langle d \wedge s \rightarrow c_1, 5 \rangle, \langle n \vee \neg s \rightarrow c_2, 5 \rangle\}$$

These preferences stand for:

- taking a day flight is important (the associated penalty is 100),
- making a stopover for a day flight, and symmetrically, not making a stopover for a night flight, is important too, but less so (the associated penalty is 30),
- the choice of the company is much less important (penalty 5): the agent has a slight preference for traveling with c_1 for a day flight with a stopover and to travel with c_2 otherwise.

[4] Equivalently, one can define the utility function of an alternative by aggregating the weights of all *satisfied* formulas, that is,

$$u_G(\mathbf{x}) = F(\{a_i \mid \mathbf{x} \models \varphi_i\}).$$

The penalty (or disutility) degrees associated to alternatives are the following:

$$p_\Delta(n\bar{s}c_1) = 105 \quad p_\Delta(n\bar{s}c_2) = 100 p_\Delta(nsc_1) = 135 \quad p_\Delta(nsc_2) = 130,$$
$$p_\Delta(d\bar{s}c_1) = 35 \quad p_\Delta(d\bar{s}c_2) = 30 p_\Delta(dsc_1) = 0 \quad p_\Delta(dsc_2) = 5.$$

Uckelman et al. (2009) give a thorough study of this representation language: they compare various sublanguages corresponding to specific syntactical restrictions on the allowed formulas and specific restrictions on the allowed weights, according to their expressivity, their succinctness, and the computational complexity of the main reasoning tasks.

The additive behaviour of penalty logic allows to express compensation between goals: the non-satisfaction of several less important goals might be compensated by the non-satisfaction of a more important goal. These valuation systems can be compared to those used in the GAI utility functions presented in Sect. 4: each weighted formula can be seen as a factor of the function GAI which takes the value 0 or the penalty depending on whether the formula is satisfied or not.

The choice of $F = \max$ leads to define the disutility of an alternative as the importance of the most important goal that it satisfies. This choice is equivalent to (up to a transformation of the utility scale) possibilistic logic (see chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" of this Volume). Its expressivity, succinctness and complexity as a preference representation language have been studied in Uckelman and Endriss (2008).

Example 8 (*Example 7 continued*) Suppose that user's preferences given in Example 7 are represented by Δ defined as follows:

$$\Delta = \{\langle d, 9\rangle, \langle d \to s, 7\rangle, \langle n \to \neg s, 7\rangle, \langle d \wedge s \to c_1, 3\rangle, \langle n \vee \neg s \to c_2, 3\rangle\}$$

The utility function induced by Δ is then:

$$u_\Delta(n\bar{s}c_1) = u_\Delta(n\bar{s}c_2) = u_\Delta(nsc_1) = u_\Delta(nsc_2) = -9,$$
$$u_\Delta(d\bar{s}c_1) = u_\Delta(d\bar{s}c_2) = -7, u_\Delta(dsc_1) = 0, u_\Delta(dsc_2) = -3.$$

and the associated preference relation is: $dsc_1 \succ_\pi dsc_2 \succ_\pi d\bar{s}c_1 \sim_\pi d\bar{s}c_2 \succ_\pi n\bar{s}c_1 \sim_\pi n\bar{s}c_2 \sim_\pi nsc_1 \sim_\pi nsc_2$.

Weighted logics have been used for the compact expression of utility functions in fair division (Bouveret and Lang 2008), bidding in combinatorial auctions (Boutilier and Hoos 2001; Nisan 2006) and more generally collective decision making (Greco and Lang 2015).

A similar language, called *marginal contribution nets* (Ieong and Shoham 2005) has been developed with the specific aim of expressing compactly coalitional value functions in cooperative games. Marginal contribution nets have been used for expressing compactly preferences in coalition structure generation (Ohta et al. 2009; Rahwan et al. 2015) and hedonic games (Elkind and Wooldridge 2009), while the

impact of compact representation by MC-nets for computing power indices in cooperative games has been studied in Ieong and Shoham (2005), Elkind et al. (2009).

5.1.2 Priorities

The ordinal counterpart of weighted logical languages is the family of *logical languages with priorities*. A *set of preferences with priorities* Δ is a n-uple $\langle G_1, \ldots, G_n \rangle$, where G_i is a multi-set of propositional formulas with priority i. By convention, G_1 contains the highest priority goals and G_n those of lowest priority. We denote by $sat(\mathbf{x}, G_i)$ and $nonsat(\mathbf{x}, G_i)$, respectively, the sets of formulas in G_i satisfied and nonsatisfied by $\mathbf{x}$, that is,

$$sat(\mathbf{x}, G_i) = \{\varphi \in G_i \mid \mathbf{x} \models \varphi\}$$

and

$$nonsat(\mathbf{x}, G_i) = G_i \setminus sat(\mathbf{x}, G_i)$$

The aim is to define a weak order over $\mathscr{X}$ from Δ, i.e. from a weak order *over formulas* define a preorder *over alternatives*. The usual choices (see Brewka 1989; Benferhat 1993; Lehmann 1995) are:

"Best-out" Alternatives are compared according to the priority level of their most important non-satisfied goals: Let $\rho(\mathbf{x}, \Delta) = \min\{i, nonsat(\mathbf{x}, G_i) \neq \emptyset\}$.

$$\mathbf{x} \geq_{\Delta}^{best-out} \mathbf{x}' \text{if and only if } \rho(\mathbf{x}, G_i) \geq \rho(\mathbf{x}', G_i)$$

This is equivalent to the choice of $F = \max$ in the previous section and doesn't bring anything new. Moreover it suffers from the so-called "drowning" effect: the presence of a non-satisfied goal with priority i inhibits the effect of all goals with degree $j > i$ (as well as all other goals with degree i). The following two refinements of "best-out" prevent the drowning effect.

"Discrimin" Two alternatives are compared according to the most important goals that one alternative satisfies but not the other one.

$$\mathbf{x} >_{\Delta}^{discrimin} \mathbf{x}' \text{ if and only if } \exists i \leq n \text{ such that } \begin{pmatrix} sat(\mathbf{x}, G_i) \supset sat(\mathbf{x}', G_i) \\ \forall j \leq i, sat(\mathbf{x}, G_j) = sat(\mathbf{x}', G_j) \end{pmatrix}$$

$$\mathbf{x} \sim_{\Delta}^{discrimin} \mathbf{x}' \text{ if and only if } \forall i \leq n, sat(\mathbf{x}, G_i) = sat(\mathbf{x}', G_i), \text{ and}$$

$$\mathbf{x} \geq_{\Delta}^{discrimin} \mathbf{x}' \text{ if and only if } \mathbf{x} >_{\Delta}^{discrimin} \mathbf{x}' \text{ ou } \mathbf{x} \sim_{\Delta}^{discrimin} \mathbf{x}'$$

"Leximin" Two alternatives are compared by first identifying the highest priority level for which they do not satisfy the same number of goals. Then the alternative which satisfies more goals at this level is preferred. Let us denote $\#sat(\mathbf{x}, G_i)$ the cardinality of $sat(\mathbf{x}, G_i)$, namely the number of goals of level i satisfied by $\mathbf{x}$.

$$\mathbf{x} >_{\Delta}^{leximin} \mathbf{x}' \text{ if and only if } \exists k \geq 1, \left(\begin{array}{l} \text{such that} \\ \text{(i)} \#sat(\mathbf{x}, G_k) > \#sat(\mathbf{x}', G_k) \\ \text{(ii)} \forall j < k,\ \#sat(\mathbf{x}, G_j) = \#sat(\mathbf{x}', G_j) \end{array} \right)$$

$$\mathbf{x} \sim_{\Delta}^{leximin} \mathbf{x}' \text{ if and only if } \forall i \leq n, \#sat(\mathbf{x}, G_i) = \#sat(\mathbf{x}', G_i)$$

$$\mathbf{x} \geq_{\Delta}^{leximin} \mathbf{x}' \text{ if and only if } \mathbf{x} >_{\Delta}^{leximin} \mathbf{x}' \text{ or } \mathbf{x} \sim_{\Delta}^{leximin} \mathbf{x}'$$

An equivalent expression of this criterion consists in defining the vector $\mathbf{s}_{\Delta}(\mathbf{x}) = \langle \#sat(\mathbf{x}, G_1), \ldots, \#sat(\mathbf{x}, G_n) \rangle$ and compare $\mathbf{s}_{\Delta}(\mathbf{x})$ and $\mathbf{s}_{\Delta}(\mathbf{x}')$ according to the lexicographic order. $\geq_{\Delta}^{leximin}$ is a total preorder.

We also have the following implications:

- $\mathbf{x} >_{\Delta}^{bestout} \mathbf{x}' \Rightarrow \mathbf{x} >_{\Delta}^{discrimin} \mathbf{x}' \Rightarrow \mathbf{x} >_{\Delta}^{leximin} \mathbf{x}'$;
- $\mathbf{x} \geq_{\Delta}^{discrimin} \mathbf{x}' \Rightarrow \mathbf{x} \geq_{\Delta}^{leximin} \mathbf{x}' \Rightarrow \mathbf{x} \geq_{\Delta}^{bestout} \mathbf{x}'$.

Lastly Brewka (2002) proposes a novel logical connector, the noncommutative disjunction $\otimes$, where $\varphi \otimes \psi$ reads "I would like to see φ satisfied, and if it is not, I would like to see ψ satisfied". Brewka (2004) goes further and proposes a more expressive representation language permitting the coexistence of different interpretation criteria of priorities in the same set of preferences.

5.2 *Preference Logics*

In the previous part, presented formalisms make use of propositional logic but are not *preference logics* in the following sense: a preference logic consists of a semantics and/or a formal system conceived to reason on dyadic preferences between propositional formulas.

Although an important part of the literature on preference logics only lies at the margin of Artificial Intelligence, this research topic has been addressed in a huge number of AI journals and conferences so that it deserves to be presented in this chapter. We first briefly present a large family of preference logics constructed on the basis of ceteris paribus principle to interpret preferences over propositional formulas. Then we present another large family of preference logics that are all based on *conditional logics*.

5.2.1 Ceteris Paribus Preferences

When an individual expresses a preference in natural language such as *I prefer an apartment on the sixth floor to an apartment on the ground floor*, she does not express that she prefers any apartment on the sixth floor to any apartment on the ground floor. The principle at use in the interpretation of such a preferential statement is that alternatives should be compared *everything else being equal* (ceteris paribus), or more generally, all properties irrelevant to the preferential statement at hand being equal.

Interpreting a statement of the form "φ is preferred to ψ", that we formally write as $\varphi \rhd \psi$, is simple when φ and ψ are "complete" formulas (each is satisfied by a unique alternative): if φ and ψ respectively correspond to alternatives $\mathbf{x}$ and $\mathbf{x}'$, then the preferential statement naturally corresponds to $\mathbf{x} \succ \mathbf{x}'$. Now, preferences expressed by individuals do not always refer to single alternatives but often to formulas representing *sets of alternatives* that are generally not singletons, neither disjoint sets. Therefore an individual may express a statement like *I prefer ice cream to cake*, even there exist several kinds of ice creams and cakes and if it is permitted to have both at the same time. This statement is generally equivalent to the statement *I prefer an ice cream and no cake to a cake and no ice cream* (Halldén 1957; von Wright 1963): $\varphi \rhd \psi$ may then be written as $\varphi \wedge \neg\psi > \neg\varphi \wedge \psi$,[5] where $>$ expresses a comparison between mutually exclusive (or contradictory) formulas. Lastly, we can consider *contexts* with conditional preferences: if γ is a propositional formula, $\gamma : \varphi \rhd \psi$ expresses that the preference of φ over ψ applies only when γ is true. Therefore we can rewrite $\gamma : \varphi \rhd \psi$ into $\gamma \wedge \varphi \rhd \gamma \wedge \psi$.

Then we have to specify how preference between contradictory formulas ($\varphi > \psi$, where $\varphi \wedge \psi$ is inconsistent) is related to preference over alternatives. A particularly intuitive principle, which goes back to von Wright (1963), is the ceteris paribus interpretation: $\varphi \rhd \psi$ is interpreted as *everything else being equal, I prefer an alternative satisfying* $\varphi \wedge \neg\psi$ *to an alternative satisfying* $\psi \wedge \neg\varphi$. Now we have to formally define the notion of *all else being equal*. The case where φ and ψ are opposed literals ($\varphi = p$ and $\psi = \neg p$, or *vice versa*) is simple: $\mathbf{x}$ and $\mathbf{x}'$ are identical ceteris paribus if they give the same valuation to all propositional symbols other than p. When φ and ψ are complex formulas, the interpretation of $\varphi > \psi$ is less obvious; several definitions have been proposed and studied both in the literature on philosophical logic (see for example von Wright 1972; Hansson 2001; Roy et al. 2009) and that of artificial intelligence (see for example Doyle and Wellman 1991; Doyle et al. 1991; Tan and Pearl 1994).

These logics share with CP-nets and their extensions the ceteris paribus principle for interpreting preferential statements. We can show that these graphical compact preference representation languages, as well as prioritized goals presented in Sect. 5.1.2, correspond to particular fragments of preference logics that are sufficiently expressive (Roy et al. 2009; Bienvenu et al. 2010).

5.2.2 Defeasible Preferences and Conditional Logics

Preferences are sometimes expressed with respect to a context which is more or less specific. Consider for example the following preferential statements:

1. *I prefer to commute by bike*;
2. *If there is a storm, I prefer to commute by metro.*

[5] This principle must be modified in the extreme case where φ is a logical consequence of ψ or *vice versa*—see Hansson (2001).

The preferential statement 1 is *defeasible*, or a *default* preference: 1 applies not only when we know that there is no storm, but also when no weather information is provided: in this case, we jump to the conclusion that the world is *normal*. However, upon further learning that a storm is predicted, 2, which is more specific than 1, is triggered and takes priority over 1. It is worth noticing that 1 and 2 are not contradictory. They should be read as: *normally, I prefer commute by bike, except when there is a storm*. Reasoning on such preferences is *nonmonotonic*: The application of a preferential statement can be revised in the light of a more specific information. This kind of reasoning has been widely addressed in Artificial Intelligence, especially for reasoning about beliefs (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" this of Volume). This principle, which consists in assuming that the world is among the most normal possible worlds given our current beliefs, allows a concise and modular description of preferences: concise because an economy of representation is made possible by avoiding to explicitly specify all exceptional circumstances in which a preferential statement is not applied; and modular because a set of such preferential statements can be completed at any moment without necessarily entirely reconsidering the description of preferences. Therefore we can add to the previous preferential statement the following statement:

3. *if there is an earthquake then I prefer commute by bike, even if there is a storm.*

This statement will have precedence over statement 2 in the double exceptional circumstance *there is a storm and an earthquake*.

The formalization of contextual and defeasible preferences use *conditional logics*. In order to simplify the presentation of these logics, we suppose that preference statements are expressed between two opposite formulas: $P(\psi > \neg\psi \mid \varphi)$, or for short, $P(\psi|\varphi)$, expresses that "in the context φ, ψ is preferred to $\neg\psi$". This preferential statement means that the set of alternatives satisfying $\varphi \wedge \psi$ is preferred to alternatives satisfying $\varphi \wedge \neg\psi$. What remains to be done is to give a precise meaning to "a set of alternatives is preferred to another set of alternatives".

Let $\succsim$ be a preference relation over $\mathscr{X}$.

- $\succsim$ satisfies $P(\psi|\varphi)$ following the *optimistic* semantics iff $\exists \mathbf{x} \models \varphi \wedge \psi$, $\forall \mathbf{x}' \models \varphi \wedge \neg\psi$, we have $\mathbf{x} \succ \mathbf{x}'$ (Pearl 1990).
- $\succsim$ satisfies $(\psi|\varphi)$ following the *pessimistic* semantics iff $\exists \mathbf{x}' \models \varphi \wedge \neg\psi$, $\forall \mathbf{x} \models \varphi \wedge \psi$, we have $\mathbf{x} \succ \mathbf{x}'$ (Benferhat et al. 2002).
- $\succsim$ satisfies $(\psi|\varphi)$ following the *strong* semantics iff $\forall \mathbf{x} \models \varphi \wedge \psi, \forall \mathbf{x}' \models \varphi \wedge \neg\psi$, we have $\mathbf{x} \succ \mathbf{x}'$ (Benferhat and Kaci 2001).

Given a set of conditional preferences $\mathscr{P} = \{P(\psi_i|\varphi_i) | i = 1, \ldots, n\}$ and a semantics, a preference relation associated with $\mathscr{P}$ must satisfy every preference $P(\psi_i|\varphi_i)$ in $\mathscr{P}$. The optimistic and pessimistic semantics, which better suit the idea of conditional logics, are particularly appropriate to express exceptions. Moreover, a unique weak order can be associated to a set of preferences following these semantics (Pearl 1990; Benferhat et al. 1992, 2002; Boutilier 1994). We can also use again the principle of ceteris paribus comparison or its generalizations, that we presented in Sect. 5.2.1, but at the price of moving away from the spirit of conditional logics.

Conditional logics go back to Lewis (1973). However, the idea of using conditionals to reason about preferences is originally due to Boutilier (1994), and then further developed in other works (Lang 1996; Lang et al. 2002; Benferhat et al. 2002; Lang and van der Torre 2003). These logics have been extended in Kaci and van der Torre (2008b) to allow users refer to several semantics at the same time; they generalize CP-theories (Wilson 2004).

Example 9 (*Example* 1, *continued*) Consider the following set of conditional preferences

$$\mathscr{P} = \{P(|d),\ P(d|s),\ P(n|\bar{s}),\ P(ds|c_1),\ P(n \vee \bar{s}|c_2)\}$$

We interpret these preferences following each of the three semantics described above:

1. *Strong semantics:* The set $\mathscr{P}$ is inconsistent, i.e., no acyclic preference relation satisfies all preferences in $\mathscr{P}$ following this semantics. This is due to preferences $P(n|\bar{s})$ and $P(n \vee \bar{s}|c_2)$ which respectively stipulate that $n\bar{s}c_1$ is preferred to nsc_2 and that nsc_2 is preferred to $n\bar{s}c_1$.
2. *Optimistic semantics:* $\mathscr{P}$ is satisfied by the weak order

$$dsc_1 \succ n\bar{s}c_2 \sim d\bar{s}c_2 \sim dsc_2 \succ n\bar{s}c_1 \sim nsc_1 \sim nsc_2 \sim d\bar{s}c_1$$

3. *Pessimistic semantics:* $\mathscr{P}$ is satisfied by the weak order

$$dsc_1 \succ dsc_2 \succ n\bar{s}c_1 \sim n\bar{s}c_2 \sim nsc_2 \sim d\bar{s}c_1 \sim d\bar{s}c_2 \succ nsc_1$$

Conditional preference logics have been used in argumentation (Kaci and van der Torre 2008a) and database queries (Chomicki 2003). They have also been used in a sustainable development application (Brockhoff 2014).

6 Conclusion

Decision making spans a diverse set of situations, each with a specific type of available information about the user's preferences and with different requirements concerning the time and the cognitive effort to be devoted to the elicitation of the user's preferences and to the computation of a suitable decision. The languages and tools we surveyed in this chapter show us that there needs to be trade-off between, on the one hand, the expressivity of the language used for eliciting and representing preferences, and its complexity in terms of communication and computation. In this chapter we have shown why it is important to represent preferences in a succinct way, and we have surveyed the main succinct preference representation languages developed in the decision sciences and the AI literature. For the sake of brevity, we have mostly left aside the *preference elicitation* task, which needs to be performed before comparing alternatives or finding an optimal alternative.

Representing and eliciting preferences, as well as reasoning about them, is crucial for various application fields of decision aid, especially in electronic commerce. A key application field is that of *recommender systems*, where the system must reason on the preferences of a user so as to recommend her some items that are likely to satisfy her. It is also an important topic for the research community interested in *user modeling*.

Preference elicitation refers to actively querying an agent so as to learn enough of their preferences, and is closely related to a field called *active learning*. More generally, preferences can be learned by observing an agent making choices or ranking alternatives: this is the field of *preference learning* (Fürnkranz and Hüllermeier 2010), which is developing rapidly and that we left out of our study.

As said in several parts of this chapter, compact representation of preference has strong links with valued constraint satisfaction (chapters "Constraint Reasoning" and "Valued Constraint Satisfaction Problems" of volume 2) and knowledge representation, especially nonmonotonic reasoning (chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" of this volume) and graphical models for uncertainty (chapter "Languages for Probabilistic Modeling over Structured and Relational Domains" of volume 2); moreover, its applications span a lot of AI fields such as planning (chapter "Planning in Artificial Intelligence" of volume 2), collective decision making (chapter "Collective Decision Making" of this volume), multicriteria decision making (chapter "Multicriteria Decision Making" of this volume), machine learning (chapter "Designing Algorithms for Machine Learning and Data Mining" of volume 2) and specifically preference learning, electronic commerce and specifically recommender systems.

References

Alanazi E, Mouhoub M, Zilles S (2016) The complexity of learning acyclic CP-nets. In: Proceedings of the Twenty-Fifth international joint conference on artificial intelligence, IJCAI 2016. New York, USA, 9–15 July 2016, pp 1361–1367

Allen TE, Goldsmith J, Justice HE, Mattei N, Raines K (2016) Generating CP-nets uniformly at random. In: Proceedings of the thirtieth AAAI conference on artificial intelligence. Phoenix, Arizona, USA, 12–17 February 2016, pp 872–878

Apt K, Rossi F, Venable B (2005) CP-nets and nash equilibria. In: Proceedings of third international conference on computational intelligence, robotics and autonomous systems (CIRAS'05). Elsevier, Amsterdam

Bacchus F, Grove A (1995) Graphical models for preference and utility. In: Proceedings of the 12th international conference on uncertainty in artificial intelligence (UAI-95), pp 3–10

Benferhat S, Kaci S (2001) A possibilistic logic handling of strong preferences. In: Proceedings of the international fuzzy systems association conference (IFSA'01), pp 962–967

Benferhat S, Dubois D, Prade H (1992) Representing default rules in possibilistic logic. In: Proceedings of the 3rd international conference of principles of knowledge representation and reasoning (KR'92), pp 673–684

Benferhat S, Cayrol C, Dubois D, Lang J, Prade H (1993) Inconsistency management and prioritized syntax-based entailment. In: Proceedings of the 13th international joint conference on artificial intelligence (IJCAI-93), pp 640–645

Benferhat S, Dubois D, Kaci S, Prade H (2002) Bipolar representation and fusion of preferences in the possibilistic logic framework. In: Proceedings of the 8th international conference on principle of knowledge representation and reasoning (KR 2002), pp 421–432

Bertele V, Brioschi F (1972) Nonserial dynamic programming. Academic, New York

Bienvenu M, Lang J, Wilson N (2010) From preference logics to preference languages, and back. In: Proceedings of the 12th international conference on principles of knowledge representation and reasoning (KR 2010)

Bigot D, Zanuttini B, Fargier H, Mengin J (2013) Probabilistic conditional preference networks. In: Proceedings of the twenty-ninth conference on uncertainty in artificial intelligence, UAI 2013. Bellevue, WA, USA, 11–15 August 2013

Bistarelli S, Fargier H, Montanari U, Rossi F, Schiex T, Verfaillie G (1999) Semiring-based CSPs and valued CSPs: frameworks, properties, and comparison. Constraints Int J 4(3):199–240

Bonzon E, Lagasquie-Schiex MC, Lang J, Zanuttini B (2009) Compact preference representation and Boolean games. Auton Agents Multi Agent Syst 1(18):1–35

Booth R, Chevaleyre Y, Lang J, Mengin J, Sombattheera C (2010) Learning conditionally lexicographic preference relations. In: ECAI 2010-Proceedings of the 19th European conference on artificial intelligence. Lisbon, Portugal, 16–20 August 2010, pp 269–274

Boubekeur F, Boughanem M, Tamine-Lechani L (2006) Towards flexible information retrieval based on CP-nets. In: Proceedings of the 7th international conference on flexible query answering systems, pp 222–231

Boutilier C (1994) Toward a logic for qualitative decision theory. In: Proceedings of the 4th international conference on principles of knowledge representation and reasoning (KR'94), pp 75–86

Boutilier C, Hoos HH (2001) Bidding languages for combinatorial auctions. In: Proceedings of the seventeenth international joint conference on artificial intelligence, IJCAI 2001. Seattle, Washington, USA, 4–10 August 2001, pp 1211–1217

Boutilier C, Bacchus F, Brafman R (2001) UCP-networks: a directed graphical representation of conditional utilities. In: Proceedings of the 17th conference on uncertainty in artificial intelligence (UAI-2001), pp 56–64

Boutilier C, Brafman R, Domshlak C, Hoos H, Poole D (2004a) CP-nets: a tool for representing and reasoning with conditional ceteris paribus statements. J Artif Intell Res 21:135–191

Boutilier C, Brafman RI, Domshlak C, Hoos H, Poole D (2004b) Preference-based constrained optimization with CP-nets. Comput Intell 20(2):137–157

Bouveret S, Lang J (2008) Efficiency and envy-freeness in fair division of indivisible goods: logical representation and complexity. J Artif Intell Res 32:525–564

Bouveret S, Endriss U, Lang J (2009) Conditional importance networks: a graphical language for representing ordinal, monotonic preferences over sets of goods. In: Proceedings of the 21st international joint conference on artificial intelligence (IJCAI-09), pp 67–72

Brafman R, Chernyavsky Y (2005) Planning with goal preferences and constraints. In: Proceedings of the international conference on artificial intelligence planning systems (ICAPS-05), pp 182–191

Brafman R, Dimopoulos Y (2004) Extended semantics and optimization algorithms for CP-networks. Comput Intell 20(2):218–245

Brafman R, Engel Y (2009) Directional decomposition of multiattribute utility functions. In: Proceedings of the 1st international conference on algorithmic decision theory (ADT 09), pp 192–202

Brafman R, Domshlak C, Shimony SE (2006) On graphical modeling of preference and importance. J Artif Intell Res 25:389–424

Braziunas D, Boutilier C (2005) Local utility elicitation in GAI models. In: Proceedings of the 22th international conference on uncertainty in artificial intelligence (UAI-05), pp 42–49

Brewka G (1989) Preferred subtheories: an extended logical framework for default reasoning. In: Proceedings of the 11th international joint conference on artificial intelligence (IJCAI-89), pp 1043–1048

Brewka G (2002) Logic programming with ordered disjunction. In: Proceedings of the 18th AAAI conference on artificial intelligence (AAAI-02), pp 100–105

Brewka G (2004) A rank-based description language for qualitative preferences. In: Proceedings of the 16th European conference on artificial intelligence (ECAI'04), pp 303–307

Brockhoff D, Hamadi Y, Kaci S (2014) Using comparative preference statements in hypervolume-based interactive multiobjective optimization. In: 8th international conference on learning and intelligent optimization (LION'04), pp 121–136

Chevaleyre Y, Koriche F, Lang J, Mengin J, Zanuttini B (2010) Learning ordinal preferences on multiattribute domains: the case of CP-nets. In: Fürnkranz J, Hüllermeier E (eds) Preference learning. Springer, Berlin, pp 273–296. https://doi.org/10.1007/978-3-642-14125-6

Chomicki J (2003) Preference formulas in relational queries. ACM Trans Database Syst 28(4):427–466. https://doi.org/10.1145/958942.958946, http://doi.acm.org/10.1145/958942.958946

Cornelio C, Goldsmith J, Mattei N, Rossi F, Venable KB (2013) Updates and uncertainty in CP-nets. In: AI 2013: Advances in artificial intelligence-Proceedings of the 26th Australasian joint conference. Dunedin, New Zealand, 1–6 December 2013, pp 301–312

Coste-Marquis S, Lang J, Liberatore P, Marquis P (2004) Expressive power and succinctness of propositional languages for preference representation. In: Proceedings of the 9th international conference on principles of knowledge representation and reasoning (KR 2004), pp 203–212

Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ (1999) Probabilistic networks and expert systems, 2nd edn. Springer, Berlin

Dimopoulos Y, Michael L, Athienitou F (2009) Ceteris paribus preference elicitation with predictive guarantees. In: Proceedings of the 21st international joint conference on artificial intelligence (IJCAI'09), pp 1890–1895

Domshlak C, Brafman R, Shimony SE (2001) Preference-based configuration of web page content. In: Proceedings of the 17th international joint conference on artificial intelligence (IJCAI-01), pp 1451–1456

Domshlak C, Prestwich S, Rossi F, Venable B, Walsh T (2006) Hard and soft constraints for reasoning about qualitative conditional preferences. J Heuristics 12(4–5):263–285

Doyle J, Wellman MP (1991) Preferential semantics for goals. In: Proceedings of the 9th national conference on artificial intelligence (AAAI-91), pp 698–703

Doyle J, Shoham Y, Wellman MP (1991) A logic of relative desire. In: Proceedings of the 6th international symposium on methodologies for intelligent systems (ISMIS'91), pp 16–31

Dupin de Saint-Cyr F, Lang J, Schiex T (1994) Penalty logic and its link with Dempster-Shafer theory. In: 10th international conference on uncertainty in artificial intelligence (UAI'94), pp 204–211

Elkind E, Wooldridge M (2009) Hedonic coalition nets. In: 8th international joint conference on autonomous agents and multiagent systems (AAMAS 2009), vol 1. Budapest, Hungary, 10–15 May 2009, pp 417–424

Elkind E, Goldberg LA, Goldberg PW, Wooldridge M (2009) A tractable and expressive class of marginal contribution nets and its applications. Math Log Q 55(4):362–376

Engel Y, Wellman M (2008) CUI networks: a graphical representation for conditional utility independence. J Artif Intell Res 31:83–112

Fishburn PC (1970) Utility theory for decision making. Wiley, New York

Fürnkranz J, Hüllermeier E (eds) (2010) Preference learning. Springer, Berlin. https://doi.org/10.1007/978-3-642-14125-6

Goldsmith J, Lang J, Truszczyński M, Wilson N (2008) The computational complexity of dominance and consistency in CP-nets. J Artif Intell Res 33:403–432

Gonzales C, Perny P (2004) GAI networks for utility elicitation. In: Proceedings of the 9th international conference on principles of knowledge representation and reasoning (KR 2004), pp 224–233

Gonzales C, Perny P (2005) GAI networks for decision making under certainty. In: Proceedings of the 2005 IJCAI workshop on advances in preference handling

Gonzales C, Perny P, Queiroz S (2008) Preference aggregation with graphical utility models. In: Proceedings of the 23rd AAAI conference on artificial intelligence (AAAI-08), pp 1037–1042

Gonzales C, Perny P, Dubus JP (2011) Decision making with multiple objectives using GAI networks. Artif Intell 175(7):1153–1179
Greco G, Lang J (2015) Group decision making via weighted propositional logic: complexity and islands of tractability. In: Proceedings of the twenty-fourth international joint conference on artificial intelligence, IJCAI 2015. Buenos Aires, Argentina, 25–31 July 2015, pp 3008–3014
Haddawy P, Hanks S (1992) Representations for decision-theoretic planning: utility functions for deadline goals. In: Proceedings of the 3rd international conference on principles of knowledge representation and reasoning (KR'92), pp 71–82
Halldén S (1957) On the logic of 'better'. Library of Theoria, Lund
Hansson SO (2001) The structure of values and norms. Cambridge University Press, Cambridge
Ieong S, Shoham Y (2005) Marginal contribution nets: a compact representation scheme for coalitional games. In: Proceedings 6th ACM conference on electronic commerce (EC-2005). Vancouver, BC, Canada, 5–8 June 2005, pp 193–202
Jensen F, Graven-Nielsen T (2007) Bayesian networks and decision graphs, 2nd edn. Springer, Berlin
Kaci S, van der Torre L (2008a) Preference-based argumentation: arguments supporting multiple values. Int J Approx Reason 48(3):730–751. https://doi.org/10.1016/j.ijar.2007.07.005, https://doi.org/10.1016/j.ijar.2007.07.005
Kaci S, van der Torre L (2008b) Reasoning with various kinds of preferences: logic, non-monotonicity and algorithms. Ann Oper Res 163(1):89–114
Keeney R, Raiffa H (1976) Decisions with multiple objectives: preferences and value tradeoffs. Wiley, New York
Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge
Koriche F, Zanuttini B (2009) Learning conditional preference networks with queries. In: Proceedings of the 21st international joint conference on artificial intelligence (IJCAI'09), pp 1930–1935
Lafage C, Lang J (2000) Logical representation of preferences for group decision making. In: Proceedings of the 7th international conference on principles of knowledge representation and reasoning (KR 2000), pp 457–468
Lang J (1996) Conditional desires and utilities-an alternative logical approach to qualitative decision theory. In: Proceedings of the 12th European conference on artificial intelligence (ECAI'96), pp 318–322
Lang J, Mengin J (2009) The complexity of learning separable ceteris paribus preferences. In: Proceedings of the 21st international joint conference on artificial intelligence (IJCAI'09), pp 848–853
Lang J, Xia L (2016) Voting in combinatorial domains. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of computational social choice. Cambridge University, Cambridge, pp 197–222
Lang J, van der Torre L, Weydert E (2002) Utilitarian desires. Int J Auton Agents Multi Agent Syst 5:329–363
Lang J, van der Torre L, Weydert E (2003) Hidden uncertainty in the logical representation of desires. In: Proceedings of the eighteenth international joint conference on artificial intelligence (IJCAI'03), pp 685–690
Lehmann D (1995) Another perspective on default reasoning. Ann Math Artif Intell 15(1):61–82
Lewis D (1973) Counterfactuals. Blackwell, Oxford
Liu J, Xiong Y, Wu C, Yao Z, Liu W (2014) Learning conditional preference networks from inconsistent examples. IEEE Trans Knowl Data Eng 26(2):376–390
Liu X, Truszczynski M (2015) Reasoning with preference trees over combinatorial domains. In: Algorithmic decision theory-4th international conference, ADT 2015, pp 19–34
Nilsson D (1998) An efficient algorithm for finding the m most probable configurations in probabilistic expert systems. Stat Comput 8(2):159–173
Nisan N (2006) Bidding languages for combinatorial auctions. In: Cramton P, Shoham Y, Steinberg R (eds) Combinatorial auctions, chap 9. MIT Press, Cambridge

Ohta N, Conitzer V, Ichimura R, Sakurai Y, Iwasaki A, Yokoo M (2009) Coalition structure generation utilizing compact characteristic function representations. In: 15th international conference on principles and practice of constraint programming-CP 2009, pp 623–638

Pearl J (1990) System Z: a natural ordering of defaults with tractable applications to default reasoning. In: Proceedings of the 3rd conference on theoretical aspects of reasoning about knowledge (TARK'90), pp 121–135

Pinkas G (1995) Reasoning, nonmonotonicity and learning in connectionist networks that capture propositional knowledge. Artif Intell 77:203–247

Rahwan T, Michalak TP, Wooldridge M, Jennings NR (2015) Coalition structure generation: a survey. Artif Intell 229:139–174

Roy O, van Benthem J, Girard P (2009) Everything else being equal: a modal logic approach to ceteris paribus preferences. J Philos Log 38(1):83–125

Tan S, Pearl J (1994) Specification and evaluation of preferences for planning under uncertainty. In: Proceedings of the 4th international conference on principles of knowledge representation and reasoning (KR'94), pp 530–539

Trabelsi W, Wilson N, Bridge DG, Ricci F (2010) Comparing approaches to preference dominance for conversational recommenders. In: 22nd IEEE international conference on tools with artificial intelligence, ICTAI 2010, vol 2, pp 113–120

Uckelman J, Endriss, (2008) Preference modeling by weighted goals with max aggregation. In: proceedings of the 11th international conference on principles of knowledge representation and reasoning (KR 2008)

Uckelman J, Chevaleyre Y, Endriss U, Lang J (2009) Representing utility functions via weighted goals. Math Log Q 55(4):341–361

von Wright GH (1963) The logic of preference. Edinburgh University, Edinburgh

von Wright GH (1972) The logic of preference reconsidered. Theory Decis 3:140–169

Wilson N (2004) Extending CP-nets with stronger conditional preference statements. In: Proceedings of the 19th AAAI conference on artificial intelligence (AAAI-04), pp 735–741

Wilson N (2009) Efficient inference for expressive comparative preference languages. In: Proceedings of the 21st international joint conference on artificial intelligence (IJCAI-09), pp 961 – 966

Wilson N (2014) Preference inference based on lexicographic models. In: Proceedings of the twenty-first European conference on artificial intelligence (ECAI 2014). IOS, Amsterdam, pp 921–926

Wilson N, George A (2017) Efficient inference and computation of optimal alternatives for preference languages based on lexicographic models. In: Proceedings of the 26th international joint conference on artificial intelligence (IJCAI 2017), pp 1311–1317

Norms and Deontic Logic

Frédéric Cuppens, Christophe Garion, Guillaume Piolle and Nora Cuppens-Boulahia

Abstract Deontic logic (from Ancient Greek *déon*, what is right) aims to formalize the links existing between the notions of obligation, prohibition, permission and optionality. Deontic logic is at the origin of normative systems which are used to model obligations, prohibitions and sanctions in organizations. In this chapter, we will first present standard deontic logic, then we will analyze its drawbacks. A synthesis of some problems tackled in normative systems is then presented: conditional obligations, norms with exceptions, violations, norms with deadlines and collective obligations. Finally, several application domains for deontic logic are examined.

1 Introduction

Deontic logics are formalisms that aim to translate philosophical notions related to norms into mathematical formulas. The first use of the expression "deontic logic" is due to the Austrian philosopher Ernst Mally (Mally 1926). This expression covers the study of normative concepts such as obligation, duty, permission, prohibition, law as well as exemption. Even if other systems have been proposed before, the work of Finnish philosopher Georg Henrik von Wright is generally considered as the foundation of deontic logic (von Wright 1951). Von Wright proposed the first viable reasoning system on deontic concepts. This system is largely based on the analogy

F. Cuppens (✉) · N. Cuppens-Boulahia
Télécom Bretagne, Brest, France
e-mail: frederic.cuppens@telecom-bretagne.eu

N. Cuppens-Boulahia
e-mail: nora.cuppens@telecom-bretagne.eu

C. Garion
Université de Toulouse, ISAE-SUPAERO/DISC, Toulouse, France
e-mail: christophe.garion@isae-supaero.fr

G. Piolle
CentraleSupélec, Rennes, France
e-mail: guillaume.piolle@centralesupelec.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_8

between the deontic concepts of obligation and permission and the alethic concepts of necessity and possibility.

Deontic logic is a branch of symbolic logic. The viability of a logical deontic system is based on its capacity to model the reasoning that both controls and coordinates our daily life. Deontic logic can thus be used to analyze reasoning in domains such as moral, law, trade or security. Applications of deontic logic traditionally include study of fundamental concepts intervening in regulatory or legislative texts (Jones and Sergot 1996; Sergot et al. 1986). However, since (Meyer et al. 1998) several works have investigated how to use deontic logic to specify information systems. A synthesis of possible applications of logic deontic in this context is given in Wieringa and Meyer (1993).

It is important to distinguish a *norm*, which has no truth value, from a *normative proposition*, which has a truth value. To understand this distinction, consider a normative expression such as "you have the permission to borrow this book and to keep it for one month". This expression can be used by an authority to grant a permission. It can also be used to describe an existing norm. In the first case, the creation of new norms corresponds to a "regulation" activity. In the second case, the objective is to specify that a normative proposition, corresponding in the previous example to a permission, exists in the current state. This activity is only descriptive. These two activities are generally regarded as being exclusive and the first objective of deontic logic is generally to define how to reason on normative propositions. However, some studies have proposed to formalize the norm creation activity with speech acts (Demolombe and Louis 2006). These speech acts allow a normative authority to create a norm and thus to ensure that some normative proposition becomes true. For instance, when a father says to one of his children "I permit you to come back at midnight tonight", the permission for the child to come back at midnight is created throughout the speech act while it did not exist before.

Deontic logic is thus a powerful tool to reason on normative propositions. Several dimensions are to be considered when specifying normative propositions:

- What: what does the normative proposition deal with? The different models can be classified into two categories: normative proposition on a state of the world (e.g. permission to have an overdraft) or on the realization of an action (e.g. obligation to repay a loan).
- When: in which circumstances is a normative proposition activated,[1] i.e. becomes true? The objective is to specify conditional or contextual normative propositions. In the case of an obligation or a prohibition, it is also to specify in which circumstances a normative proposition is violated. In particular, recent works have investigated the concept of obligations with deadlines.
- For whom: who is concerned by the normative proposition? In the case of obligation,[2] this leads to specify individual obligations or group (or collective)

[1] The date of creation of a normative proposition is distinguished from its activation date: a regulation can be activated after its creation.

[2] The same reasoning applies on permissions and prohibitions.

obligations. The entity concerned by the obligation can be a physical entity (a physical person) or a legal person.
- To whom: to which authority should we be accountable and what is the normative system the proposition belongs to? The aim is to specify responsibilities and activities (i.e. communication acts) allowing the transfer of these responsibilities, for instance delegation activities.

This chapter is organized as follows. Section 2 presents the "what?" dimension and examines the logics representing obligations to be and obligations to do. Section 3 studies the "when?" dimension with the conditional and contextual obligations perspective, as well as the problem of exceptions and violations handling. Section 4 completes the "when?" dimension with the analysis of obligations with deadlines. Section 5 deals with collective obligations. Finally, Sect. 6 concludes and proposes a summary of possible applications of deontic logic in information systems.

2 Obligation to Be and Obligation to Do

There are many deontic logic variants, each one with its specific features, advantages and drawbacks regarding adequacy for a given class of problems. The major distinction among these logics lies on the object of obligations. An *obligation to be* is a constraint over a state, a formula that the system must satisfy. It can be seen as an obligation of result. Standard deontic logic is the most prominent illustration of this vision. Conversely, an *obligation to do* imposes the realization of an action or a process, allowing for the representation of an obligation of means.

2.1 Standard Deontic Logic

Standard deontic logic (SDL) is a normal modal logic (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume) in which the universal modality, Ob, represents the concept of obligation. A primitive version of deontic logic was introduced by von Wright in von Wright (1951). However, SDL takes into account several improvements proposed by both the scientific community and von Wright himself. In particular, the initial proposal lied on an "obligation to do", whereas the stable version of SDL is based on an "obligation to be", easier to manipulate. The SDL formula $Ob\ \varphi$ should then be understood as an obligation for φ to be true, φ being, possibly, the resulting state of one or several actions not represented in the formula. The deontic concept covered by an SDL obligation is itself subject to interpretation and adaptation. It might represent a notion of duty, of a norm imposed by an external entity or even of the ideality of a given state ($Ob\ \varphi$ representing the fact that situation φ is ideal for an unspecified reason).

As we will see in the rest of the section, standard deontic logic only imperfectly represents the notion of obligation, yet its mechanisms make up a commonly used basis for the design of more complex logics.

2.1.1 Modalities and Axiomatics

If obligation (Ob) is considered as the primitive (universal) modality, then permission (Per), prohibition (For), optionality (Opt) or gratuity (Gra) of a formula can be defined as syntactic abbreviations using Ob (formulas (1) to (4)). In SDL, prohibition is therefore an obligation on the negation of a formula and permission is an absence of prohibition. Optionality denotes the absence of an obligation (or a permission over the negation of a formula) and gratuity is the conjunct absence of obligation and prohibition.

$$Per\ \varphi \stackrel{def}{=} \neg Ob \neg \varphi \tag{1}$$

$$For\ \varphi \stackrel{def}{=} Ob \neg \varphi \tag{2}$$

$$Opt\ \varphi \stackrel{def}{=} \neg Ob\ \varphi \tag{3}$$

$$Gra\ \varphi \stackrel{def}{=} \neg Ob\ \varphi \wedge \neg Ob \neg \varphi \tag{4}$$

Base modality Ob has a KD axiomatics (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume), which means that obligation is distributive over implication (axiom K) and that SDL obligations must be coherent, in the sense that obligation implies permission (axiom D). As a consequence, the simultaneous obligation and prohibition of a same formula (situation known as a *dilemma*) lead to the logical inconsistency of any SDL system. Furthermore, SDL obligations satisfy the necessitation rule: theorems are obligatory formulas.

As stated in the introduction, one can build an analogy between the concepts of obligation and necessity on the one hand, and of permission and possibility on the other hand. It should be noted however that the concept of necessity is generally associated with a KT axiomatics: if a formula is necessary, then this formula is true (axiom T). By contrast, formula $Ob\ \varphi \wedge \neg \varphi$ is satisfiable in SDL and represents the violation of an obligation on φ.

2.1.2 Semantics

Being a normal modal logic, standard deontic logic may be interpreted over Kripke models (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume), in which the accessibility relation (the deontic accessibility relation), is serial: from any world there is at least one

accessible world. That is to say, if a given (consistent) set of deontic formulas is attached to a given world (for instance the world actually representing the current state of a system), then there is at least one world in which the obligations and prohibitions in those formulas are fulfilled. This feature is the consequence of the hypothesis of coherency of obligations. It expresses the fact that a set of SDL formulas cannot prevent (through the obligations, prohibitions, permissions…it contains) its own realization, under pain of being logically inconsistent. Since all the deontic formulas of a given world are fulfilled in the worlds reachable through the deontic accessibility relation, those worlds are sometimes called "ideal".

Using a possible world semantics for standard deontic logic relies on a simple representation. However, the benefits are not as clear as with linear temporal logic, for instance. Indeed, in the case of deontic logic it is hardly conceivable that an agent or a program maintains a representation of possible worlds according to the deontic context, whereas it may directly represent and reason on instants in time. The standard semantics attached to SDL is therefore mostly conceptual and hardly appeals to intuition.

2.1.3 Limitations

Standard deontic logic suffers from a number of limitations. Some of them are restrictions in terms of expressivity imposed by the initial design choices of the language. For instance, the structure of the obligation modality does not allow for the representation of the entity issuing the obligation, or of the entity subject to it. Furthermore, no temporal notion is natively included in the language, and any conflicting situation between two or several formulas formally leads to the inconsistency of the whole system. This kind of limitation can be at least partially overcome by the means of extensions such as obligation modality differentiation or articulation with a temporal logic (Åqvist 2004).

Other limitations are more fundamentally linked to the logical structure of SDL and are less easily avoided. They are traditionally called "paradoxes" and are the main source of criticism regarding SDL. Paradoxes are either logically correct formulas or derivations with counter-intuitive interpretations, or sets of formulas looking intuitively sound but leading to logical contradictions.

For instance, the Ross paradox, results of a dissonance between the intuitive meaning of disjunctions and their formal definition. It refers to the formula $Ob\ p \rightarrow Ob\ (p \vee q)$, which may be interpreted as "if I ought to post a letter, then I ought either to post it or to destroy it", which is not an intuitive reasoning. In a similar fashion, the paradox of derived obligation (Prior 1954) stems from the structure of logical implication, while the good samaritan Paradox (Prior 1958) is a consequence of the necessitation rule characterizing normal modal logics. A more detailed analysis of formal and philosophical limitations of standard deontic logic can be found in McNamara (2006).

2.2 Deontic Logic of Actions

Obligations to do are another possible representation choice for deontic logics. The idea is to make obligations bear not on states, but essentially on actions (and possibly on their outcomes). A first approach consists of mixing operators from standard deontic logic (or one of its variants) with operators from a dynamic logic, i.e. based on the notion of action. A first proposal in this direction was proposed by Meyer in Meyer (1988). It is an instance of a "reductionist" approach, which reduces the definition of deontic concepts down to the notion of "violation".[3] This approach uses the $[\alpha]\varphi$ operator of dynamic logic, indicating that proposition φ becomes necessarily true as soon as α is executed, as well as a deontic constant V representing a state of violation. The obligation to execute action α, noted $Ob\ \alpha$, and the interdiction to execute α, noted $For\ \alpha$, are respectively defined as follows:

$$Ob\ \alpha \stackrel{def}{=} [\bar{\alpha}]V \tag{5}$$

$$For\ \alpha \stackrel{def}{=} [\alpha]V \tag{6}$$

The formula $[\bar{\alpha}]V$ says that the state of violation V is a consequence of the absence of execution of action α (noted $\bar{\alpha}$). Permission and optionality modalities are then respectively defined as the negation of interdiction and the negation of obligation.

More recently, a non-reductionist approach of the deontic logic of actions has been proposed in Broersen et al. (2001). Reference (Balbiani 2005) and (Cuppens et al. 2005) proposed as well, in 2005, a multi-modal logic associating, among others, SDL obligations to modalities of an extension of Propositional Dynamic Logic (PDL). In this formalism, obligations and permissions bear on process combinations. Since those works also consider the possibility of obligations with delays, we will refer to them again in Sect. 4. A recent work of Knobbout et al. (2016) mixes state-based norms and action-based norms.

Another way of taking actions into account in deontic logic is by relying on *stit* ("see to it that") operators. Those operators, like Pörn's E_a modality (Pörn 1977), Belnap and Perloff's "achievement stit" (Belnap and Perloff 1988) or von Kutschera and Horty's "deliberative stit" (von Kutschera 1986; Horty 1989), build a direct relation between an agent and a state formula, resulting from its actions. The structure of those operators aims to capture the notion of responsibility that an agent may have for the realization of the target formula, through the causality between its choices (the actions it undertakes) and the possible resulting states. Thus, a "stit" operator cannot link an agent to a state which could have occurred independently from its actions, or which it cannot influence (like tautologies or antilogies). The ability of an agent to perform an action or to ensure the occurrence of a state is a key concept

[3] As early as 1958, Anderson (Anderson 1958) had proposed to define the obligation to be Ob by using the alethic necessity operator $\Box$ and a deontic constant V, representing a state of violation: $Ob\ \varphi \stackrel{def}{=} \Box(\neg\varphi \longrightarrow V)$. Intuitively, this definition says that if φ is not realized, then a violation occurs.

of "stit" operators. In this way, when an SDL deontic operator is applied to "stit" formulas, the distinction between the obligation for a proposition to be true in a given state and the obligation to ensure that this proposition becomes true in this state is made explicit. The representation of absurd or impossible obligations can then be avoided. Even though the obligation modality remains subject to the necessitation rule or the nature of logical connectors, using elaborate "stit" operators allows one to avoid some of the paradoxes of SDL, like the Ross paradox. The comfort guaranteed by "stit" operators may however become a handicap when it becomes necessary to work with obligations impossible to fulfil, for instance when working with conflicting norm sets.

3 Conditional and Contextual Obligations

One of the earliest problem tackled in deontic logic is the study of conditional and contextual obligations. A conditional or contextual obligation is an obligation which is taken into account only under certain conditions or in a particular context. For instance, "*if it is raining, you must turn on your headlights*" is a conditional obligation: the obligation of turning on your headlights must be taken into account only when it is raining.[4] Reasoning on conditional and contextual obligations leds to three problems that we will present in this section:

- the difficulty to define a *dyadic* deontic operator to model such obligations. Such an operator should allow to take into account not only the proposition on which the deontic notion is directed, but also a proposition representing the application context of the norm. We will present in Sect. 3.1 several models for dyadic operators and the expected properties of these operators.
- management of exceptions: what happens if a general obligation is contradicted by a more particular obligation? We will present in Sect. 3.2 the notions of conflicts and defeasible norms using a semantics of preferred worlds.
- management of violations, particularly through the problem of contrary-to-duties that will be presented in Sect. 3.3.

3.1 Dyadic Deontic Logic

We want to define a dyadic modal operator for obligation $Ob(\varphi|\psi)$ where φ is a proposition on which the obligation holds and ψ is the context of the application of the obligation. In the following, the remarks and expected properties about this operator may also be applied to other deontic notions like permission or prohibition.

[4] Of course, there may be other cases when turning on your headlights is mandatory: at night, when going through a tunnel etc.

There are two possible models for such an operator:

- define $Ob(-|-)$ using an operator modelling a non-conditional obligation
- define $Ob(-|-)$ without using an unary obligation operator.

3.1.1 Using an Unary Operator to Define a Dyadic Obligation Operator

Using an unary obligation operator $Ob(-)$, there are two ways to define $O(\psi|\varphi)$:

$$Ob(\varphi|\psi) \stackrel{def}{=} \psi \rightarrow Ob(\varphi) \tag{7}$$

$$Ob(\varphi|\psi) \stackrel{def}{=} Ob(\psi \rightarrow \varphi) \tag{8}$$

If we consider that SDL is used to model the unary obligation operator $Ob(-)$, both definitions allow to derive the following properties for all propositions φ, ψ and γ:

$$\vdash Ob(\varphi) \leftrightarrow Ob(\varphi|\top) \tag{9}$$

$$\vdash Ob(\varphi|\psi) \rightarrow Ob(\varphi|\psi \wedge \gamma) \tag{SA}$$

Property (9) indicates that it is possible to represent non conditional obligations with the dyadic operator: simply use a tautology as the conditional part of the dyadic obligation. Property (SA), called *strengthening of the antecedent* is more problematic: it denotes that this representation is not sufficient to represent norms with exceptions as we will see in Sect. 3.2.

If definition (7) is chosen to represent the dyadic obligation, the two following properties can be derived:

$$\vdash \neg\psi \rightarrow Ob(\varphi|\psi) \tag{10}$$

$$\vdash \psi \wedge Ob(\varphi|\psi) \rightarrow Ob(\varphi) \tag{DF}$$

The first theorem is problematic. For instance, if it is not raining, then "if it is not raining, then is is obligatory to put a bathing suit" can be deduced. Notice that "put a bathing suit" can be replaced by *any satisfiable proposition*. The second theorem, called *factual detachment*, shows how to derive non conditional obligations from conditional obligations and facts. For instance, if it is mandatory to drive on the left side of the road if you are in Great Britain, then if you are effectively in Great Britain, it is obligatory to drive on the left side of the road. Factual detachment allows one to derive obligations that apply in a particular situation: these obligations will be called *effective obligations*.

If definition (8) is chosen to represent dyadic obligations, the two following properties are obtained:

$$\vdash Ob(\neg\psi) \rightarrow Ob(\varphi|\psi) \quad (11)$$

$$\vdash Ob(\psi) \wedge Ob(\varphi|\psi) \rightarrow Ob(\varphi) \quad (\text{DD})$$

As in the previous case, the first theorem is problematic: any proposition is mandatory if some proposition is forbidden. For instance, if it is forbidden to drive on the left side of the road, then it is mandatory to kill your neighbors if your are driving on the left side of the road. The second theorem, called *deontic detachment*, allows one to derive non conditional obligations from both conditional and non conditional obligations. Using this theorem, *ideal obligations* are derived: these obligations do not consider the actual situation, but only the set of normative sentences representing a regulation.

The definition of a dyadic deontic operator from an unary deontic operator leads to problems, whatever the chosen option is. Let us however notice that two properties are interesting and should be verified by any dyadic deontic operator:

- factual detachment, used to find what are the obligations that must effectively be applied in a particular situation;
- deontic detachment, that allows one to deduce the obligations that should ideally be applied from a set of normative sentences representing a regulation.

3.1.2 Direct Definition of a Dyadic Operator

The problems raised by the use of SDL and its limitations suggest a more complete and semantically richer understanding of obligation by directly defining dyadic deontic operators. We will rely here on Hansson (1971) and more details can be found in Spohn (1975), van der Torre and Tan (1997), Lewis (1974).

In Hansson (1971), Hansson presents three deontic logics, DSDL1, DSDL2 and DSDL3, whose semantics is based on the notion of *ideality*: some worlds are more ideal than others and the propositions true in these ideal worlds are the obligations to be modelled. There may be several ideality levels representing "degraded" situations and thus solving the problem of representing norms with exceptions (cf. Sect. 3.2). DSLDL1 and DSDL2 are rather weak systems and the focus is put on DSDL3, whose accessibility relation, representing ideality, is transitive and total and corresponds to a *preference relation*. Let us notice that the three DSDL logics do not validate property (SA).

Several logics use the same semantics (cf. (Hansson 1990; van der Torre and Tan 1999b) for instance) and a complete axiomatization of DSDL2 can be found in Parent (2009). We will come back to these logics with preferential semantics in Sects. 3.2 and 3.3, because they may be used to manage exceptions and violations in normative systems. On a more "proof-theoretic" point of view, (Parent and van der Torre 2017a) presents a logical framework to reason on detachment in normative systems.

3.2 Exceptions

In a normative system, there may be a general rule and a rule that applies in a particular context such that both rules allow to deduce inconsistent normative statements. For instance, a regulation $\mathscr{R}$ in a university may stipulate that "*it is forbidden to take emergency exits*", but that "*in case of fire, it is mandatory to take the emergency exits to go out the building*". If we use a dyadic operator *Ob* that represents obligation, then regulation $\mathscr{R}$ may be modelled as follows:

$$Ob(\neg exits|\top) \tag{12}$$

$$Ob(exits|fire) \tag{13}$$

Formula 12 represents the non-conditional prohibition of taking the emergency exits (cf. formula 9). Formula (13) represents the conditional obligation of taking the emergency exits in case of fire.

The norms we should *actually* take into account in such a regulation should be determined in a particular context. We will suppose in the following that $\mathscr{C}$ is a consistent set of propositional formulae representing the context in which the regulation should be evaluated. To find what are the obligations to be taken into account, we should find formulae φ such that $\mathscr{R} \models Ob(\varphi|\mathscr{C})$ (denoting by $\mathscr{C}$ the conjunction of formulas in $\mathscr{C}$).[5]

In our example, whatever model based on SDL among the two presented in Sect. 3.1 is chosen, we can show that $\mathscr{R} \models Ob(exits|fire) \wedge Ob(\neg exits|fire)$. A *dilemma* is obtained: it is both mandatory and prohibited to take the emergency exits in case of fire.

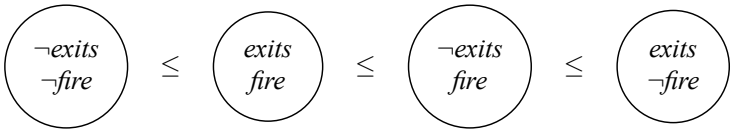
This problem is well-known in logic and is usually solved using a non-monotonic logic (cf. chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" in this volume). The prohibition of taking the emergency exits must apply in all situations *except* when there is a fire. The obligation of not taking the emergency exits should not be derivable from regulation when there is a fire, as it does not apply anymore. Van der Torre and Tan speak about *overridden defeasibility* (van der Torre and Tan 1997), because the rule represented by formula (13) is more specific than the rule represented by formula (12) and should be cancelled when the context allows it.

From a more technical point of view, dyadic operators models based on SDL presented in Sect. 3.1.1 imply that the rule (SA) may be used. $Ob(\neg exits|fire)$ may then be deduced from $Ob(\neg exits|\top)$, which is not acceptable. The use of (SA) should therefore be blocked if norms with exceptions should be taken into account, because it represents the monotonicity of obligations derivation according to the context.

To address this issue, a *preferential* Kripke semantics can be used for the *Ob* operator as presented in Sect. 3.1.2: the accessibility relation between possible worlds is then a preference relation, denoted here by $\leq$ (most preferred worlds are classically

[5]Notice that the reasoning is the same for the other deontic notions (prohibition, permission etc).

the minimal worlds for $\leq$). An obligation may be seen as a constraint on preferences between the worlds. For instance, the obligation of not taking the emergency exits may be represented by the following constraint: the worlds in which $\neg exits$ is true are preferred to those in which *exits* is true *in most cases*. The obligation to take the emergency exits in case of a fire may be represented by the following constraint: the worlds in which $exits \wedge fire$ is true are preferred to those in which $\neg exits \wedge fire$ is true. A possible model of $\mathscr{R}$ with such semantics would then be:



To find the actual obligations with such semantics, the algorithm is pretty simple: find the most preferred world among those which satisfy the context. In our example, when there is no particular context, the most preferred world is one in which $\neg exits$ is true: it is forbidden to take the emergency exits. On the other hand, if we consider that there is a fire, then the most preferred satisfying the context is the one in which *exits* is true: it is mandatory to take the emergency exits.

3.3 Violations

Violations detection may seem rather simple: it seems sufficient to verify that if φ is actually mandatory, then $\neg\varphi$ is not true or else the obligation is violated. If the question of violations management in deontic logic is examined more accurately, the problem of Contrary-to-Duties (CTD) arises rapidly. Lots of works are tackling this problem, see for instance (Carmo and Jones 2002; van der Torre and Tan 1999b, 1998; Tan and van der Torre 1997; Prakken and Sergot 1996, 1997; Cholvy and Garion 2001; van Benthem et al. 2014; Calardo et al. 2014).

A CTD is an obligation that must be applied only in sub-ideal situations. For instance, the obligation to apologize when not keeping your promise is only effective when the obligation of keeping your promises has been violated. The most famous examples of CTD can be found in Chisholm (1963), Forrester (1984) and lead to paradoxes. For instance, Chisholm's paradox may be presented as follows:

1. Mr X must help his neighbors
2. if Mr X does not help his neighbors, then Mr X must not let them know he is coming
3. if Mr X does help his neighbors, then Mr X must let them know he is coming
4. Mr X does not help his neighbors.

First obligation (sentence 1) is called *prima facie* and the second one (phrase 2) is the Contrary-to-Duty. Intuitively, the four sentences constituting the paradox are consistent. The objective is thus to find a formalism in which a *prima facie* rule and its CTD can be represented consistently and independently (i.e., the two

formulas representing the *prima facie* rule and its CTD are logically independent), while allowing to represent the contextual obligation 3.

Several approaches have been proposed to solve the problem of Contrary-to-Duties. For instance, some temporal approaches (cf. (Åqvist and Hoepelman 1981)) distinguish two moments in time: when an obligation is violated and when the secondary obligation of the CTD is fulfilled. There exist also approaches based on action logics that distinguish the condition of a conditional obligation which is considered as a state and the obligation itself that is considered as an action (Castañeda 1981; Meyer 1988). Chisholm's paradox can perfectly be represented in these two kinds of logics, because it is easy to identify in this paradox a temporal dimension and distinguish "it is obligatory to do something" from "it is obligatory to be in a particular state".

However, let us consider the following example proposed by Prakken and Sergot in Prakken and Sergot (1996) (this example could be part of a subdivision by-law):

there must be no dog	(PF)
if there is a dog, there must be a warning sign	(CTD)
there must be no warning sign	(DM)
there is a dog	(F)

This example is interesting because time and action notions are not present in the subdivision by-law. The three sentences are applicable at the same time and no action appears in the regulation. Let us notice that formalisms presented in Sect. 3.1 are not sufficient to correctly represent these four sentences: SDL reductions using definitions (7) and (8) are too strong and make (DM), (CTD) and (F) inconsistent (if using (7)) or (DM), (CTD) and (PF) inconsistent (if using (8)). Let us notice similarly that DSDL3 is too weak to detect the dilemma.

Several approaches claim that reasoning with CTD is simply a particular kind of defeasible reasoning and thus that techniques for non-monotonic reasoning could be applied directly on this scenario (cf. (McCarthy 1994) for instance). If the principles of non-monotonic reasoning can be easily adapted to reason on moral dilemmas or conflicting obligations (Prakken 1996; van der Torre and Tan 1998, 1999b; Horty 1994), it is not the case for CTD: prima facie rule does not "defer" to the CTD, it is always in effect and is in particular violated if the CTD must be applied.

In the dog example, the sentence "if there is a dog, there must be a warning sign" expresses the fact that some non-ideal worlds are more ideal than other non-ideal worlds. The worlds in which there is a dog are clearly non-ideal, but among them the worlds in which there is a warning sign are "better" than the worlds in which there is no warning sign. It seems therefore interesting to use a preferential semantics to represent CTD, like for instance in Prakken and Sergot (1997), van der Torre and Tan (1999b), Cholvy and Garion (2001). An hybrid approach using defeasible logic and preference logic is also proposed in van der Torre and Tan (1997).

The diversity of approaches used to tackle the problem of representation and reasoning with CTD shows that this problem is still open. In particular, there is no

consensus on the formal representation of a scenario involving CTD. Carmo and Jones suggested in Carmo and Jones (2002) eight postulates which, according to them, should be verified by an approach trying to model CTD. For instance, one of these postulates states that the logical representations of the four sentences must be logically independent, another one that the formalism must be applicable on non-temporal examples etc. This initiative provides a "minimal" consensus on what should be expected from logical formalisms dealing with CTD.

The logic they have developed from these postulates uses a dyadic operator Ob to represent conditional obligations and two monadic operators Ob_i and Ob_a to represent respectively what is *ideally* obligatory and what is *actually* obligatory (i.e. in the current context). This approach distinguishes ideal and effective obligations and is interesting because it has a behavior similar to (DD) when it is correct (deriving ideal obligations) and a behavior similar to (DF) to deduce effective obligations. They also introduce a simple agency model that manages violations given what the agent is able to do or not. Cholvy and Garion rely on this approach to develop a formalism using a logic of conditional preferences in Cholvy and Garion (2001). Recent works of van Benthem et al. (2014), Calardo et al. (2014) use also a logic of preferences to deal with CTD. Notice that a sound and complete implementation of Carmo and Jones logic in HOL has been recently presented in Benzmüller et al. (2018). Finally, some solutions to the CTD problem has been proposed in Input/Output logics with a norm-based semantics (instead of a semantics based on preferred world), see Parent and van der Torre (2018), Parent and van der Torre (2017b).

4 Obligations with Delays

Deontic logics are often associated with temporal logics (Åqvist 2004), in order to gain in expressivity. Indeed, obligations, prohibitions or permissions generally apply to states or actions anchored in a given history, and norms themselves often bear time-related constraints. When reasoning on obligations situated in time, it becomes paramount to associate them with a *deadline*, a delay before which the obligation must be fulfilled to avoid a violation. An obligation without deadline can indeed be considered as void (Dignum et al. 2004): either it is an immediate obligation, in which case one can only acknowledge its fulfilment or non-fulfilment without any means to act on it, or it is a standing obligation on an unspecified future, in which case one always has the possibility to postpone, without the obligation ever being formally violated. Only the specification of a deadline or a delay tightly linked to the obligation allows one to characterize its fulfilment or violation, while constraining the means of action of the agent subject to it. This notion of delay or deadline fits the legal notion of *term*, as the "modality of an obligation related to the occurrence of a future event of certain realization" (translated from (Cornu 1987)).

4.1 The Several Models for Obligations with Delays

The design of an operator of obligation with delay depends much on the chosen deontic and temporal formalisms. For instance, Dignum *et al.* have proposed an operator for a branching temporal logic associated to a *stit*-like operator (Dignum et al. 2004), Demolombe et al. for an "obligation to do" formalism introduced by Demolombe (2005), Brunel et al. (2006) work on an extension of the product between SDL and a linear temporal logic, while (Cuppens et al. 2005) introduce the concept via a multi-modal logic combining SDL obligations with the modalities of an extension of dynamic logic embedding action durations.

One may classify the different proposals in three families: those using a notion of *delay* (reasoning on the basis of a duration), those using a *deadline* (reasoning on the basis of a date) and those using a *contextual event* (reasoning on the basis of an abstract temporal constraint). The last kind may model the legal concept of *condition*, which is distinct from *term* in that its realization is not certain. We will refer to all these concepts by the generic term of obligations with delays.

4.2 Criteria and Choice Points for Designing an Operator

Dignum et al. propose in Dignum et al. (2004) six useful choice points regarding the design of an obligation with delay operator. These six questions may find different answers according to the variant of the philosophical context to model and to the structure of the base language, even though some answers may appear more natural than others.

1. Deadline definition: it may be practical to consider the deadline as a certain and non ambiguous event which will occur once and only once in the time flow. It then fits what Åqvist calls "systematic frame constants" (Åqvist 2004). However, specific needs may lead to language enrichments allowing repeatable deadlines, ambiguous deadlines or never-occurring deadlines. One can also, like Dignum et al., choose to use tautological deadlines to express immediate obligations;
2. Deadline in the future: it may seem obvious, for a human, that a deadline in the past (or even at the present instant) leaves one with no means of action. The fulfilment of the obligation is then independent from the actions and capabilities of the agent. For this reason, one may choose to constrain deadlines to be situated in a strict future;
3. Obligations on tautologies: should they be tautologies themselves, as in SDL? Or should they be unsatisfiable, as it is the case when expressing obligations to do with a *stit* operator, because of the agent's absence of control on the obligated formula?

4. Obligations on antilogies: it is known that, in SDL, a situation in which $Ob\ \varphi$ and $Ob\ \neg\varphi$ are simultaneously true corresponds to a dilemma, which should be avoided since it leads to logical inconsistency. Dignum et al. consider in Dignum et al. (2004) that introducing delays may lead to a modification of the notion of dilemma. So, if formula $Ob(\varphi < \delta)$ expresses the fact that an obligation on φ must be fulfilled before a delay δ, then one may wonder whether a situation in which $Ob\ (\varphi < \delta)$ and $Ob\ (\neg\varphi < \delta)$ are simultaneously true corresponds to a dilemma. Dignum *et al.* consider that that situation is coherent provided it is possible to successively fulfil both obligations while respecting delay δ;
5. Obligations not fulfilled at the deadline: the persistence of an unfulfilled obligation past its deadline is discussed in Brown (1996), Elrakaiby et al. (2012). An obligation is said to be persistent if it remains true after its violation. For instance, let us consider an obligation to submit the review of a publication before the acceptance notification deadline. One might consider that this obligation is not persistent, i.e. that, even though unfulfilled, it disappears on the day of the deadline. However, an obligation to pay one's taxes is definitely persistent: the obligation will remain after its violation. In this case, the only way to remove this obligation is to pay. In both cases, persistent or not, it is useful to associate the violation of an obligation with sanctions or *contrary-to-duty* obligations;
6. Nature of violations: as the case may be, one can use either punctual violations, corresponding to an event situated in time, "violation states" in which the agent may be after having violated the obligation and until a specific condition is met (such as a later fulfilment), or a combination of both. The temporal expressivity of the chosen formalism may constrain this choice: event-like violations are not directly manipulable without the temporal operators to reason on past.

In addition, Brunel et al. point out two fundamental principles that any operator for obligations with delays should verify. The *monotony* principle says that from an obligation with a given delay, one can logically derive the same obligation with any longer delay. The *propagation principle* states that the obligation must be maintained over time, until either it is fulfilled or the delay is expired. This specification of the essence of obligations with delays leads to the design of several operators around the mechanics of a temporal *Until*.

One should also note that the concept of prohibition maintained over a period of time may be defined jointly with the one of obligation with delay, by adapting the monotony and propagation principles as well as the six choice points. However, the notion of maintained prohibition is much easier to design, as the simple holding of an immediate prohibition. The two operators exhibit some weak form of duality (between the obligation with delay and the negation of the maintained prohibition). Indeed, if there is no obligation on $\neg\varphi$ with delay δ, then one should be able to infer that there is no prohibition on φ maintained until δ. Conversely, if there is a prohibition on $\neg\varphi$ maintained until δ, then one should be able to infer an obligation on φ with delay δ.

5 Collective Obligation

A collective obligation applies to a group of agents, in such a way that this group as a whole is obliged to achieve a given state or to realize a given task. When none of the different members of a group cannot individually fulfil a collective obligation, it is necessary to assign sub-tasks to the different members of the group and coordinate the fulfilment of these sub-tasks in order to fulfil this collective obligation.

Several works have investigated this type of obligation (Royakkers and Dignum 2000; Grossi et al. 2004; Garion and Cholvy 2007; Elrakaiby et al. 2009; Cuppens et al. 2013). These works have specially focused on the following issues:

- Impact of a collective obligation over individual obligations. Grossi et al. (2004) notice that fulfilling a collective obligation generally requires a planning activity to decompose the collective obligation into sub-tasks to be fulfilled by different members of the group. Grossi et al. thus suggest a model based on the concept of plan and on the distribution of tasks to agents that contribute to the plan. However, Garion et Cholvy (Garion and Cholvy 2007) consider that deriving several individual sub-tasks from the collective obligation may generally depend on several parameters, including the capability of the members of the group to contribute to the collective obligations as well as personal will of each member of this group to be involved in the collective obligation fulfilment. The concept of commitment (Royakkers and Dignum 2000; Garion and Cholvy 2007) is proposed to model an agent that is both capable and willing to contribute to the fulfilment of some sub-tasks of a collective obligation. In case of commitment of an agent, then sub-obligations can be derived, i.e. there is an obligation to fulfil the derived sub-tasks.
- Group coordination. Grossi et al. (2004) investigate this problem and propose a model in the form of additional obligations for the agents involved in the collective obligation fulfillment to be informed of the obligations they have to fulfil. Moreover, these coordination obligations also include the obligation to inform other members of the group when a given sub-obligation that contributes to the fulfilment of the collective obligation is actually fulfilled. These coordination obligations are then integrated in the plan allowing the fulfilment of the collective obligation.
- Responsibility for the group members in case of violation. In Garion and Cholvy (2007), the authors investigate how to analyze responsibility for group members in case of violation of a collective obligation. In one hand, a group member cannot be held responsible for non fulfilling the sub-tasks they are not capable to fulfil. In the other hand, if several members of a group are capable to fulfil a given sub-task, then agents that are committed to fulfil the obligation are firstly responsible for non fulfilling this sub-task.
- Collective obligation with deadlines. Must this deadline be equal for all the group members? This problem is addressed in Elrakaiby et al. (2009). This paper presents a model where members of a group may have different deadlines. For example, one collective obligation may have a deadline corresponding to the end of the working day, a deadline that may depend on the working hours of the different group members. Elrakaiby et al. (2009) also investigate how the state of a

group may change over time, because some members may join or leave the group. In this case, the violation of a collective obligation is raised only when the deadline is achieved for all the active members of the group.

In Carmo and Pacheco (2001), Carmo and Pacheco consider that assignment of obligations to a group of agents require creating a new agent, called institutional agent. In particular, these agents can join and leave the group and thus, the group of agents can change. In contrast, the institutional agent remains the same. It is then necessary to specify how this agent can interact with the members of the group that this institutional agent represents as well as agents external to the group.

The approach suggested in Carmo and Pacheco (2001) formalize the difference from a normative point of view between a group of agents and the institutional agent that the group. It is clearly possible to consider situations where a group of agents jointly executes an action, for example moving an heavy table. However, the concept of institutional agent occurs when a single collective entity is used in the norm definition, as it is the case for a company for example.

Thus, the objective of Carmo and Pacheco (2001) is to model how an institutional agent interacts with the external world from a normative point of view. It is especially important to specify how the actions executed by the individual agents "count as" actions executed by the institutional agent. We refer the interested reader to the work by Jones and Sergot (1996) for a logical formalisation of this *count as* logical operator. Let us notice that the actions executed by the individual agents on behalf of the institutional agent depend on the role assigned to this institutional agent when it executes this action. Indeed, an agent may be permitted to execute an action when she is assigned to a given role but this same action would be prohibited when she is assigned to another role. Thus, in Carmo and Pacheco (2001), the authors suggest an extension of the modal operator *stit* to explicitly include the role assigned to the agent when she executes an action. It is then possible to assign norms to role and interpret the actions executed by individual agents as actions executed on behalf of the institutional agent. Using the work presented in Demolombe and Louis (2006), it is also possible to formalise speech acts having the effects to create normative institutional facts, such as, for example, the creation of an obligation. For more recent works on the subject of collective obligation, (Porello 2018) uses non normal modal logics to relate collective responsibility to individual responsibility by distinguishing common, aggregative and corporate actions. Finally, (Pigozzi and van der Torre 2017) presents deontic challenges with a focus on multiagent systems which is clearly related to collective obligations.

6 Conclusion

We have presented the principal dimensions to be taken into account to formalize the deontic concepts of obligation, prohibition and permission. Deontic logic is a very active research domain and plenty of propositions lead to important advances

in recent years. We should also mention the results on defeasible deontic reasoning and its application to violation and "Contrary-to-Duties" management.

However, and contrary to other modal logics like temporal logic, there is no stable model for deontic concepts. Several difficult problems, for which there is no universally accepted solution, still have to be explored. Consider for instance open problems concerning conflicting deontic rules, obligations with delays and formalization of responsibility and delegation concepts. Modelling sanction and reparation processes has been little explored also. Gabbay et al. (2013) is a recent effort to present a detailed overview of current research on deontic logic and should be taken as starting reference for someone interested in deontic logic (see also the Deontic Logic website at http://www.deonticlogic.org).

Nevertheless, the synthesis work realized in Wieringa and Meyer (1993) shows that there are plenty of applications of deontic logic to information systems, e.g.:

- analysis of laws and regulations. The first works in this area were not based on deontic logic and used first-order logic to model a set of norms (cf. for instance (Gray 1985; Sergot et al. 1986)). This approach, so-called "factual", cannot distinguish the actual world from the "ideal" world represented by the regulation. Factual approaches reach their limits when it is necessary to reason on situations in which norms are violated and Contrary-to-Duty norms are activated. Deontic logic has the advantage to be able to represent consistently violation situations.
- contractualization process. Languages based on speech acts play an important role to formalize interactions between agents in lots of applications such as electronic commerce (Dignum 2002). Kimbrough et al. (1984) have considered in 1984 the possibility to use deontic logic to represent performative speech acts related to contract negotiation. More recently, Demolombe and Louis (2006) studied the formalization of speech acts with normative effects such as creation of an obligation, assignment of a role to an agent or declaration of bidding opening. Deontic logic allows therefore to formally specify exchanges between agents in applications such as online shopping and to verify the compliance of these exchanges with respect to contractual and legal commitments.
- representation of integrity constraint in databases (cf. chapter "Databases and Artificial Intelligence" of Volume 3). Works like (Wieringa et al. 1989; Demolombe and Jones 1996) distinguish two kinds of constraints: those corresponding to alethic constraints (constraints on the real world) and those corresponding to deontic constraints. For instance, the constraint "January has 31 days" is a necessity of real world whereas "the salary of an employee cannot decrease" is a deontic constraint corresponding to a prohibition. It is important to distinguish these two types of constraints and to consider the case when a deontic constraint is violated. Indeed, if the second constraint is modelled as a necessity, then some updates of the database may be blocked or the database may become inconsistent if the update is accepted when the constraint is violated.
- security policies representation. A security policy can be modeled as a set of norms which controls the system functioning. It is generally considered that a security policy contains different types of rules, particularly access and usage control

rules (Cuppens-Boulahia and Cuppens 2008). An access control policy corresponds to permissions and prohibitions that filter access requests to the resources of the information system. An usage control policy applies when somebody has been granted access to a resource and can be represented by a set of obligations that should be fulfilled conjointly with the resource use (Hilty et al. 2007; Cuppens and Cuppens-Boulahia 2010). Deontic logic allows then the separation between the functional specification of a system from the security constraints (corresponding to deontic conditions) (Khosla and Maibaum 1987).

- security properties expression (Bieber and Cuppens 1992; Glasgow et al. 1992; Cuppens and Demolombe 1996; Aucher et al. 2010; Balbiani and Seban 2011). Several works focused on formalizing security properties like confidentiality by using deontic concepts. Intuitively, the confidentiality property should ensure that some secret information will not be divulged to agents that are not authorized to know them. This property can be formalized in a multi-modal logic associating epistemic modalities to represent the agents knowledge and deontic modalities to represent permissions and prohibitions of agents. Confidentiality property is then expressed by the guarantee that the agents can only acquire information they are permitted to know.

As Wieringa and Meyer explain in Wieringa and Meyer (1993), the use of deontic logic in information systems not only causes formalization problems, some of them are difficult to solve, but also technical, philosophical, legal and social problems. These different dimensions open countless possibilities of inter-disciplinary research works on taking into account deontic concepts in information systems.

References

Anderson AR (1958) A reduction of deontic logic to alethic modal logic. Mind 67:100–103

Åqvist L (2004) Combinations of tense and deontic modalities. In: Lomuscio A, Nute D (eds) Proceedings of the 7th international workshop on deontic logic in computer science (DEON 2004), vol 3065, LNCS, Madeira, Portugal, pp 3–28

Åqvist L, Hoepelman J (1981) Some theorems about a "tree" system of deontic tense logic. In: Hilpinen R (ed) New studies in deontic logic. Reidel, Dordrecht, pp 187–221

Aucher G, Boella G, van der Torre L (2010) Privacy policies with modal logic: the dynamic turn. In: Governatori G, Sartor G (eds) Deontic Logic in Computer Science, 10th International Conference, DEON 2010. Lecture notes in computer science, vol 6181. Springer, Berlin, pp 196–213

Balbiani P (2005) Constitution et développement d'une logique des modalités aléthiques, déontiques, dynamiques, et temporelles en vue de la formalisation du raisonnement sur les actions et sur les normes. In: Herzig A, Lespérance Y, Mouaddib A-I (eds) Modèles formels de l'interaction, Caen (France), 01/05/2005-31/05/2005. Cépaduès éditions, pp 23–33

Balbiani P, Seban P (2011) Reasoning about permitted announcements. J Philos Log 40(4):445–472

Belnap N, Perloff M (1988) Seeing to it that: a canonical form for agentives. Theoria 54:175–199

Benzmüller C, Farjami A, Parent X (2018) A dyadic deontic logic in hol. In: Proceedings of the 14th international conference on deontic logic and normative systems

Bieber P, Cuppens F (1992) A logical view of secure dependencies. J Comput Secur 1(1):99–130

Broersen J, Wieringa R, Meyer J-JC (2001) A fixed-point characterization of a deontic logic of regular action. Fundam Inform 48(2–3):107–128

Brown MA (1996) Doing as we ought: towards a logic of simply dischargeable obligations. In: [Brown and Carmo, 1996], pp 47–65

Brown MA, Carmo J (eds) (1996) Deontic logic, agency and normative systems, DEON '96: third international workshop on deontic logic in computer science. Workshops in Computing, Sesimbra, Portugal, 11–13 January 1996. Springer, Berlin

Brunel J, Bodeveix J-P, Filali M (2006) A state/event temporal deontic logic. In: Goble L, Meyer J-J (eds) Deontic logic and artificial normative systems. Lecture notes in computer science, vol 4048. Springer, Berlin, pp 85–100

Calardo E, Governatori G, Rotolo A (2014) A preference-based semantics for CTD reasoning. Springer International Publishing, Cham, pp 49–64

Carmo J, Jones A (2002) Handbook of philosophical logic, 2nd edn. Extensions to classical systems 2, chapter Deontic logic and contrary-to-duties, vol 8, Kluwer Publishing Company, pp 265–343

Carmo J, Pacheco O (2001) Deontic and action logics for organized collective agency, modeled through institutionalized agents and roles. Fundam Inform 48(2–3):129–163

Castañeda H-N (1981) The paradoxes of deontic logic: the solution to all of them in one fell swoop. In: Hilpinen R (ed) New studies in deontic logic. Reidel, Dordrecht, pp 37–85

Chisholm RM (1963) Contrary-to-duty imperatives and deontic logic. Analysis 24:33–36

Cholvy L, Garion C (2001) An attempt to adapt a logic of conditional preferences for reasoning with contrary-to-duties. Fundam Inform 48(2,3):183–204

Cornu G (1987) Vocabulaire juridique. Quadrige - Presses universitaires de France, France

Cuppens F, Cuppens-Boulahia N (2010) Spécification et gestion des obligations pour le besoin de contrôle d'usage (un aperçu). Revue Génie Logiciel 94:2–5

Cuppens F, Cuppens-Boulahia N, Elrakaiby Y (2013) Formal specification and management of security policies with collective group obligations. J Comput Secur 21(1):149–190

Cuppens F, Cuppens-Boulahia N, Sans T (2005) Nomad: a security model with non atomic actions and deadlines. In: Proceedings of the 18th IEEE workshop on computer security foundations, IEEE Computer Society, pp 186–196

Cuppens F, Demolombe R (1996) A deontic logic for reasoning about confidentiality. In: [Brown and Carmo, 1996], pp 66–79

Cuppens-Boulahia N, Cuppens F (2008) Specifying intrusion detection and reaction policies: An application of deontic logic. In: van der Meyden R, van der Torre L (eds) DEON. Lecture Notes in Computer Science, vol 5076. Springer, Berlin, pp 65–80

Demolombe R (2005) Formalisation de l'obligation de faire avec délais. Actes des Troisièmes journées francophones des modèles formels de l'interaction (MFI'05). Caen, France, pp 103–111

Demolombe R, Jones AJI (1996) Integrity constraints revisited. Log J IGPL 4(3):369–383

Demolombe R, Louis V (2006) Speech acts with institutional effects in agent societies. In: Goble L, Meyer J-JC (eds) DEON, Lecture notes in computer science, vol 4048, Springer, Berlin, pp 101–114

Dignum F (2002) Software agents and e-business, Hype and Reality. In: Wieringa R, Feenstra R (eds) Enterprise information systems III. Kluwer

Dignum F, Broersen J, Dignum V, Meyer J-J (2004) Meeting the deadline: Why, when and how. In: Hinchey MG, Rash JL, Truszkowski W, Rouff C (eds) Third international workshop on formal approaches to agent-based systems (FAABS'04). Lecture Notes in Computer Science, vol 3228, Greenbelt, MD, USA, pp 30–40

Elrakaiby Y, Cuppens F, Cuppens-Boulahia N (2009) Formalization and management of group obligations. In: Proceedings of the 2009 IEEE international symposium on policies for distributed systems and networks, IEEE Computer Society, pp 158–165

Elrakaiby Y, Cuppens F, Cuppens-Boulahia N (2012) Formal enforcement and management of obligation policies 71(1):127–147

Forrester JW (1984) Gentle murder, or the adverbial Samaritan. J Philos 81:193–197

Gabbay D, Horty J, Parent X, van der Meyden R, van der Torre L (eds) (2013) Handbook of deontic logic and normative systems. College Publications

Garion C, Cholvy L (2007) Deriving individual obligations from collective obligations. In: Boella G, van der Torre L, Verhagen H (eds) Normative Multi-agent Systems. Dagstuhl Seminar Proceedings, vol 07122, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany

Glasgow JI, MacEwen GH, Panangaden P (1992) A logic for reasoning about security. ACM Trans Comput Syst 10(3):226–264

Gray GB (1985) Statutes enacted in normalized form: the legislative experience in tennessee. In: Computer power and legal reasoning. West Publishing Co., pp 467–493

Grossi D, Dignum F, Royakkers LMM, Meyer J-JC (2004) Collective obligations and agents: Who gets the blame? In: Lomuscio A, Nute D (eds) DEON. Lecture notes in computer science, vol 3065. Springer, Berlin, pp 129–145

Hansson B (1971) An analysis of some deontic logics. In: Hilpinen R (ed) Deontic logic: introductory and systematic readings, Dordrecht: Reidel, pp 121–147 (Originally published 1969 in *Noûs* 3: 373-398)

Hansson SO (1990) Preference-based deontic logic (PDL). J Philos Log 19:75–93

Hilty M, Pretschner AA, Basin DA, Schaefer C, Walter T (2007) A policy language for distributed usage control. In: Proceedings of ESORICS'07, pp 531–546

Horty J (1994) Moral dilemmas and non-monotonic logic. J Philos Log 23:33–65

Horty JF (1989) An alternative stit operator. Technical report. Philosophy Department, University of Maryland

Jones AJI, Sergot M (1996) A formal characterisation of institutionalised power. Log J IGPL 4(3):427–443

Khosla S, Maibaum T (1987) The prescription and description of state based systems. In: Temporal logic in specification. Lecture notes in computer science, vol 398. Springer, Berlin, pp 243–294

Kimbrough SO, Lee RM, Ness D (1984) Performative, informative and emotive systems: the first piece ofthe PIE. In: Maggi L, King J, Kraenens K (eds) Proceedings of the fifth conference on information systems, pp 141–148

Knobbout M, Dastani M, Meyer JC (2016) A dynamic logic of norm change. In: ECAI 2016 - 22nd european conference on artificial intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including prestigious applications of artificial intelligence (PAIS 2016), pp 886–894

Lewis D (1974) Logical theory and semantical analysis, chapter Semantic analysis for dyadic deontic logic. D. Reidel Publishing Company, pp 1–14

Mally E (1926) Grundgesetze des Sollens. Elemente der Logik des Willens. Graz: Leuschner and Leubensky

McCarthy L (1994) Defeasible deontic reasoning. Fundamenta Informaticae 21:125–148

McNamara P (2006) Deontic logic. In: Zalta EN (ed) The stanford encyclopedia of philosophy. Stanford University

Meyer J-J (1988) A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic. Notre Dame J Form Log 21(1):109–136

Meyer J-JC, Wieringa RJ, Dignum FPM (1998) The role of deontic logic in the specification of information systems. In: Logics for databases and information systems. Kluwer, pp 71–115

Parent X (2009) A complete axiom set for Hansson's deontic logic DSDL2. Log J IGPL 18(3):422–429

Parent X, van der Torre L (2017a) Detachment in normative systems: examples, inference patterns, properties. IfCoLog J Log Appl 4(9):2996–3039

Parent X, van der Torre L (2017) The pragmatic oddity in a norm-based deontic logic. In: Governatori G (ed) 16th international conference on artificial intelligence and law (ICAIL-2017). ACM Publications

Parent X, van der Torre L (2018) Input/output logics with a consistency check. In: Proceedings of the 14th international conference on deontic logic and normative systems

Pigozzi G, van der Torre L (2017) Multiagent deontic logic and its challenges from a normative systems perspective. IfCoLoG J Log Appl 4(9):2929–2993

Porello D (2018) A logic for reasoning about group norms. In: Proceedings of the 14th international conference on deontic logic and normative systems
Pörn I (1977) Action theory and social science: some formal models. Synthese Library:120
Prakken H (1996) Two approaches to the formalization of defeasible deontic reasoning. Stud Log 57:73–90
Prakken H, Sergot M (1996) Contrary-to-duty obligations. Stud Log 57(1):91–115
Prakken H, Sergot M (1997) Dyadic deontic logic and contrary-to-duty obligations. In: Nute D (ed) Defeasible Deontic Logic. Synthese Library, pp 223–262
Prior AN (1954) The paradoxes of derived obligation. Mind 63:64–65
Prior AN (1958) Essays in moral philosophy, chapter escapism. University of Washington Press, USA, pp 135–146
Royakkers LMM, Dignum F (2000) Organizations and collective obligations. In: Ibrahim MT, Küng J, Revell N (eds) DEXA. Lecture notes in computer science, vol 1873. Springer, Berlin, pp 302–311
Sergot M, Sadri F, Kowalski R, Kriwaczek F, Hammond P, Cory HT (1986) The british nationality act as a logic program. Commun ACM 29(5):370–386
Spohn W (1975) An analysis of Hansson's dyadic deontic logic. J Philos Log 4(2):237–252
Tan Y-H, van der Torre L (1997) Contextual deontic logic: violation contexts and factual defeasability. In: Formal models of agents. Lecture notes in artificial intelligence, vol 1760. Springer, Berlin, pp 240–251
van Benthem J, Grossi D, Liu F (2014) Priority structures in deontic logic. Theoria 80:116–152
van der Torre L, Tan Y (1997) The many faces of defeasibility in defeasible deontic logic. In: Nute D (ed) Defeasible deontic logic, Synthese library, vol 263 . Kluwer, pp 79–121
van der Torre L, Tan Y (1998) An update semantics for prima facie obligations. In: Prade H (ed) Proceedings of the thirteenth european conference on artificial intelligence (ECAI'98), pp 38–42
van der Torre L, Tan Y (1999a) Contrary-to-duty reasoning with preference-based dyadic obligations. Ann Math Artif Intell 27(1–4):49–78
van der Torre L, Tan Y (1999b) An update semantics for defeasible obligations. In: Laskey K, Prade H (eds) Proceedings of the fifteenth conference on uncertainty in artificial intelligence (UAI'99), pp 628–631
von Kutschera F (1986) Bewirken. Erkenntnis 24:253–281
von Wright GH (1951) Deontic logic. Mind 60:1–15
Wieringa R, Meyer J-JC, Weigand H (1989) Specifying dynamic and deontic integrity constraints. Data Knowl Eng 4:157–189
Wieringa RJ, Meyer J-JC (1993) Applications of deontic logic in computer science: A concise overview. In: Deontic logic in computer science: normative system specification. Wiley, New York, pp 17–40

A Glance at Causality Theories for Artificial Intelligence

Didier Dubois and Henri Prade

Abstract Causality plays a key role in the understanding of the world by humans. As such, it has been considered by artificial intelligence researchers from different perspectives ranging from the use of causal links in diagnosis or in reasoning about action to the ascription of causality relations and the assessment of responsibility. In the last two decades, some formal models of causality, such as those proposed by Pearl and Halpern, have been much influential beyond the field of artificial intelligence because they account for the distinction between actual causality and spurious correlations. Yet other aspects of causality modeling are worth of interest, such as the role played by the notion of abnormality, since what we need to explain are often deviations from the normal course of things. The chapter provides a brief but extensive overview of the artificial intelligence literature dealing with causality, albeit without the ambition of giving a complete account of works by philosophers and psychologists that have influenced it.

1 Introduction

The notion of causality is important for artificial intelligence as it is essential when reasoning about events occurring in the outside world, but also to devise artifacts that efficiently act on it. Causality plays a key role in many human activities from making predictions about potential diseases due to smoking or pollution of cities, to deciding who is responsible for a car accident and should be blamed for it. Causality is also of primary importance in advanced data analysis (Gammerman 1999).

There is a huge philosophical literature on causality, because it seems to be very hard to grasp the nature of causal relations, let alone to formally model them. There has been various attitudes about causation that range from philosophical skepticism to the claim that it rules the world in a deterministic way. Some have even deemed

D. Dubois (✉) · H. Prade
IRIT, CNRS and Université Paul Sabatier, Toulouse, France
e-mail: dubois@irit.fr

H. Prade
e-mail: prade@irit.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_9

causality as being undefinable, as too complex a notion, like Zadeh (2002), or because it is so basic that only other concepts could be explicated using it as a primitive notion. There has been a tendency to exclude the concept of causation from scientific approaches in the early XX^{th} century, but things are less clear to-day. The small book by Mumford and Anjum (2013) surveys the various points of view across the History. See also (Paul and Hall 2013) for an overview, with discussions, of approaches to causality, (Cartwright 2007) for a discussion of notions of causality appropriate to the sciences, and (Chambaz et al. 2014) for a debate between a philosopher, a statistician and a medical doctor about causality. A thorough and renewed discussion of causality with many examples can be found in the recent book by Pearl and Mackenzie (2018).

There is actually an opposition between those scholars who claim that causality is just a way to explain how things occur in the world, and those who claim that causality is really at work in the outside world. David Hume (1748), a philosopher from the XVIIIth century, considered that one can only observe regularities in the world, namely that some events are always followed by other events, but that we cannot observe the causal connection between them, if any. In that case, we can speak of perceived causation, whereby two events are just perceived to occur conjointly, one preceding the other. Temporality is essential to distinguish the cause from the effect, and lays bare the fundamental dissymmetry in the causal connection, not to be confused with mere correlation. This view of causality may be sufficient for prediction purposes.

As opposed to the Humean view, some, like Spinoza (1992) in the XVIIth century, see causality as *necessary* connection in the outside world. Here, the concern is to lay bare what could be named the real essence of causality. One claim often found in the literature is that causation basically consists in a transfer of energy (like when one moving billboard ball hits another previously still one). Following this trend, one may speak of the actual cause of an event. It is clear that "necessitarian causality" can be the basis of actions that will modify the world, and that it opens the way to the assessment of responsibility for blaming or praising purposes.

These two major trends can be distinguished in nowadays uses of causation. Actual causation appears to be essentially concerned with the influence of some variables over other variables, a conception that is at work in the perspective of scientific discovery (Salmon 1984) or modeling, as well as for instance in daily medicine or even in epidemiology. At the opposite, commonsense (perceived) causation, dealing more often with events, is ubiquitous in all sorts of matters and activities. Not only are cause and effect relations essential to understanding, explaining, and generally making sense of the world, but also they play a pivotal role in a wide range of other processes, among which prediction and diagnosis, praise and blame, goal-directed reasoning, and persuasion. Artificial intelligence is concerned by both the actual and the commonsense views of causality, depending on whether its goal is to build an automatic system for diagnosis, or to assist people in making sense of a particular situation.

One interesting issue is whether causality is an all-or-nothing concept or not. Actual causality can be a matter of degree, as for instance fever caused by a cold. One may have that the stronger the cold, the higher the fever. Another issue is to

compute the degree of certainty that an effect seems to follow a cause. It is clear that if causation is primarily understood as a matter of observing regularities, the frequency with which an effect follows from a cause can be less than one. Yet another type of degree of causation is encountered when several causes jointly contribute to an effect. The problem may be then to determine the degree of contribution of each component of the cause to the resulting effect; see (Chockler and Halpern 2004), (Alechina et al. 2017). This is especially of interest in blaming or praising tasks, when it is important to determine the contribution of each actor to the resulting effect. See the recent paper by Kleiman-Weiner and Halpern (2018) defining a degree of blameworthiness.

It has been often implicitly assumed that the study of causality should aim at a grand model that would encompass all uses of causation. However, the weaknesses of many proposed models, and the subtleties of others (for instance, see the variants of actual causality in (Halpern 2017), plus the fact that each model never truly addresses all uses of causation, suggest that the various facets of the concept might require different treatments. As a matter of fact, various scholars in the fields of philosophy, artificial intelligence, and psychology (e.g., Lewis 1986; Pearl 2000; Keil 2006) have acknowledged the existence of multiple issues in causal analysis.

This chapter is organized as follows. The next section reviews some basic notions and questions often involved in the study of causation. Then, it provides a reasoned list of artificial intelligence problems where causality plays an important role. Section 3 then presents a number of formal models for causality that, for the main part have been proposed in the area of Artificial intelligence: relational models, logical models, probabilistic approaches, graphical models, qualitative nonmonotonic approaches to perceived causation, action logics, and approaches based on structural equations that focus on actual causes.

2 Causality in Artificial Intelligence: Issues and Problems

In this section, we consider basic notions involved in causal relations, that are essential when formalizing them: the difference between causality and correlation or logical inference, the use of counterfactuals, the role of time, and abnormality, the notion of intervention, the idea that the simpler the alleged cause the more likely it is, the question of the interaction between causes, and whether causality is transitive or not. Then we discuss AI problems where causality and these various aspects play a major role.

2.1 Basic Issues and Principles Underlying Causal Links

Causality, material implication and conditionals. It is tempting, naively, to model the statement 'A causes B', where A and B are events, by the material implication $A \rightarrow B = \neg A \vee B$, since from A we can deduce B. However, this model is

reversible, in the sense that it implies that $\neg B$ would cause $\neg A$, which is certainly not the case if we accept the idea that the effect occurs later than the cause as pointed out by Simon (1952). At best we could claim that observing $\neg B$ suggests that the cause A is not present. So, formal models of causality refrain from using material implication most of the time, or from using it in both ways (like in model-based diagnosis, see the chapter "Diagnosis and Supervision: Model-based Approaches" in this volume, where causes and observations are represented by specific literals, and computing causes come down to a problem of abduction in classical logic). Psychological experiments indicate that material conditionals hardly account for perceived causality (Over et al. 2007).

As an alternative, one may model a causal link by a three-valued conditional $B|A$ introduced by De Finetti (1936). It differs from the material conditional by its truth-valuation when the cause A is absent (the conditional is then not applicable, which is modeled by a third truth-value). This approach accounts for the asymmetry of the causal conditional since $B|A$, $\neg B|\neg A$, and $A|B$ have different truth-tables (see chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" in this volume).

Probability, correlation and time Unfortunately, when applying probability theory and conditioning to the modeling of causality, the asymmetric nature of 3-valued conditionals becomes immaterial. The usual rendering of A causes B in probability theory is $P(B|A) > P(B)$ (after Good 1961, 1962). With this definition, causation is viewed as a simple probability increase, and is indistinguishable from correlation, since the above condition equivalently writes $P(A \wedge B) > P(A)P(B)$, which is clearly equivalent to $P(A|B) > P(A)$, i.e., that B causes A. This inequality just indicates that something has caused the conjoint appearance of A and B, and only temporality can distinguish the cause from the effect (e.g., that B occurs later than A). Namely as done for example by Suppes (1970): A is a prima facie cause of B if and only if (i) B is somewhat possible, $P(B) > 0$, (ii) B becomes more probable in the context of A, $P(B|A) > P(B)$, and (iii) the occurrence of A happens before the occurrence of B. Precedence thus acts as a filter for acceptable causal relations. However, the statement that causality is correlation plus a precedence constraint may be seen as unconvincing, especially in the scope of capturing actual causality. One way to remedy this situation is to use the notion of intervention (von Wright 1971; Pearl 1994) discussed later on in this paper.

Counterfactual causality Although not all counterfactual statements count as examples of causality, it is one of the most important notions in the representation of causality in the present time. A counterfactual is a statement of the form: had A not occurred, B would not have occurred either. Uttering such a statement implies pragmatically that A did actually occur, so counterfactuals can be easily disparaged, on the ground that no one knows for sure what would have happened if A had not occurred. Nevertheless, our knowledge of the world gives us strong intuitions about how things would evolve if the circumstances were different, and we feel this kind of statements as deeply connected to our conception of cause.

The relevance of counterfactuals in causal reasoning has been pointed out by many authors in artificial intelligence (e.g., Lewis 1973; Stalnaker 1968; Pearl 2000; Pearl et al. 2016), but also in philosophy (e.g., Mackie 1974; Woodward 2003; Hitchcock 2001) and psychology (e.g., Hilton and Slugoski 1986; Spellman and Mandel 1999; McEleney and Byrne 2006). The idea these authors have strived to capture is that all we know has not the same status: some parts of our knowledge enjoy a greater stability than others; if some things we believe to be true were in fact false, we could still make use of the more stable parts to derive some conclusions about what would then happen. This mental operation allows us to give content to counterfactual statements.

Despite the importance of counterfactuals in the study of causation, one cannot reduce it to counterfactual knowledge. The counterfactual approach to asserting causality fails in two situations. First, A may be necessary for B to occur, but it may not be sufficient. Then it cannot stand as the cause for B. Yet, had A not occurred, B would not have occurred either. For instance, although it may be true that "If Peter had not been born, he would not have got an accident", Peters parents cannot be considered as being a cause of the accident (in which they were not involved), even if they certainly contributed to the fact that Peter was born. However A may still be viewed as the cause, if other circumstances required to make B occur are normal expected ones while A is something unusual or the result of some voluntary act.

A second case is when two facts A and A' are each sufficient for making B occur and it turns out that both facts are true (causal redundancy). For instance, after (Halpern 2017), a forest fire may be caused by both dropping a lighted match (A) and by a lightning strike (A'). Then even if A had not occurred, B would still be the case (overdetermination). Here, counterfactual reasoning does not apply as it is not sufficient for distinguishing if a potential cause is an actual cause or not. Sometimes, other aspects of causality are useful to detect the actual cause, like temporality. For instance, consider the case after (Lewis 2000) (see also Halpern 2017), where a glass bottle was shattered by Billy and Suzy throwing stones, while Suzy's shot reached the bottle before Billy's, and Billy would have shattered the bottle, had not it been for Suzy's stone (preemption of a cause with respect to another one). While the counterfactual is false, only the real shot that led to the present state (Suzy's shot) is indeed the actual cause in this case.

Interaction between causes Sometimes, concurrent causes may produce effects that could not be obtained by a subset of them. In such a case, they are said to act interactively. For instance, whereas rain or wind alone cannot knock down a tree, the combination of rain and wind may. Another well-known example is the interaction of several medical drugs.

Thus, an event B can be caused by a combination of elementary causes $\mathscr{A} = \{A_1, \ldots A_k\}$. The set of causes $\mathscr{A}$ is then called the causal complex of B (Hobbs 2005). Let $\mathscr{E}(\mathscr{A})$ be the set of effects of $\mathscr{A}$. A cause-effect relation is said to be monotonic if $\mathscr{A} \subseteq \mathscr{A}'$ implies $\mathscr{E}(\mathscr{A}) \subseteq \mathscr{E}(\mathscr{A}')$. In other words, if you know that B is an effect of the conjunction of causes in $\mathscr{A}$, you can be sure that it is still an effect of $\mathscr{A}$ plus whatever fact C you happen to know to be true besides of $\mathscr{A}$. If $\mathscr{A}$

is a causal complex of B, this means that $\mathscr{A}$ contains everything that guarantees the occurrence of B, and except if there are interacting factors with a negative synergy, adding items to $\mathscr{A}$ will not prevent B from occurring.

If the causal relation is a matter of degree, a graded form of interactivity may be observed (reinforcement), whereby two causes have the same effect, but the concurrent occurrence of the two causes significantly increases the probability or the magnitude of the effect, over and beyond what would be expected by each elementary cause. It appears that human experts have some ability to detect interaction effects (Novick and Cheng 2004).

In contrast with this positive form of interaction, a form of negative synergy (also called prevention) may be encountered if adding a cause to the causal complex of B prevents B from occurring. We say that C prevents $\mathscr{A}$ from having its usual effect B when the causal complex $\mathscr{A}$ should have had effect B, but the occurrence of C has canceled effect B. For instance if the wind blows, the shutter will flap unless it is hooked. Handling negative interaction may require to specify the effects of all combinations of causes (wind and rain, rain and lightning, wind and rain and lightning, etc.), or it may require refined causal rules (e.g., the rain soaks the ground, the wind uproots a tree on a soaked ground). The difficulty here relates to the very large set of qualifications that can apply to a given cause.

Parsimonious covering of effects An event B can be caused by any among several combinations of elementary causes. In other words there can be a disjunction of causal complexes. However one often considers that a small set of elementary causes (especially a single cause) is more likely, or a more plausible explanation of B than a larger set of potential elementary causes. A causal relation satisfies minimality if when a set of events $A_1, \ldots A_k$ is a cause of B, then no other smaller set of causes (in the sense of cardinality) can also be considered as a cause of B. This is the parsimonious covering principle (Reggia et al. 1985a, b).

Minimality is useful to limit the set of candidate causes for an event. Especially, the unique cause assumption is very often made and is an extreme example of this principle. A non-minimal conjunction of causes of B is sometimes called "sufficient cause".

Abnormality Another important notion invoked in several conceptions of causal relations has been pointed out by two legal philosophers, Hart and Honoré (1985), namely, the notion of abnormality. According to these authors, abnormality is one of the chief criteria for selecting causal conditions in everyday life inference. Indeed, in the investigation of human affairs, what we want to explain is generally a deviation from the normal course of events, and abnormal facts are generally privileged when providing causal explanations.

The expected course of the world, and in particular the expected behavior of agents is governed by so-called norms. A behavior that follows a norm can be normal (unsurprising) or normative (mandatory), see, e.g., (von Wright 1963) for a more complete analysis. An approach to normative causal analysis is proposed in Kayser and Nouioua 2009, applied to the analysis of car accident reports. We may consider

that, generally speaking, the normative understanding of causality is a special case of the normal one, because we generally expect agents to respect their duties.

Abnormality can trigger the search for a cause, as it commonly makes more practical sense to look for causes of abnormal situations, as compared to situations that are perceived as completely normal. Furthermore, abnormality can be a filter when looking for a cause, because abnormal situations are more easily attributed to causal factors that are themselves abnormal. Indeed, one may argue that an event is more likely to be interpreted as a cause if it is abnormal than if it is usual. For instance, the presence of oxygen is a necessary condition for an arson, but it is not pointed out as its cause, while the act of the arsonist will be considered as the cause because the act of setting a forest on fire is considered as abnormal, while the presence of oxygen is always taken for granted. Namely, in the common use of the cause-effect relation, a principle of cognitive economy is at play: if, as it is almost always the case, the causal complex contains a large number n of propositions, we cannot afford to memorize, to use or to communicate a proposition such as $A_1 \wedge \wedge A_n$ causes B. As among the n propositions, it turns out that most of them are almost always true (like oxygen present in the fire example), checking their truth value is a waste of time.

Suppose, all A_i with $i \geq 2$ are usually true. It is then by far more convenient to use the rule "A_1 causes B", even if we know that the joint occurrence of A_1 and the negations of A_i for any $i \geq 2$ will no longer cause B. In other words, A causes B does not mean that "every time A occurs, B must occur too", but it is used as a shorthand for "every time A occurs, along with a list of other factors (which can safely be assumed to be true), B must occur too". This is clearly not monotonic. Even worse, a causal rule is allegedly never complete, since we cannot know in advance the number of usually present factors that play a role in the occurrence of B. The idea that causal reasoning is intimately bound to nonmonotonic reasoning can already be found in Shoham (1990, 1991), Simon (1991).

Interventions Interventions are commonly viewed as external actions that force some variables to take some specific values. They play an important role in distinguishing causation from mere correlation, more significantly than mere precedence. As we already observed, a (global) joint probability alone can help to determine correlated or independent events but cannot lay bare causal relations. This is also true for other approaches based on counterfactuals or nonmonotonic reasoning, where background knowledge is concerned with the representation of normality, rather than objective actual causality. To overcome this limitation, intervention can be used to arbitrate between several causal structures that fit the correlation data equally well. This core notion was introduced in early works (e.g., von Wright 1971), but was given its most prominent role by Pearl (2000), whose definition of "A causes B" requires that the forced occurrence of A, by means of an intervention, increases the probability of the occurrence of B.

Intervention clearly plays a useful role in causal analysis: it provides a natural way of understanding actual causation by proceeding to a set of interventions and manipulations. Identifying a causal relationship between different elements of a system is much easier if the agent can directly intervene in the manner of an experimenter and evaluate the effects of such manipulations. An intervention is determined as

coming from outside. In learning causal models interventions are handled using a new operator called "do" that allows to distinguish intervention $do(A)$ from a simple static observation of A. Although intervention is mostly considered in probabilistic frameworks, it can also augment counterfactual-based definitions of causality.

Partial transitivity A causal relation is transitive if and only if whenever "A causes B" and "B causes C" it follows that "A causes C". Transitivity is often expected because people often reason using causal chains of events. The intuition that causation should be transitive has been increasingly questioned by philosophers (Björnsson 2007; Hall 2000; Hitchcock 2001).

There are situations where transitivity looks natural. For instance, suppose the kettle is on the stove. The kettle whistles because the water is boiling. The water is boiling because it is hot enough. It looks natural to conclude that the cause of the whistling of the kettle is that its water has been heated enough. One explanation for such transitivity can be the fact that the only possible cause for the water boiling is that it has been heated enough.

But there are other situations where transitivity sounds debatable. For instance, you are asleep under an apple tree. You wake up because an apple falls on your head. It falls because it was ripe enough. It sounds strange to consider the ripeness of the apple as the cause of your waking up, as there are many other possible causes to make the apple fall, and the ripeness is not sufficient to conclude that the apple will fall on your head. More generally, the transitivity of long chains of causal connections can be challenged. For instance, if you see your life as such a long chain of events, one causing the other, you may conclude that your birth is a cause of your eventual death, a statement that few people would endorse. A famous saying questioning transitivity of causality is due to Montaigne: Eating ham makes you thirsty; drinking quenches your thirst: therefore eating ham quenches your thirst.[1]

So, depending on the intended purpose of the causal analysis, transitivity may look natural, optional, or problematic. See Eells and Sober (1983), Bonnefon et al. (2012) for discussions of conditions for causal transitivity in the probabilistic and the qualitative nonmonotonic setting, and (Halpern 2016) for the transitivity of actual causality.

2.2 The Use of Causality in AI

There are two main classes of problems addressed by Artificial Intelligence: representing knowledge and reasoning about the world so as to make sense of it or to predict future observations in a reasoned way, and devising intelligent machines that can make decisions and act on the world. Both classes of problems need a form of causality to be embedded. In the first class, perceived causes are important as causal

[1] In French, "le jambon fait boire; le boire désaltère: par quoi le jambon désaltère", Michel de Montaigne, Les Essais, Chap. 15, 1580.

reasoning is one way of explaining situations or decisions to a user and causal explanations are a source of arguments. In the second class, the notion of actual cause is very useful to figure out the potential consequences of actions. In the following we review a number of AI problems, some of which are addressed in other chapters of this collection, where causality plays an important role.

- **Prediction** The problem of prediction needs the knowledge of regularities in the world so as to apply this generic knowledge to particular situations where only partial observations are available. It is clear that for this purpose, perceived causality, if validated by sufficient evidence, is enough to predict the values of some variables from the observation of others. Sometimes, even the mere knowledge of correlations may be enough to make plausible predictions. For instance, Bayesian networks and their non-additive variants (like possibilistic or credal networks), reviewed in chapter "Belief Graphical Models for Uncertainty Representation and Reasoning" of Volume 2, are often used for such prediction tasks.
- **Diagnosis** Diagnosis problems are an important class of application of Artificial intelligence. While early expert systems for diagnosis contained uncertain rules that directly predict faults or diseases based on observed symptoms, this type of deductive approach was later on given up, with the emergence of more standard relational, probabilistic or logical approaches that adopt a causal representation. The relational approach is the most basic one where for each fault, the set of possible symptoms is described by means of a relation (Reggia et al. 1985a, b). In a more sophisticated approach the conditional probabilities of observing symptoms are provided and possibly prior probabilities on faults. The problem is then to find combinations of faults that cover the observed symptoms in a case.
  The idea of abnormality is often used in diagnosis, whose aim is to identify the source of an anomaly, faults being considered abnormal. In model-based diagnosis, the idea is to provide a description of how a system normally works when no faults are present (for instance by means of a set of logical formulas, see Reiter (1987)), and the diagnosis method proceeds by abduction, looking for fault variables that must be made active to justify (via deduction, that can be causally interpreted) the anomalous observations. See chapter "Diagnosis and Supervision: Model-based Approaches" of this volume for an overview of model-based diagnosis.
  Minimality (parsimony) is desirable for the purpose of diagnosis. Diagnostic reasoning is indeed based on the conjecture that the set of faulty components is minimal (often with respect to set inclusion criterion, or to cardinality-based criterion). This is the usual assumption made in relational models (see, e.g., Dubois and Prade 2000). A similar case can be made for the more general situation of postdiction. In model-based diagnosis, the description of the normal behavior of the system is supposed to cover possible interactions between faults. When the system fails, however, and particularly in situations of temporal diagnosis, the interaction of the components in the system may be too poorly defined to assess which hypothesis of correct behavior is inconsistent with the observed discrepancies.

- **Action theories: how to achieve a goal for sure** In planning problems, the question is to find a sequence of actions that change the state of a system from an initial state to a goal state. Each elementary action can be interpreted in terms of causal connection between two states, given some preconditions. See chapter "Planning in Artificial Intelligence" in Volume 2 for a survey on planification methods. More generally, we deal with dynamic systems where actions are chosen based on partial observations on the current state (see chapter "Reasoning about Action and Change" of this volume). It is clear that it is the notion of actual causality that is at work here, in what can be termed *goal-directed reasoning*, where an agent can act based on its beliefs, desires and intentions. Namely, an action can be viewed as an actual cause of some change in the system.
  For the purpose of goal-directed reasoning, interaction between actions can be safely ignored only when the plan to the goal can be construed as purely sequential. If, however, two actions are required in parallel (perhaps to simultaneously achieve different parts of the plan), then it would be risky to ignore it, as the two causes might interact in an untoward way, creating unwanted effects.
- **Making Sense and Explanation** Given a stream of data, one may use causal knowledge to understand what is the logic governing these time-stamped observations, thus making sense of them. Such causal and/or taxonomic information can be useful to produce explanations for a user. This is called causal ascription. A general pattern proposed by Besnard et al. (2008a, b) is the following: "If A causes B, and A is not impossible based on the current knowledge, then A explains B". In this approach the causal relations are supposed to be given. Causes are assumed to always produce their effects (there is no uncertainty). If A is a conjunction of elementary causes, one may consider a minimal subset of elementary causes to be a more plausible explanation. For example, Halpern and Pearl (2005a, b) model of causality (see the next section for more details) distinguishes between "sufficient causes" and "actual cause", and only the latter satisfy minimality.
  Agents are more likely to engage in explanation tasks when they are confronted to abnormal situations, and they are likely to focus on abnormal factors as the cause of these abnormal situations (for behavioral evidence, see Hilton and Slugoski 1986; Gavanski and Wells 1989).
- **Causal information helps learning** Learning from data, especially using statistical methods rely on studying correlation. There is hardly any way of detecting causality from mere data. However having causal information about a phenomenon is instrumental in structuring a model prior to learning its parameters. For instance, in learning Bayesian networks, having causal knowledge may help ordering the variables in a proper way and to guess the conditional independence relations. In the last ten years, scholars in machine learning have become aware of the role causality could play in the understanding of data collections such as image repositories. A broad overview of the issues can be found in Guyon et al. (2010). Causality may be searched for in time-stamped data (Guyon et al. 2011), or in data that result from interventions (e.g. medical data obtained by testing the efficacy of drugs on patients) or that can be interpreted as such (images where, for instance,

the observation of a car is seen as causing the presence of wheels (Lopez-Paz et al. 2017)).

In this context, the minimality property states that if a directed graph satisfies the Markov condition with respect to a probability distribution, then no sub-graph of it also satisfies the Markov condition with respect to this probability distribution. For instance, if a causal graph contains an arc from a variable A to a variable B while they are probabilistically independent, then this graph violates the minimality condition. However the minimality criterion (finding the simplest graph) leads to best exploit conditional independences that lie in the data, but it is not clear that the direction of arcs in the obtained Directed Acyclic Graph indicates causality at all, even if there is experimental evidence from cognitive psychology that people infer causal relations at least partly from covariation information.
Finally, there is an impact of the interaction between variables on learning. This impact is in proportion of the strength of the interactions themselves. Weak interactions will appear as noise and be hard to detect, while strong interactions will jam the data and make it hard to learn from.

- **Assigning responsibility** Determining actual causes helps in deciding who's to blame or reward when analyzing a given abnormal situation where something went wrong. Minimality is clearly required for the purposes of praising or blaming. Consider the situation where we blame some agent for her role in the set of events $\mathscr{A}$, which caused an unfortunate outcome B. We certainly want to ensure that there is no subset of $\mathscr{A}$ that caused B, in which the agent played no role. Abnormality plays a role for praising and blaming, because we usually praise or blame agents for extraordinary, unexpected, abnormal courses of action (with good or bad results). In order to assign responsibility, we need to be sure that the resulting effect B was caused by some voluntary intervention, if this intervention made a difference and was intentional. Intentionality is then an important notion to be taken into account for determining whom to praise or to blame.
  Interaction between causes poses hard problems for the purposes of praising and blaming. When the concurrent actions of two agents achieve an untoward effect that would not have been achieved by either action, the issue of who is to blame can be the object of intricate legal contests, going into detailed considerations about the particular knowledge that the agents may have had about the potential interaction. Allowing for interacting causal relations, while advisable, will not by itself solve these hard problems.
  Finally, for the purpose of praise and blame, transitivity is clearly required in some cases, and it is hard to find examples where it would do any harm. Consider first the "do not blame the messenger" situations: Imagine that Tom does something outrageous, which caused Billy to tell Suzy about it, which caused Suzy to be shocked (she would not have known without Billy telling). Clearly, we do not want to blame Billy only for the feelings of Suzy. We would at least want Tom to share the blame, but for this, we need to be able to say that Tom caused the feelings of Suzy, for which we need causal transitivity.

3 Formalizing Causality

In this section we review various mathematical formulations of causal relations that can be found in the literature.

3.1 Relational Models of Causality

The parsimonious covering theory developed by Peng and Reggia (1990) proposes a relation-based formulation of causality dedicated to diagnosis problems where disorders and symptoms can be directly connected. They assume the knowledge of a relation R between potential causes (disorders) A_j and effects B_i (symptoms). A related pair $(A_j, B_i) \in R$, means that A_j *may* directly cause B_i. However this does not mean that A_j necessarily causes B_i. This representation for the "causal" association between A_j and B_i can be understood as a qualitative counterpart to assigning a non-extreme probability to it.

The detection of causes proceeds as follows: given a set $\mathscr{E}$ of effects known to be present (but possibly incomplete) one determines the set of potential causes of $\mathscr{E}$ as

$$\mathscr{C}(\mathscr{E}) = \{A_j, \exists B_i \in \mathscr{E}, (A_j, B_i) \in R\} = \cup_{B_i \in \mathscr{E}} R B_i,$$

where RB_i is the set of causes for which B_i is a possible effect. An explanation of $\mathscr{E}$ is a subset $\mathscr{C} \subseteq \mathscr{C}(\mathscr{E})$ of potential causes that cover the set of effects, i.e., such that

$$\mathscr{E} \subseteq \{B_i, \exists A_j \in \mathscr{C}, (A_j, B_i) \in R\} = \cup_{A_j \in \mathscr{C}} A_j R$$

where $A_j R$ is the set of possible effects of A_j. The set of causes $\mathscr{C}$ is then called a *cover* of $\mathscr{E}$. Peng and Reggia more particularly look for so-called "parsimonious covers", i.e. covers of $\mathscr{E}$ that are relevant, i.e., each element in $\mathscr{C}$ potentially causes one effect in $\mathscr{E}$ (i.e., $\forall A_j \in \mathscr{C}$, $A_j R \cap \mathscr{E} \neq \emptyset$), irredundant (none of the proper subsets of $\mathscr{C}$ is also a cover of $\mathscr{E}$) and minimal (the cardinality is the least among all covers of $\mathscr{E}$). In this very elementary model, it is supposed that disorders are independent so that their effects accumulate and do not interfere (otherwise relation R should associate a *subset* of elementary causes to each effect).

In the above approach, there is no idea of sufficient cause, which would need a wider framework using another relation where causes or groups thereof *necessarily* produce some effects. For instance, Dubois and Prade (1995, 2000) use two relations R^+ and R^- that to each cause A_j respectively associate the set $A_j R^+$ of effects that are certainly produced by A_j alone, and the set $A_j R^-$ of effects that certainly cannot be caused by A_j alone. The causal relation R of the previous paragraph corresponds to the complement of R^-. This bipolar modeling leaves room for expressing that an effect B_j is either a sure effect or is only a possible effect of A_j. Moreover, the available evidence is now composed of two parts: a set $\mathscr{E}^+$ of certainly present

effects, and a set of certainly absent ones $\mathscr{E}^-$. A cover of $(\mathscr{E}^+, \mathscr{E}^-)$ is a set $\mathscr{C}$ of causes such that their sure effects are not among the absent effects $\mathscr{E}^-$, and such that their impossible effects are not among the observed effects $\mathscr{E}^+$.

There exist graded versions of these models. The degree of causal dependence $R(A_j, B_i)$ between the cause A_j and the effect B_i is interpreted as the probability of the causation event "A_j *causes* B_i" when A_j is present by Peng and Reggia (1990), and it comes close to Bayesian diagnosis methods, although this probability differs from $P(B_i|A_j)$. In the valued extension of the bipolar relational model, the degrees of association correspond to the degree of certainty and of possibility that B_i be an effect caused by A_j in the setting of possibility theory. In both cases, the degree of causal relationship reflects uncertainty. The available observations may be also pervaded with uncertainty. Fuzzy relational models, as proposed and developed by Sanchez (1977) and his followers, rather handle the *intensity* of presence of manifestations or disorders. See Dubois and Prade (2000) for an overview of relational approaches. Although such an approach could in principle be extended to the study of a cascade of causal relations, transitivity is not an issue here, since only direct causal relationships are exploited.

3.2 Modal Logic Setting for Counterfactual Causality

A modal logic approach to counterfactual causality is due to von Wright (1963); see Demolombe (2000) for an introductive summary and a discussion. Events considered by von Wright may involve explicit agents, and thus actions as well. He gives a modal account of conditions under which it can be said that an action causes property B to be true, in terms of possible worlds. Two different accessibility functions are used to account for one single action performed by an agent i: let $d_i(w)$ be the world which obtains when agent i performs the action in world w, and $e_i(w)$ the one which obtains when not doing it (e_i stands for an empty action, which does not preclude spontaneous changes). For instance, when agent i opens an (unlocked) window, let w_q be a quiet world and w_w a windy one, then in $d_i(w_q)$ as well as in $d_i(w_w)$ the door is open (by the agent), in world $e_i(w_q)$ it is not, while it is in $e_i(w_w)$ (since opened by the wind). The claim that then action has caused B can hold in two different ways:

- when $\neg B$ holds in w and in $e_i(w)$, but B holds in $d_i(w)$ (using modality Br_i, where $Br_i B$ stands for "i's action brings about B"),
- or when B holds in w and $d_i(w)$ but not in $e_i(w)$ (using modality Ss_i, where $Ss_i B$ stands for "i's action sustains B), i.e., the agent's action keeps B true, while some other cause would make it false.

Continuing the example, $Br_i open$ is true at w_q and $Ss_i open$ should be true for the action "keeping the window open" in the world w_w.

Von Wright also considers the case of "omitting to bring about B" when, B is false in w and $e_i(w)$, and remains false in $d_i(w)$, as well as a modality "omitting to

sustain". There are altogether eight possible situations, according to whether B is true or false in w, $d_i(w)$ and $e_i(w)$.

This description of cause involves a form of counterfactual situation through $e_i(w)$ since $Br_i B$ can be read: but for the action of agent i, B would have remained false. Kanger (1972), Pörn (1977) consider non deterministic actions which can access a set of worlds (described with the help of accessibility relations), e.g., after pushing the window, it can be open or broken, and when the window is not pushed, it may remain closed or be open by the wind. Hilpinen (1997) considers complex actions made of simultaneous elementary ones, i.e., interaction between actions.

Von Wright analysis also involves intentionality through the distinction made between action and event. This is particularly clear in the definition of omission: when $\neg B$ holds in w, $e_i(w)$, $d_i(w)$, it could plainly be said that the action has no effect at all. Stating that the agent omitted to bring about B rather considers that he has free will and could have done another action. Through the Ss_i modality, the approach can easily account for prevention.

Demolombe (2012) has recently proposed an extension to several agents acting together of the "bringing it about" operators. A joint action operator is defined which possesses the property of non monotonicity with respect to sets of agents. It is refined in a restricted joint operator for cases where several sets of agents independently cause a state of affairs and it is extended to sets of agents who are acting indirectly.

3.3 Probabilistic Modeling of Causality

Probability theory plays an important role in the modeling of uncertain cause and effect relations (Eells 1991). In practice, causes are not always followed by their effects, and effects may appear without their specified causes. Causation is not distinguished from correlation in the standard quantitative, probabilistic definition of causation originally discussed by Good (1961, 1962) as already seen in the previous section ("A causes B" when the probability of B increases in the light of the new information A or equivalently $P(B|A) > P(B|\neg A)$), since this condition is in fact symmetrical. One solution to distinguish causation from mere correlation is to augment the definition with temporality. Other augmentations address the issue of spurious causation, i.e., that two correlated events may actually be the effects of some common cause (see, e.g., Simon 1954). Spurious correlation is avoided provided that (i) A is a prima facie cause of B, and (ii) there is no event Z being observed before A and B such that A and B are independent given Z.

A major problem in using the last definition in practice is related to the fact that if an event Z has not been reported, it does not mean that Z has not occurred. Worse, if Z has been reported, with $P(B|Z) \neq P(B|Z, A)$, then one should not discard Z. Indeed, it may happen that another event Y has been observed such that $P(B|Z, Y) = P(B|Z, A, Y)$ prior to observing A and B. In this case, the conjunction $Y \wedge Z$ inhibits the fact that A causes B. The problem is solved if the set of variables involved in the problem is fully determined, and the independence

relations are completely described. This is the case with Bayesian networks, often used for representing uncertain causal relations, as described in next section. However, in the Bayesian network formalism, causality is captured by the directions of arcs in the graph, not by the multidimensional probability distribution encoded by the conditional probability tables.

Probabilistic causality viewed as correlation (plus a precedence constraint) is not necessarily transitive (Eells and Sober 1983). The fact that $P(B|A) > P(B|\neg A)$, namely, A probabilistically "causes" B, together with the fact that $P(C|B) > P(C|\neg B)$, namely, B probabilistically "causes" C, does not always imply that $P(C|A) > P(C|\neg A)$, namely, A probabilistically "causes" C; see Bonnefon et al. (2012) for counter-examples. It is shown by Eells and Sober (1983) that this probabilistic causation is transitive as soon as the causal chain (on events and not on variables) is Markovian, namely if the two following conditions hold: $P(C|A, B) = P(C|B)$ and $P(C|A, \neg B) = P(C|\neg B)$, namely C and A are independent in both contexts B and $\neg B$. These conditions hold in Bayesian networks, except that in Bayesian networks independence relations are expressed in terms of variables.

However, probabilistic causation is nonmonotonic. Indeed, one may have $P(B|A) > P(B) > 0$ but $P(B|A \wedge C) \leq P(B)$ (and even $P(B|A \wedge C) = 0$ (a case of prevention)). In the presence of several potential causes $A_1, \ldots, A_n$ of an event B, one may select an event A_i that optimizes the difference $P(B|A_i) - P(B)$ as the best cause.

3.4 Causal Bayesian Networks and Interventions

A causal Bayesian network is a Bayesian network where directed arcs of the graph are interpreted as elementary causal relations between variables (Pearl 1988). When there is an influence relation between two variables, intervention allows to determine the causality relation between these variables. In this case, arcs between variables should follow the direction of the causal process. Pearl (1994, 2000), following ideas of manipulation on probability distributions (Spirtes et al. 1993), has proposed an approach for handling interventions in causal graphs based on a *do* operator that sets a group of variables to prescribed values, and checks its effect on the probability of their direct children. Note that causal relations expressed by graphs only concern variables, not complex events. Causal Bayesian networks organize causal knowledge in terms of a few basic mechanisms, each involving a relatively small number of variables. Each intervention entails local change at the level of only one parents-child relation (Pearl 1994, 2000). The parents-child relation at the level of each variable A_i is governed by a local probability distribution $P(A_i|Par(A_i))$ where $Par(A_i)$ is the parents set of A_i. The joint probability distribution is computed using the chain rule:

$$P(A_1, \ldots, A_n) = \prod_{i=1,\ldots,n} P(A_i|Par(A_i)).$$

An intervention forcing a variable A_k to take the value a_k is denoted by $do(A_k = a_k)$ or $do(a_k)$ for short. The intervention consists of making A_i (say) true independently from all its other direct causes (i.e., parents). Graphically, this modification is represented by the deletion of links from the set of variables $Par(A_k)$ pointing to A_k. The resulting graph is said to be *mutilated* and its joint probability is defined by:

$$P(A_1, \ldots, A_n)_{do(a_k)} = \begin{cases} \frac{P(A_1,\ldots,A_n)}{P(A_k|Par(A_k))} & \text{if } A_k = a_k, \\ 0 & \text{otherwise.} \end{cases}.$$

It is clear that the result of an intervention depends on the structure of the causal graph and not only on the joint probability.

There is a difference between observing $A_k = a_k$ and enforcing it. In the first case, one can predict the probability of other variables by conditioning the joint probability associated to the original Bayesian network. In the case of an intervention, we must condition the joint probability associated to the mutilated graph. With respect to the initial graph, the result of an intervention, is similar to Lewis (1976)'s imaging, where the mass of an excluded state is transferred to the closest remaining state, for some closeness relation to be defined according to the application. Here the masses of states that are ruled out by the intervention are transferred to a special set of states that share the same values of parent variables $Par(A_k)$, as pointed out by Pearl (2000) (see also Kyburg 2005). The difference between observations and interventions has been confirmed by psychological studies (Lagnado and Sloman 2005).

Interestingly enough, the "do" operator has been first proposed by Goldszmidt and Pearl (1992) (see also Goldszmidt and Pearl 1996) within the framework of Spohn (2012)'s ranking functions. Besides, Spohn (2006)'s view of causation, which follows the tradition of the probabilistic paradigm (Glymour et al. 1987) (including intervention see Spohn (2000)) is also quite in agreement with Lewis' view in terms of counterfactuals (Huber 2011). Spohn's ranking functions have strong relationships with possibility theory (see chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" in this volume). So it should not come as a surprise that the same intervention technique can be applied as well to possibilistic networks described in Chapter "Belief Graphical Models for Uncertainty Representation and Reasoning" in volume 2 of this treatise (see Benferhat 2010; Benferhat and Smaoui 2011; Ayachi et al. 2014 for details). Moreover, the idea of intervention has been also applied to evidential networks (based on Shafer's belief functions) (Boukhris et al. 2014).

3.5 Shafer Trees Approach to Causal Conjectures

Somewhat in the spirit of Lewis (1986)'s idea of "causal explanation", Shafer (1996, 1998) (see also Shafer et al. 2000) has proposed a "causal logic" that aims at describing the possible relations of concomitance between events when actions take place. The approach is based on the notion of "event tree", where a node corresponds to

a situation (which is not supposed to be a fully detailed state of the world), and the tree represents the temporal chain between instantaneous events, and thus encodes a precedence relation. A situation can be represented by a unique node, or by a set of nodes, called "clade", appearing on different paths. This enables a more refined description of contexts. Moreover, without any reference to a particular tree of events, an "event space" is introduced, where refinement relations and precedence relations can be stated between situations. Then five basic relations between two instantaneous events S, T in a tree are defined:

1. S *refines* T (if S happens, then T happens as well at the same time);
2. S *requires* T (S can only happen if T has already happened);
3. S *foretells* T (if S happens, then T must happen later);
4. S *forebears* T (if S happens, then T may happen);
5. S *diverges* T (if S happens, T cannot happen).

More generally S may be a unique node and T a clade, or both S and T are clades. Many other relations can be defined from these five primitive ones. For instance, "S entails T" if and only if when S happens, T already happened, happens at the same time or is going to happen.

The associated logic has proposition formulas (to represent propositions that are true, false, or indeterminate in situations), and event formulas (to represent situations themselves, thought of as instantaneous events that happen or fail, rather than being true or false). In this approach, the relation "A causes B" is limited to the case where A is an action and B is an instantaneous event. Shafer causal logic is about reasoning on what takes place, what took place, what will take place, what may take place, what cannot take place, depending on instantaneous events. It clearly involves ideas of temporality (through precedence), necessary connexion, and correlation as core notions.

This framework, whose intuitions partly come from probability theory, can be augmented with probabilistic information (Shafer 1996, 1998). However, the use of upper and lower probabilities rather than exact probabilities (expressed in terms of a class of cautious gambles) is necessary to provide the general representation of probability in event spaces. Moreover, Shafer (1999) emphasizes the point that the phrase "X causes Y", where X and Y are variables, is vague, and that very different causal relations can be defined in the probabilistic setting from regularities in terms of conditional probability, conditional expectation, or linear regression, in particular.

The causal logic approach, which shares concerns with the logics of action, but is also motivated by foundational issues for probability, seems to have had a limited impact in the artificial intelligence literature on causality until now. It should however be of interest for prediction, and explanation purposes especially. This setting might be also useful for assessing responsibility; see Shafer (2000) for a preliminary discussion on this issue.

3.6 The Preferential Approach to Plausible Causality and Abnormality

This approach, proposed in Dubois and Prade (2005), and further developed in Bonnefon et al. (2008), especially addresses the task of making sense of a reported sequence of facts, in the light of beliefs supposedly entertained by an agent as to what is normal and what is not. More precisely, consider a sequence of time-stamped facts such as $\neg B_t, A_t, B_{t'}$ is reported, where t' denotes a time instant strictly after t (B_t means that B is known to be true at time t). This reads: B was false, A took place, then later B became true.

Besides, the agent that receives the sequential information $\neg B_t, A_t, B_{t'}$ is supposed to have some knowledge on what is the normal course of the world in context C, which is the conjunction of all other facts known by, or reported to, the agent, and maybe also in context $C \wedge A$, regarding B. Thus the approach explicitly refers to what the agent holds as background knowledge. The agent knowledge is supposed to be made of default rules. Namely, the agent may either believe that $C \vdash\!\!\sim B$ (B is expected to be true in context C), or that $C \vdash\!\!\sim \neg B$ (B is expected to be false), or that $C \not\vdash\!\!\sim B$ and $C \not\vdash\!\!\sim \neg B$ (the truth or the falsity of B is contingent in context C), where $\vdash\!\!\sim$ is a nonmonotonic consequence relation (Kraus et al. 1990) describing what is normal, and $\not\vdash\!\!\sim$ stands for its negation. $C \not\vdash\!\!\sim B$ means that $C \vdash\!\!\sim \neg B$ is not in the background knowledge of the agent. Similarly, in context $C \wedge A$, the agent may have the same form of belief on the normal course of things. It is assumed, that $C \wedge A$ is consistent (otherwise, the fact that A becomes true would be incompatible with context C). In case the agent knows $C \wedge A \vdash\!\!\sim B$, B is expected to be true after A took place, and the sequence $\neg B_t, A_t, B_{t'}$ is in conformity with the agent's knowledge; on the contrary, if the sequence $\neg B_t, A_t, \neg B_{t'}$ is reported, it would mean that event A had an abnormal, surprising behavior.

In Bonnefon et al. (2008), the two following definitions of *facilitation* and *causality ascriptions* are proposed. They rely on pieces of default knowledge (so the approach heavily uses abnormality as a core notion), and a sequence where a change of the form $\neg B_t, A_t, B_{t'}$ is reported (thus making also reference to temporality). Given a context C, and the knowledge base of the agent,

- if the agent believes that $C \vdash\!\!\sim \neg B$, and if for the agent $C \wedge A \not\vdash\!\!\sim \neg B$, the agent will perceive A as having *facilitated* the occurrence of B in context C (denoted $C : A \Longrightarrow_{fa} B$);
- if the agent believes that $C \vdash\!\!\sim \neg B$, and that $C \wedge A \vdash\!\!\sim B$, the agent will perceive A as *being the cause* of B in context C (denoted $C : A \Longrightarrow_{ca} B$).

Bonnefon et al. (2008) report experiments that indicate the cognitive validity of these notions of facilitation and causation. These definitions have noticeable properties. In particular, it has been shown that

- Each of $C : A \Longrightarrow_{ca} B$ and $C : A \Longrightarrow_{fa} B$ implies $C \not\vdash\!\!\sim A$. This means that only abnormal events in a context may be regarded as a cause, or a facilitation.

- A restricted transitivity property holds: If $C : A \Longrightarrow_{ca} B$, $C : B \Longrightarrow_{ca} D$ and $B \wedge C \mid\sim A$ hold, then $C : A \Longrightarrow_{ca} D$.
- The two properties hold for $\Longrightarrow_{ca}$ provided that $\mid\sim$ is a preferential entailment in the sense of Kraus et al. (1990). The first property holds for facilitation ($\Longrightarrow_{fa}$) if $\mid\sim$ is a rational closure entailment (Lehmann and Magidor 1992).

Note that here transitivity requires the "saliency" condition $B \wedge C \mid\sim A$, i.e. it means that the normal way to have B (in context C), is to have A. For instance, for $A =$ *drinking*, $B =$ *inebriate*, D: *staggering*, we have *drinking* $\Longrightarrow_{ca}$ *inebriate* and *inebriate* $\Longrightarrow_{ca}$ *staggering* entail *drinking* $\Longrightarrow_{ca}$ *staggering*, since *inebriate* $\mid\sim$ *drinking*. See Bonnefon et al. (2012) for a thorough discussion and empirical material supporting the fact that humans do not always endorse transitivity for causality relations, but largely acknowledge it in case the saliency condition holds.

Other types of cognitive situations can be captured by this model. For instance, a companion case to the two previous situations is the interpretation of the sequence $\neg B_t, A_t, B_{t'}$ for an agent for whom neither $C \mid\sim \neg B$ nor $C \mid\sim B$ holds, while $C \wedge A \mid\sim B$ is part of the agent's background knowledge; then A may be perceived by the agent as a kind of justification for having $B_{t'}$(Bonnefon et al. 2008). The approach also captures the ideas of necessary (or enabling) condition (as oxygen for a fire), of prevention to persist, and of prevention to take place, and has also some potential for the assessment of responsibility (according to agent's background knowledge (Prade 2008)). Generally speaking, the approach appears to be suitable for making sense of a sequence of reported events; see Chassy et al. (2012) for a cognitive discussion and an implementation.

Besides, this approach does not embed the notion of intervention and thus cannot readily distinguish spurious correlation from causation. Nonetheless, the approach could be extended in that direction, since both this approach and graphical models can be encoded in a possibilistic setting, see (Benferhat et al. 2009).

Lastly, it has been noticed, (e.g., in Besnard et al. 2008b) that if "A causes B", and $B \models B'$ (classical logic entailment), it does not follow that "A causes B'". For instance, saying that a disease may give a fever in the range $[38 - 39]$°C, does not mean that the disease may give a fever in the range $[38 - 40]$°C. The nonmonotonic consequence relation $\mid\sim$ is such that $A \mid\sim B$ entails $A \mid\sim B'$ if $B \models B'$. The two conditions involved in the definition of causality, namely $C \mid\sim \neg B$ and $C \wedge A \mid\sim B$ are precisely encoded in a possibilistic setting by $N(\neg B|C) > 0$ and $N(B|A \wedge C) > 0$ respectively, where N is a necessity measure. If we want to express that (i) the counter-models of B are not possible effects of A and that (ii) any model of B is an effect guaranteed to be possible, we have to add a third condition, namely $\Delta(B|A \wedge C) > 0$ where Δ is a guaranteed possibility measure, as suggested in Dubois and Prade (2005) (see chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" in this volume for set functions Δ that are decreasing with respect to inclusion).

3.7 Actual Causality: Action Logic

In the scope of reasoning about actions, in contrast with other tasks involving causality like diagnosis or explanation, the problem is to compute the results of actions. In order to describe actions, in terms of their preconditions and their consequences, many formalisms propose to encode the link between an action and its consequences by "causal implication". This choice seems natural since the material implication is not well suited to encode actions, especially due to the "frame problem" (see chapter "Reasoning about Action and Change" in this volume), which amounts to finding a way to express the persistence of the values of all the fluents that are not affected by an action. Indeed this problem is closely related to causal relations, because it is important to select as potential cause of the change of value of a fluent only the actions that might affect that fluent. We are not going to describe all the approaches proposed in order to address the frame problem but we only focus on the formalisms that explicitly refer to causality and we briefly outline how this notion has been handled in these formalisms.

- In Stein and Morgenstern (1994) a "motivated action theory" is proposed, which is based on situation calculus but extended to allow for concurrency of actions and to integrate causation as primitive. A causal statement has the following form CAUSES(preconditions, act, postconditions) meaning that if "preconditions" holds when "act" occurs then "postconditions" will hold in the resulting situation. Given a description of what holds and what occurs in a scenario, and a background knowledge consisting of generic knowledge true in every situation (that may include CAUSES statements) the aim is to deduce the facts that follows from all this information. In order to to deal with the frame problem Guyon they use a kind of frame axiom stating that the value of a fluent persists if either it is not in a CAUSES statement or the preconditions of this statement did not hold or the action did not occur. In order to ensure that CAUSES statements differ from material implication the authors introduce the notion of motivated facts and actions. Roughly speaking a formula is motivated if it has to happen.
- Lin (1995) has introduced a new predicate $Caused(f, v, s)$ meaning that the fluent f is caused to get the truth-value v in state s. This predicate is used in the situation calculus framework. The situation calculus language is a many-sorted first-order one, with three particular sorts: "situation", "action" and "fluent". Once state constraints and direct effects of actions are described, persistence assumptions are required in order to achieve prediction. This is done by assuming that unless caused, the truth value of a fluent will persist, in other words, things that are not caused do not change.
- Another type of approach is to use a unary modal operator "is caused". Geffner (1990) proposes to define a causal theory by a default theory augmented with such a causal operator. The modal operator allows us to distinguish between a fact that holds and a fact that has a causal explanation. Defeasibility is handled by using

abnormality atoms that are expected to be false. On this basis, Giordano (1998) has defined a dynamic logic for actions and causality in order to address the ramification problem and to deal with persistency. Two modalities are used in their work, the first one is Geffner's modal operator, and a second one is used for representing actions. Turner (1999)'s proposal, called Universal Causation Logic (UCL), can be embedded in Geffner's theory but it is less complex since it does not aim (like Geffner's) at working on a unifying framework for nonmonotonic inference but, more specifically, at providing a mathematically simple modal nonmonotonic logic designed for representing commonsense knowledge about actions. This is done on the basis of a "principle of universal causation" expressing that what one obtains in the world is exactly what is caused in it.

- In the proposal presented by Giunchiglia et al. (2004) a modal logic formalism is also used. The semantics of their causal theory is based on the principle of universal causation: "every fact that is obtained is caused and vice versa" as in Turner's semantics. In their proposal, the default persistence assumption used to handle the frame problem ("normally, the values of fluents do not change") is replaced by: "the value of a fluent does not change unless a different value is caused by performing an action".
- The proposal of Thielscher (1997) is based on propositional logic. This approach starts from a different input: instead of having a set of causal laws, the causal laws are derived from a given relation I called "Influence Information" and from a set of propositional constraints. The domain knowledge is encoded as a set of constraints, and causal laws are applied iteratively until the result obtained violates some constraint.
- Combining ordinary actions encounters the difficult problem of indirect effects: breaking the glass of wine also means that the wine inside spills out. Giving a systematic account of these effects is difficult, and it matters for ascription: Billy's clumsiness, which caused the glass of wine to break, also caused the tablecloth to be spoiled. In order to overcome the difficulty, numerous researchers such as Pearl (1988), Geffner (1990), Lin (1995) have introduced a distinction between caused and uncaused changes. In the modal framework, McCain and Turner (1995) use a modal structure to semantically account for indirect effects by reasoning at the individual world level. The difference is made by the notion of "causally explained world", i.e. not only the accessed world differs from the source world as predicted by causal rules, but they must not have unpredicted differences. In the same vein, Giordano et al. (2000) use a modal operator to mean that a new proposition is caused. This modal operator follows the laws of the modal logic K. In White (2002), caused transformations are taken as an accessibility relation between worlds and sequent calculus proof rules are proposed. In Bochman (2003), the author uses yet another modal approach equipped with a quasi-reflexive accessibility relation, where accessibility is limited to causally explained worlds.

3.8 The Halpern and Pearl Approach

Halpern and Pearl (2001, 2005a) propose a model devoted to the identification of "actual causes." The model borrows from an approach due to Galles and Pearl (1997, 1998), who distinguish between "exogenous" and "endogenous" variables forming respective sets $\mathscr{U}$ and $\mathscr{V}$. Exogenous variables have their values determined by outside factors. The values of endogenous variables are determined by the values of exogenous variables. The values assigned to endogenous variables are governed by a so-called (non-linear) functional model described in terms of structural equations (Pearl 2000) of the form $X =: f(V_1, \ldots V_i, U_1, \ldots U_j)$ where $\{V_1, \ldots V_i\} \subseteq \mathscr{V}$ and $\{U_1, \ldots U_j\} \subseteq \mathscr{U}$. Structural equations are directed in the sense that we cannot compute the value of any $Y \in \{V_1, \ldots V_i\}$ from the value of X and other variables.

A causal model, after (Galles and Pearl 1997, 1998), is denoted by $M = (\mathscr{U}, \mathscr{V}, \mathscr{F})$ where $\mathscr{U}$ and $\mathscr{V}$ are sets of endogenous and exogenous variables, respectively, and $\mathscr{F}$ is a set of structural equations determining all endogenous variables. There is one structural equation per endogenous variable, and no equation determining the exogenous variables. Given the values $U = u$ of all exogenous variables $U \in \mathscr{U}$, which is called a context, the values of endogenous variables are completely determined by the structural equations. Only endogenous variables can be causes or be caused.

The relationship between a functional model and a Bayesian network whose arcs are causally interpreted is that if we consider all variables in the Bayesian net as endogenous, we can replace each node X and its parents by a structural equation of the form $X = f(PX, U_X)$, where PX represents the parents of X and U_X is a new exogenous variable, and define a joint probability P on exogenous variables $\mathscr{U}$ such that for each node X, the conditional probability table $P(X|PX)$ is determined by P (Druzdzel and Simon 1993).[2] In this way, probabilities (e.g., randomness) only bear on exogenous variables and the Bayesian network is replaced by deterministic structural equations. Structural equations leave room for interventions, where the values of some endogenous variables are fixed independently of the values of exogenous variables.[3]

The causal model described above can be thus represented using a graph, in which nodes correspond to variables in $\mathscr{V}$, and an edge from X to Y exists if the value of Y depends on the value of X. This graph is a directed acyclic graph (DAG) representing the relationships between variables that are fully specified by structural equations.

When an endogenous variable X depends on another endogenous variable Y, it means that if we fix the values of all variables other than X and Y, varying the value

[2]This work has its roots in works on causal ordering, emphasizing the directed nature of causation, in econometrics, with the pioneering works of Wright (1921), Haavelmo (1943) (see Pearl 2015), continued in early works by Simon (1952, 1953, 1954), Simon and Rescher (1966). Later on, a debate took place about causal ordering and its use for qualitative reasoning in diagnosis (Iwasaki and Simon 1986a, b; de Kleer and Brown 1986).

[3]The interest of *linear* structural equation models for analyzing non-trivial causal phenomena has been emphasized in the recent years by Pearl and his co-authors, see, e.g., Pearl (2013), Chen et al. (2014).

of Y results in a variation in the value of X through the structural equations. This kind of dependence is not transitive. It captures the counterfactual view of dependence.

A causal setting is a causal model M with a context $\mathbf{u}$. An elementary event consists in observing the value $X = x$ of an endogenous variable. An event ϕ is a Boolean combination of elementary events. A cause will take the form of a conjunction of elementary events, namely, $\mathbf{X} = \mathbf{x}$ for a tuple $\mathbf{X}$ of endogenous variables. Halpern (2017) writes $(M, \mathbf{u}) \models \phi$ when ϕ is true (actually happens) in the causal setting $(M, \mathbf{u})$. The intervention consisting in setting $\mathbf{Y}$ to $\mathbf{y}$ in a causal setting is denoted by $\mathbf{Y} \leftarrow \mathbf{y}$. Halpern denotes the fact that ϕ is true in the causal setting $(M, \mathbf{u})$ under intervention $\mathbf{Y} \leftarrow \mathbf{y}$ by $(M, \mathbf{u}) \models [\mathbf{Y} \leftarrow \mathbf{y}]\phi$.

Halpern (2015, 2017) (originally with Pearl (Halpern and Pearl 2005a)) has proposed several variants of the definition of what an actual cause is, based on the structural equation approach. Indeed, the pure counterfactual definition becomes insufficient in cases when several facts can each stand as a cause for a given state of affairs. One of these definitions goes as follows: The event $\mathbf{X} = \mathbf{x}$ is said to be an actual cause of an event ϕ if and only if:

(1) $(M, \mathbf{u}) \models \phi$ and $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x}$ (ϕ and $\mathbf{X} = \mathbf{x}$ hold in the causal setting $(M, \mathbf{u})$).

(2a) There is a set $\mathscr{W}$ of endogenous variables not appearing in $\mathbf{X}$, and a setting $\mathbf{x}'$ of $\mathbf{X}$ such that if $(M, \mathbf{u}) \models \mathbf{W} = \mathbf{w}$ then $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}]\neg\phi$, namely, if $\mathbf{X}$ is set to $\mathbf{x}'$ and $\mathbf{W}$ is set to $\mathbf{w}$ then ϕ becomes false.

(2b) There is a set $\mathscr{W}$ of endogenous variables not appearing in $\mathbf{X}$ such that $(M, \mathbf{u}) \models Z = z$ for all endogenous variables $Z \notin \mathscr{W}$, and for all subsets $\mathscr{W}'$ of $\mathscr{W}$ and $\mathscr{Z}' \subseteq \mathscr{V} \setminus (\mathscr{W} \cup \mathscr{X})$, we have $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W}' \leftarrow \mathbf{w}, \mathbf{Z}' \leftarrow \mathbf{z}]\phi$. Namely, ϕ remains true if all variables in $\mathscr{Z}'$ and in $\mathbf{X}$ take their values in the causal setting $(M, \mathbf{u})$, even if those in $\mathbf{W}'$ are chosen otherwise.

(3) The subset X is minimal.

In the condition $2(a)$, the set $\mathscr{W}$ stands for variables that may influence the value of ϕ and this condition expresses a counterfactual effect. However, as pointed out in Halpern and Hitchcock (2015), it leaves room to the fact that it may be necessary to intervene on the value of some variable(s) in $\mathscr{W}$ to allow the effect of the value of $\mathbf{X}$ on the value of ϕ to manifest itself. Condition $2(b)$ is supposed to deal with overdetermination (each of two elementary variables influences the value of ϕ) and preemption (like in the bottle shattering case where Suzy hits before Billy does). To quote reference (Halpern and Hitchcock 2015): "The role of condition $2(b)$ is to limit the "permissiveness" of condition $2(a)$ by ensuring that the change in the value of $\mathbf{X}$ alone suffices to bring about the change from ϕ to $\neg\phi$; setting $\mathbf{W}$ to $\mathbf{w}$ merely eliminates possibly spurious effects that may mask the effect of changing the value of $\mathbf{X}$". As mentioned earlier, in this model, causality is not transitive, see Halpern and Pearl (2005a) for a counter-example. The computational complexity of structural equations-based causality is studied by Aleksandrowicz et al. (2017).

Halpern (2017) shows that each variant of the above definition has its own subtleties and merits in dealing with tricky examples, but he does not make a definite choice between the variants. The difficulty of the approach stems from the rather

complicated statements of each variant of the definition of an actual cause in a considered setting, where causes are viewed as conjunctions of facts, and effects as general propositions. See the recent paper by Batusov and Soutchanski (2018) for a renewed definition of actual cause in the context of situation calculus basic action theories.

Further discussions on the limitations of structural equations, especially for capturing causality in preemption problems can be found in Hall (2007), Hitchcock (2009). In particular, the distinction between default and deviant situations (Hall 2007) may help applying the counterfactual criterion of causality, when in the absence of the main cause, the presence other potential causes is perceived as exceptional. In the case when an effect can only be produced by a conjunction of causes, this distinction is useful as well as it is the deviant event that will be regarded as the cause. The framework based on structural equations can be thus augmented with a theory of "normality" or "typicality" accounting for the fact that people privilege abnormal facts in their causal ascriptions (Hart and Honoré 1985; Knobe and Fraser 2008), leaving "normal" things aside (as, e.g., the presence of oxygen, in a fire case) (Halpern and Hitchcock 2015).

In companion papers, (Halpern and Pearl 2001a, 2005b) (see also Halpern 2017) provided a definition of explanation based on their definition of causality. An explanation is a fact which is not known as certain, but such that if it were, would constitute a *sufficient* cause (Halpern 2017) of the fact to explain. As an explanation depends on the epistemic state of an agent, the inference of an explanation supposes to consider every context judged possible by the agent.

3.9 Psychological Models

Although this chapter focuses on AI models of causation, it is worth noting that cognitive psychology also offers several such models. We mention two well-known ones, which rely on very different settings: classical logic and neural nets.

Thagard (1989)'s theory of explanatory coherence (see also Thagard and Verbeurgt 1998; Thagard 2000) views causal ascriptions as attempts to maximize explanatory coherence between propositions. Maximizing coherence would lead to accept the most plausible hypotheses that explain an event that took place (and reject the alternative hypotheses). In this model, if one proposition explains another, then there is a positive constraint between them. Negative constraints result from events that prevent or are inconsistent with other events. See Benferhat et al. (2008) for a short presentation of this approach, including its connectionist implementation, where nodes in the neural network represent propositions, that can be more or less accepted or rejected, as a result of the computation.

A mainstream psychological model of causality is the mental model theory of *naive causality* (Goldvarg and Johnson-Laird 2001). This theory offers a psychological model of how people mentally represent causal relations and reason from them. Its definition of causality is primarily based on the ideas of temporality and necessary

connection. According to the principle of temporal constraint, the causal claim "A causes B" implies that B does not precede A in time. Such a claim is represented as a triple of possibilities (i.e., Boolean valuations over the pair of formulas A and B) that would make it true, namely: $\{AB, \neg AB, \neg A\neg B\}$. This representation, based on material implication, thus assumes a necessary connection between A and B. It defines "A causes B' as A logically implies B and B does not precede A. As a consequence, causation is unrestrictedly transitive.

In this approach, the statement "A allows B is represented by the triple $\{AB, A\neg B, \neg A\neg B\}$. Thus, "$A$ allows B" is taken to mean that B cannot happen without A, but that A is insufficient to produce B on its own. Note that the definition of "A allows B" is formally equivalent to the definition of "not-A causes not-B". Similarly, prevention is represented by a different triple of possibilities, in fact, by the triple of possibilities $\{A\neg B, \neg AB, \neg A\neg B\}$ corresponding to "A causes not-B".

3.10 Towards Comparing Models

As can be seen, there are quite a number of proposals for modeling causality ranging from simplistic to sophisticated ones, using different representation settings (classical logic, modal logic, probability theory, possibility theory, relation calculus, neural nets, ...), and aiming at the treatment of specific or more general artificial intelligence tasks. The idea of actual causality has been also used in databases, where it has been proposed to take advantage of the lineage of answers to a query for finding their causes and computing a degree of responsibility of a tuple with respect to an answer, as a basis for explaining unexpected answers to a query. The idea there is that "tuples with high responsibility tend to be interesting explanations to query answers" (Meliou et al. 2010, 2014). See also (Bertossi 2018) for ongoing research on database causality.

A rare example of an attempt at comparing different approaches on the same task can be found in Benferhat et al. (2008). The chosen task is causality ascription in a car accident report. It amounts to determining what elements in a sequence of reported facts can be related in a causal way, on the basis of some knowledge about the course of affairs. The study compares six approaches, corresponding to a large span of formal models, respectively based on structural equations (Halpern and Pearl 2005a), nonmonotonic consequence relations (Bonnefon et al. 2008), preference relations between trajectories (Dupin de Saint-Cyr 2008),[4] identification of violated norms (Kayser and Nouioua 2009), possibilistic graphical representations and interventions

[4]This proposal is based on the idea that counterfactuality involves the computation of two kinds of evolutions of the world, namely extrapolation and update. If we want to know whether an action is a counterfactual cause of an event, given a reported sequence of events, we need to (i) compute the most normal evolutions of the world (called trajectories) that correspond to the sequence. This computation is called extrapolation, it is a process of completing initial beliefs sets stemming from observations by assuming minimal abnormalities in the evolution of the world with respect to generic knowledge; (ii) compute what would have happened if some event had not been true. This is done

(Benferhat 2010), and connectionism (Thagard and Verbeurgt 1998). Interestingly enough, the compared approaches focus on different aspects of the problem by either identifying all the potential causes, or selecting a smaller subset thereof by taking advantages of contextually abnormal facts, or by modeling interventions to get rid of simple correlations. The paper also proposes a battery of criteria for judging the approaches.

4 Conclusion

Causality is not a notion which is easy to grasp in spite of its intuitive appeal. This fact explains the existence of a huge literature on the topic. The main ambition of this chapter is to introduce the different facets and issues associated with this notion and to provide a broad overview of the different proposals that have appeared in the artificial intelligence literature. It is also worth pointing out that the different ideas surveyed here can be handled in various qualitative or quantitative settings.

Acknowledgements This paper has partly benefited of the collective work, performed a decade ago, in the 3-year ANR research project MICRAC (2005-2008) dedicated to the study of causality modeling, whose participants included S. Benferhat, J.-F. Bonnefon, Ph. Chassy, R. Da Silva Neves, D. Dubois, F. Dupin de Saint-Cyr, D. Hilton, D. Kayser, F. Levy, F. Nouioua, S. Nouioua-Boutouhami and H. Prade.

References

Alechina N, Halpern JY, Logan B (2017) Causality, responsibility and blame in team plans. In: Larson K, Winikoff M, Das S, Durfee EH (eds) Proceedings of 16th conference on autonomous agents and multiagent systems (AAMAS'17), São Paulo, ACM, pp. 1091–1099

Aleksandrowicz G, Chockler H, Halpern JY, Ivrii A (2017) The computational complexity of structure-based causality. J Artif Intell Res 58:431–451

Ayachi R, Ben Amor N, Benferhat S (2014) Inference using compiled min-based possibilistic causal networks in the presence of interventions. Fuzzy Sets Syst 239:104–136

Batusov V, Soutchanski M (2018) Situation calculus semantics for actual causality. In: Proceedings of 32nd AAAI Conference on Artificial Intelligence, AAAI Press, New Orleans

Benferhat S (2010) Interventions and belief change in possibilistic graphical models. Artif Intell 174(2):177–189

Benferhat S, Bonnefon J, Chassy P, Da Silva Neves R, Dubois D, Dupin de Saint-Cyr F, Kayser D, Nouioua F, Nouioua-Boutouhami S, Prade H, Smaoui S (2008) A comparative study of six formal models of causal ascription. In: Greco S, Lukasiewicz T (eds) Scalable uncertainty management, (Proceedings SUM'08). LNCS, vol 5291. Springer, Berllin, pp 47–62

Benferhat S, Dubois D, Prade H (2009) Interventions in possibilistic logic. In: Godo L, Pugliese A (eds) Scalable Uncertainty Management (Proc. SUM'09), LNCS, vol 5785. Springer, Berlin, pp 40–54

---

by updating using a distance between trajectories that takes into account the date of the change, and normality.

Benferhat S, Smaoui S (2011) Inferring interventions in product-based possibilistic causal networks. Fuzzy Sets Syst 169(1):26–50

Bertossi LE (2018) Causality in databases: answer-set programs and integrity constraints. In: Olteanu D, Poblete B (eds) Proc. 12th Alberto Mendelzon Int. Workshop on foundations of data management, Cali, Colombia, CEUR workshop proceedings, vol 2100, CEUR-WS.org

Besnard P, Cordier M-O, Moinard Y (2008a) Deriving explanations from causal information. In: Ghallab M, Spyropoulos CD, Fakotakis N, Avouris NM (eds) Proceedings of 18th European conference on artificial intelligence (ECAI'08), IOS Press, Patras, pp 723–724

Besnard P, Cordier M-O, Moinard Y (2008b) Ontology-based inference for causal explanation. Integr Comput-Aided Eng 15(4):351–367

Björnsson G (2007) How effects depend on their causes, why causal transitivity fails, and why we care about causation. Philos Stud 133(3):349–390

Bochman A (2003) A logic for causal reasoning. In: Gottlob G, Walsh T (eds) Proceedings of 18th international joint conference on artificial intelligence (IJCAI'03), Morgan Kaufmann, Acapulco, pp 141–146

Bonnefon J, Da Silva Neves R, Dubois D, Prade H (2008) Predicting causality ascriptions from background knowledge: model and experimental validation. Int J Approx Reason 48(3):752–765

Bonnefon J, Da Silva Neves R, Dubois D, Prade H (2012) Qualitative and quantitative conditions for the transitivity of perceived causation - Theoretical and experimental results. Ann Math Artif Intell 64(2–3):311–333

Boukhris I, Benferhat S, Elouedi Z (2014) Ascribing causality from observational and interventional belief function knowledge modeling. Mult-Valued Log Soft Comput 22(4–6):459–480

Cartwright N (2007) Hunting causes and using them: approaches in philosophy and economics. Cambridge University Press, Cambridge

Chambaz A, Drouet I, Thalabard J-C (2014) Causality, a trialogue. J Causal Inference 2(2):201–241

Chassy P, de Calmès M, Prade H (2012) Making sense as a process emerging from perception-memory interaction: a model. Int J Intell Syst 27(8):757–775

Chen B, Tian J, Pearl J (2014) Testable implications of linear structural equation models. In: Brodley CE, Stone P (eds) Proceedings of 28th AAAI Conference on Artificial Intelligence Québec City, July 27–31. AAAI Press, pp. 2424–2430

Chockler H, Halpern JY (2004) Responsibility and blame: a structural-model approach. J Artif Intell Res 22:93–115

De Finetti B (1936) La logique de la probabilité. Actes du Congrès international de philosophie scientifique. Hermann et Cie, Paris, pp 1–9

de Kleer J, Brown JS (1986) Theories of causal ordering. Artif Intell 29(1):33–61

Demolombe R (2000) Action et causalité : Essais de formalisation en logique. In: Prade H, Jeansoulin R, Garbay C (eds) Le Temps, l'Espace et l'Evolutif en Sciences du Traitement de l'Information, Cépaduès, Toulouse, pp 209–223

Demolombe R (2012) Causality in the context of multiple agents. In: Ågotnes T, Broersen JM, Elgesem D (eds) Proceedings of 11th international conference on deontic logic in computer science (DEON'12), Bergen, LNCS, vol 7393. Springer, Berlin, pp 1–15

Druzdzel MJ, Simon HA (1993) Causality in Bayesian belief networks. In: Heckerman D, Mamdani EH (eds) Proceedings of 9th conference on uncertainty in artificial intelligence (UAI'93), Providence, July 9–11, Morgan Kaufmann, pp 3–11

Dubois D, Prade H (1995) Fuzzy relation equations and causal reasoning. Fuzzy Sets Syst 75(2):119–134

Dubois D, Prade H (2000) An overview of ordinal and numerical approaches to causal diagnostic problem solving. In: Gabbay DM, Kruse R (eds) Abductive reasoning and learning, Kluwer Academic Publication, pp 231–280

Dubois D, Prade H (2005) Modeling the role of (ab)normality in the ascription of causality judgements by agents. In: Morgenstern L, Pagnucco M (eds) Working notes of IJCAI-05 workshop on nonmonotonic reasonning, action and change (NTAC'05), Edinburgh, pp 22–27

Dupin de Saint-Cyr F (2008) Scenario update applied to causal reasoning. In: Brewka G, Lang J (eds) Proceedings of 11th international conference on principles of knowledge representation and reasoning (KR'08), AAAI Press, Sydney, pp 188–197
Eells E (1991) Probabilistic causality. Cambridge University Press, Cambridge
Eells E, Sober E (1983) Probabilistic causality and the question of transitivity. Philos Sci 50:35–57
Galles D, Pearl J (1997) Axioms of causal relevance. Artif Intell 97(1–2):9–43
Galles D, Pearl J (1998) An axiomatic characterization of causal counterfactuals. Found Sci 3(1):151–182
Gammerman A (ed) (1999) Causal models and intelligent data management. Springer, Berlin
Gavanski I, Wells GL (1989) Counterfactual processing of normal and exceptional events. J Exp Soc Psychol 25:314–325
Geffner H (1990) Causal theories for nonmonotonic reasoning. In: Shrobe HE, Dietterich TG, Swartout WR (eds) Proceedings of 8th national conference on artificial intelligence, AAAI/MIT Press, pp 524–530
Giordano L, Martelli A, Schwind C (1998) Dealing with concurrent actions in modal action logics. In: Prade H (ed) Proceedings of 13th European conference on artificial intelligence (ECAI'98), Brighton, Wiley, New York, pp 537–541
Giordano L, Martelli A, Schwind C (2000) Ramification and causality in a modal action logic. J Log Comput 10(5):625–662
Giunchiglia E, Lee J, Lifschitz V, McCain N, Turner H (2004) Nonmonotonic causal theories. Artif Intell 153(1–2):49–104
Glymour C, Scheines R, Spirtes P, Kelly K (1987) Discovering causal structure. Academic, New York
Goldszmidt M, Pearl J (1992) Rank-based systems: a simple approach to belief revision, belief update, and reasoning about evidence and actions. In: Nebel B, Rich C, Swartout WR (eds) Proceedings of 3rd international conference on principles of knowledge representation and reasoning (KR'92). Cambridge, MA, Oct. 25–29, Morgan Kaufmann, pp 661–672
Goldszmidt M, Pearl J (1996) Qualitative probabilities for default reasoning, belief revision, and causal modeling. Artif Intell 84:57–112
Goldvarg Y, Johnson-Laird PN (2001) Naive causality: a mental model theory of causal meaning and reasoning. Cogn Sci 25:565–610
Good IJ (1961) A causal calculus i. Br J Philos Sci 11:305–318
Good IJ (1962) A causal calculus ii. Br J Philos Sci 12:43–51
Guyon I, Janzing D, Schölkopf B (2010) Causality: objectives and assessment. JMLR Proc 6:1–38
Guyon I, Statnikov AR, Aliferis CF (2011) Time series analysis with the causality workbench. In: Popescu F, Guyon I (eds) Neural information processing systems (NIPS) mini-symposium on causality in time series, Vancouver, Dec. 10, 2009, JMLR Proceedings, vol 12, pp 115–139
Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1-2):1–12. Reprinted in D. F. Hendry and M. S. Morgan (eds.), The Foundations of Econometric Analysis, Cambridge Univ. Pr., 477-490, 1995
Hall N (2000) Causation and the price of transitivity. J Philos 97:198–222
Hall N (2007) Structural equations and causation. Philos Stud 132(1):109–136
Halpern J (2016) Sufficient conditions for causality to be transitive. Philos Sci 583:213–226
Halpern JY (2015) A modification of the Halpern-Pearl definition of causality. In: Yang Q, Wooldridge M (eds) Proceedings of 24th international joint conference on artificial intelligence (IJCAI'15), Buenos Aires, July 25–31, AAAI, pp 3022–3033
Halpern JY (2017) Actual causality. MIT Press, Cambridge
Halpern JY, Hitchcock C (2015) Graded causation and defaults. Br J Philos Sci 66 (2):413–457
Halpern JY, Pearl J (2001a) Causes and explanations: a structural-model approach - Part II: explanations. In: Nebel B (ed) Proceedings of 17th international joint conference on artificial intelligence (IJCAI'01), Morgan Kaufmann, Seattle, pp 27–34

Halpern JY, Pearl J (2001b) Causes and explanations: a structural-model approach: Part 1: Causes. In: Breese JS, Koller D (eds) UAI ' 01: Proceedings of 17th conference in uncertainty in artificial intelligence, Morgan Kaufmann, Seattle, pp 194–202
Halpern JY, Pearl J (2005a) Causes and explanations: a structural-model approach. Part I: causes. Br J Philos Sci 56 (4):843–887
Halpern JY, Pearl J (2005b) Causes and explanations: a structural-model approach. Part II: explanations. Br J Philos Sci 56 (4):889–911
Hart HLA, Honoré T (1985) Causation in the Law. Oxford University Press, Oxford
Hilpinen R (1997) On action and agency. In: Ejerhed E, Lindström S (eds) Logic, action and cognition - essays in philosophical logic, Kluwer Academic Publication, pp 3–27
Hilton DJ, Slugoski BR (1986) Knowledge-based causal attribution: the abnormal conditions focus model. Psychol Rev 93:75–88
Hitchcock C (2001) The intransitivity of causation revealed in equations and graphs. J Philos 98(6):273–299
Hitchcock C (2009) Structural equations and causation: six counterexamples. Philos Stud 144(3):391–401
Hobbs JR (2005) Toward a useful concept of causality for lexical semantics. J Semant 22(2):181–209
Huber F (2011) Lewis causation is a special case of Spohn causation. Br J Philos Sci 62:207–210
Hume D (1748) An inquiry concerning human understanding. 3rd 1777 ed., republished in: Enquiries Concerning Human Understanding and Concerning the Principles of Morals, Oxford University Press, Oxford, 1971
Iwasaki Y, Simon HA (1986a) Causality in device behavior. Artif Intell 29(1):3–32
Iwasaki Y, Simon HA (1986b) Theories of causal ordering: reply to De Kleer and Brown. Artif Intell 29(1):63–72
Kanger S (1972) Law and logic. Theoria 38:105–132
Kayser D, Nouioua F (2009) From the textual description of an accident to its causes. Artif Intell 173(12–13):1154–1193
Keil FC (2006) Explanation and understanding. Ann Rev Psychol 57:227–254
Kleiman-Weiner M, Halpern JY (2018) Towards formal definitions of blameworthiness, intention, and moral responsibility. In: Proceedings of 32nd AAAI conference on artificial intelligence (AAAI'18), New Orleans
Knobe J, Fraser B (2008) Causal judgement and moral judgement: two experiments. In Sinnott-Armstrong W (ed) Moral psychology, vol 2: the cognitive science of morality, MIT Press, pp 441–447
Kraus S, Lehmann D, Magidor M (1990) Nonmonotonic reasoning, preferential models and cumulative logics. Artif Intell 44:167–207
Kyburg Jr HE (2005) Book review – Judea Pearl, causality, Cambridge University Press, Cambridge, 2000. Artif Intell, 169:174–179
Lagnado DA, Sloman SA (2005) Do we "do"? Cogn Sci 29:5–39
Lehmann D, Magidor M (1992) What does a conditional knowledge base entail? Artif Intell 55:1–60
Lewis D (1973) Causation. J Philos 70:556–567
Lewis D (1976) Probabilities of conditionals and conditional probabilities. Philosl Rev 85:297–315
Lewis D (1986) Philosophical papers, volume II. Oxford University Press, Oxford. Contains: 'Causation', with 6 postscripts to the original 1973 paper, pp. 159–213; 'causal explanation', pp. 214–240; 'Events', pp.241–270'
Lewis D (2000) Causation as influence. J Philos 97(4):182–197
Lin F (1995) Embracing causality in specifying the indirect effects of actions. In: Proceedings of 14th international joint conference on artificial intelligence (IJCAI'95), Morgan Kaufmann, Montréal, pp 1985–1993
Lopez-Paz D, Nishihara R, Chintala S, Schölkopf B, Bottou L (2017) Discovering causal signals in images. In: Proceedings of 2017 IEEE conference on computer vision and pattern recognition, (CVPR'17), IEEE Computer Society, Honolulu, pp 58–66
Mackie JL (1974) The cement of the universe: a study of causation. Oxford University Press, Oxford

McCain N, Turner H (1995) A causal theory of ramifications and qualifications. In: Proceedings of 14th international joint conference on artificial intelligence (IJCAI'95). Morgan Kaufmann, Montréal, pp 1978–1984
McEleney A, Byrne RMJ (2006) Spontaneous counterfactual thoughts and causal explanations. Think Reason 12(2):235–255
Meliou A, Gatterbauer W, Halpern JY, Koch C, Moore KF, Suciu D (2010) Causality in databases. IEEE Data Eng Bull 33(3):59–67
Meliou A, Roy S, Suciu D (2014) Causality and explanations in databases. PVLDB 7(13):1715–1716
Mumford S, Anjum RL (2013) Causation. A Very Short Introduction. Oxford University Press, Oxford
Novick LR, Cheng PW (2004) Assessing interactive causal influence. Psychol Rev 111(2):455–485
Over DE, Hadjichristidis C, Evans JSBT, Handley SJ, Sloman SA (2007) The probability of causal conditionals. Cogn Psychol 54:62–97
Paul LA, Hall N (2013) Causation. A User's Guide, Oxford University Press, Oxford
Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publication
Pearl J (1994) A probabilistic calculus of actions. In: de Mántaras RL, Poole D (eds) Proceedings of 10th annual conference on uncertainty in artificial intelligence (UAI'94), Seattle, July 29–31, Morgan Kaufmann, pp 454–462
Pearl J (2000) Causality, 2nd revised edn. 2009 Cambridge University Press, Cambridge
Pearl J (2013) Linear models: a useful "microscope" for causal analysis. J Causal Inference 1(1):155–170
Pearl J (2015) Trygve Haavelmo and the emergence of causal calculus. Econ Theory 31(1):152–179
Pearl J, Mackenzie D (2018) The book of why. The new science of cause and effect. Basic Books, New York
Pearl J, Glymour M, Jewell NP (2016) Causal inference in statistics: A Primer, Wiley, New York
Peng Y, Reggia JA (1990) Abductive inference models for diagnostic problem-solving. Springer, Berlin
Pörn I (1977) Action Theory and Social Science. D, Reidel, Synthese Library, Some Formal Models, p 120
Prade H (2008) Responsibility judgments: Steps towards a formalization. In: Magdelena L, Ojeda-Aciego M, Verdegay J-L (eds) Proceedings of 12th international conference on information processing and management of uncertainty in knowledge-based systems (IPMU'08), Málaga, pp 145–152
Reggia JA, Nau DS, Wang PY (1985a) A formal model of diagnosis inference i. problem formulation and decomposition. Inf Sci 37:227–256
Reggia JA, Nau DS, Wang PY (1985b) A formal model of diagnosis inference ii. algorithmic solution and application. Inf Sci 37:257–285
Reiter R (1987) A theory of diagnosis from first principles. Artif Intell 32:57–95
Salmon WC (1984) Scientific explanation and the causal structure of the world. Princeton University Press, Princeton
Sanchez E (1977) Solutions in composite fuzzy relation equations: Application to medical diagnosis in Brouwerian logic. In: Gupta MM, Saridis GN, Gaines BR (eds) Fuzzy automata and decision processes, North-Holland, pp 221–234
Shafer G (1996) The art of causal conjecture. MIT Press, New York
Shafer G (1998) Causal logic. In: Prade H (ed) Proceedings of 13th European conference on artificial intelligence (ECAI'98), Brighton, Wiley, New York, pp 711–720
Shafer G (1999) Causal conjecture. In: Gammerman A (ed) Causal models and intelligent data management. Springer, Berlin, pp 17–32
Shafer G (2000) Causality and responsibility. Cardozo Law Rev 22:101–123
Shafer G, Gillett PR, Scherl RB (2000) The logic of events. Ann Math Artif Intell 28(1–4):315–389
Shoham Y (1990) Nonmonotonic reasoning and causation. Cogn Sci 14(2):213–252

Shoham Y (1991) Remarks on simon's comments. Cogn Sci 15(2):301–303
Simon H (1952) On the definition of causal relation. J Philos 49:517–528
Simon H (1953) Causal ordering and identifiability. In: Hood WC, Koopmans TC (eds) Studies in econometric methods. Wiley, New York, pp 49–74
Simon H (1954) Spurious correlation: a causal interpretation. J Am Stat Assoc 49:467–479
Simon H (1991) Nonmonotonic reasoning and causation: comment. Cogn Sci 15(2):293–300
Simon H, Rescher N (1966) Cause and counterfactual. Philos Sci 33(4):323–340
Spellman BA, Mandel DR (1999) When possibility informs reality. counterfactual thinking as a cue to causality. Curr Dir Psychol Sci 8 (4):120–123
Spinoza B (1677) Ethics. Everyman paperbacks. Republished in 1992
Spirtes P, Glymour C, Schienes R (1993) Causation, prediction and search. Springer, Berlin
Spohn W (2000) Bayesian nets are all there is to causal dependence. In: Costantini D (ed) Stochastic dependence and causality. CSLI Publication, Stanford, pp 157–172
Spohn W (2006) Causation: an alternative. Br J Philos Sci 57:93–119
Spohn W (2012) The laws of belief: ranking theory and its philosophical applications. Oxford University Press, Oxford
Stalnaker RC (1968) A theory of conditionals. In: Rescher N (ed) Studies in logical theory, Blackwell, pp 98–112
Stein LA, Morgenstern L (1994) Motivated action theory: a formal theory of causal reasoning. Artif Intell 71(1):1–42
Suppes P (1970) A probabilistic theory of causality. North-Holland Publication Company
Thagard P (1989) Explanatory coherence. Behav Brain Sci 12:435–467
Thagard P (2000) Probabilistic networks and explanatory coherence. Cogn Sci Q 1:91–114
Thagard P, Verbeurgt K (1998) Coherence as constraint satisfaction. Cogn Sci 22:1–24
Thielscher M (1997) Ramification and causality. Artif Intell 89(1–2):317–364
Turner H (1999) A logic of universal causation. Artif Intell 113(1–2):87–123
von Wright GH (1963) Norm and Action. Routledge and Keagan
von Wright GH (1971) Explanation and understanding. Cornell University Press
White G (2002) A modal formulation of McCain-Turner's theory of causal reasoning. In: Flesca S, Greco S, Leone N, Ianni G (eds) Logics in artificial intelligence (Proceedings JELIA'02), LNCS, vol 2424. Springer, Berlin, pp 211–222
Woodward J (2003) Making things happen: a theory of causal explanation. Oxford University Press, Oxford
Wright S (1921) Correlation and causation. J Agric Res 20:557–585
Zadeh LA (2002) Causality is undefinable. toward a theory of hierarchical definability. In: Meech JA, Veiga MM, Kawazoe Y, LeClair SR (eds) Intelligence in a materials world: selected papers from IPMM-2001, CRC, Boca Raton, pp 29–34

Case-Based Reasoning, Analogy, and Interpolation

Béatrice Fuchs, Jean Lieber, Laurent Miclet, Alain Mille, Amedeo Napoli, Henri Prade and Gilles Richard

Abstract This chapter presents several types of reasoning based on analogy and similarity. Case-based reasoning, presented in Sect. 2, consists in searching a case (where a case represents a problem-solving episode) similar to the problem to be solved and to adapt it to solve this problem. Section 3 is devoted to analogical reasoning and to recent developments based on analogical proportion. Interpolative reasoning, presented in Sect. 4 in the formal setting of fuzzy set representations, is another form of similarity-based reasoning.

1 Introduction

Charles S. Peirce (1839–1914) distinguished three main forms of logical inference, namely deduction, abduction and induction, in relation with scientific inquiry

B. Fuchs (✉)
Université de Lyon, IAE-Université Lyon 3, CNRS, LIRIS, 69372 Lyon Cedex 08, France
e-mail: beatrice.fuchs@liris.cnrs.fr

J. Lieber · A. Napoli
Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
e-mail: jean.lieber@loria.fr

A. Napoli
e-mail: amedeo.napoli@loria.fr

L. Miclet
IRISA, 22300 Lannion, France
e-mail: laurent.miclet@gmail.com

A. Mille
Université Lyon 1, CNRS, LIRIS UMR 5205, 69622 Villeurbanne, France
e-mail: alain.mille@liris.cnrs.fr

H. Prade
IRIT, CNRS and Université Paul Sabatier, Toulouse, France
e-mail: prade@irit.fr

G. Richard
IRIT, CNRS and Université, 31062 Toulouse Cedex 9, France
e-mail: richard@irit.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_10

(see Peirce 1955). Each of these three inference forms involves generic knowledge in their patterns, in a way or another. There exist other modes of reasoning that only deal with factual information and that are still useful for producing plausible conclusions, although they may turn to be false. These latter modes are based on the idea of comparing cases and the notion of similarity. This chapter covers two important forms of such inference: case-based reasoning and analogical reasoning. The chapter also includes another form of similarity-based reasoning that provides interpolation capabilities. It is based on fuzzy rules, where a fuzzy set may be viewed as a particular value associated with the values that are more or less close to this value.

The paper is organized into three main sections that are respectively devoted to case-based reasoning, analogical reasoning, and interpolative reasoning.

2 Case-Based Reasoning

Case-based reasoning (CBR) relies on experience in the form of problem-solving episodes (or cases) in order to solve new problems (Riesbeck and Schank 1989). It can be differentiated from other approaches of problem-solving in artificial intelligence (AI) which mainly exploit general domain knowledge to generate solutions. By contrast, a CBR system is mainly based on concrete chunks of experience, with specific contexts. Such chunks are represented by *source cases* stored in a *case base*. When a new problem—the *target problem*—is given as input to a CBR system, this latter searches in the case base a source case (or, sometimes, several source cases) similar to the target problem that is reused in order to solve it thanks to an *adaptation* process. The new chunk of experience (the target problem together with its solution), once validated, can be stored in the case base and the system knowledge can gain problem-solving competence this way.

CBR is based on the idea that for solving a problem, the problem-solving experience is often useful, when a "direct" solution is not easily found. For example, if someone wants a pear pie recipe, has not the experience of such a recipe, but has the similar experience of an apple pie recipe, he/she can adapt this latter to cook a pear pie. The underlying principle relates to the *analogical proportion* "A is to B as C is to D". In the framework of CBR, A and C are problems and B and D are solutions. Figure 1 illustrates this idea. An overview of works on analogical reasoning which is concomitant with the emergence of CBR is given in (Hall 1989). Analogical reasoning in itself, independently from CBR, is presented in Sect. 3.

The origins of CBR can be found in works of M. Minsky and R. Schank. In the work about perception of M. Minsky, a knowledge representation formalism able to explain to some extent the efficiency of human mental activities has been defined (Minsky 1975). This formalism is based on structures called frames that can be dynamically reused and that represent models of situations. The matching of frames can be used to recognize situations. The studies of R. Schank on natural language understanding (Schank 1982) argued that cognitive processes of understanding and learning are linked with the way the human experience is organized. In his theory, meaning is captured thanks to semantic constructs that are indepen-

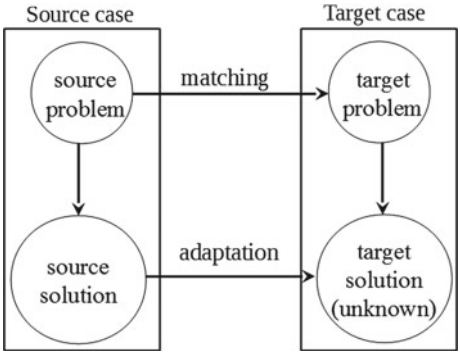


Fig. 1 The analogical proportion applied to CBR

dent from syntax and are represented by sequences that are used to predict how future sequences can be extended. Then, he designed the model of scripts for an improved description of episodes by a set of actions structured by relations. This model has then evolved towards the model of dynamic memory, able to reorganize itself dynamically as new episodes are learned, generating generalized episodes that factorize the common features of actual specific episodes (actual in the sense that they are representation of actual facts). In (Riesbeck and Schank 1989), the episodes are described with the help of memory organization packets (MOPs) and the understanding of a situation depends on the way MOPs are related in the memory. Later, Janet Kolodner has implemented one of the first CBR systems based on the model of dynamic memory (Kolodner 1993).

2.1 Basic Notions Related to CBR

In a given application domain, the notions of *problem* and of *solution* are given. `Problems` denotes the *problem space* and `Solutions`, the *solution space*: a problem is by definition an element of `Problems`, a solution is by definition an element of `Solutions`. The existence of a binary relation on `Problems` $\times$ `Solutions` that is read "has for solution" is assumed though the complete knowledge of this relation is usually not known. *Solving* a problem `pb` amounts to find (or build) `sol` $\in$ `Solutions` such that `pb` *has for solution* `sol`. Since the problem-solution relation is usually not completely known, `sol` is, for most CBR systems, only a solution hypothesis.

CBR systems can be categorized according to the type of problems they aim at solving. For example, if a problem is given by an initial stage `init` and a goal to reach `goal`, and if a solution is a path in the search space from `init` to a state satisfying `goal`, this is a planning problem (see chapter "Planning in Artificial Intelligence" of Volume 2) and the use of CBR to tackle such a problem is called case-based planning (see Sect. 2.4). A decision problem is described by a situation for which a decision is required. Other types of problems can be distinguished like configuration diagnosis, or scheduling problems (Riesbeck and Schank 1989; Stefik 1995).

A *case* is the representation of a problem-solving episode. Let pb ∈ Problems. In general, a case is given by an ordered pair (pb, sol(pb)) where pb ∈ Problems, sol(pb) ∈ Solutions and pb has for solution sol(pb). Often, pieces of information useful to its reuse are associated with a case. In particular, the available information on the links between pb and sol(pb) is called *dependency*.

A *case base*, denoted CaseBase in the following, is a finite set of cases. A *source case* (srce, sol(srce)) is an element of CaseBase and srce is called a *source problem*. The target problem, denoted by tgt, is the problem to be solved.

2.1.1 The Process Model of CBR

CBR is usually modeled by a "cycle" that specifies the sequence of its steps. This cycle contains four general steps having profit of a knowledge base including a case base (Aamodt and Plaza 1994). This cycle has been enriched by an *elaboration* step, which gives the cycle presented in Fig. 2.

During the elaboration step, the query expressed by the user is transformed into a problem understandable by the system, and the target problem tgt is generated. During the retrieval step, a case (srce, sol(srce)) similar to the target problem tgt is searched in the case base. Then this case is modified during the *adaptation* step (also known as reuse step). The solution sol(tgt) can be validated (e.g., by experts) and, if validated or corrected, the newly formed case (tgt, sol(tgt)) can be stored in the case base (*validation* and *case storage* steps).

This process model has variants. One of them is the possibility to retrieve and adapt (or combine) several source cases similar to the target problem.

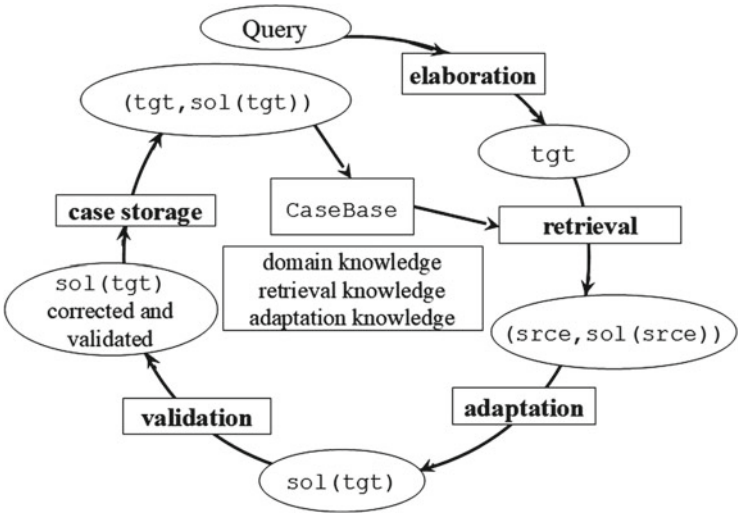


Fig. 2 The CBR cycle

2.1.2 The Knowledge Model of CBR

A CBR system is based on several *knowledge containers* (see (Richter 1998; Richter and Weber 2013)). One of them is the case base. Another one constitutes the domain knowledge (or domain ontology), that contains the vocabulary used to express the cases and also expresses sufficient conditions for a problem, a solution or a case to be licit (for the notion of ontology, see chapters "Reasoning with Ontologies" and "Knowledge Engineering" of this volume). The third one is the retrieval knowledge or similarity, that enables to prefer a source case to another, given the target problem. Similarity is often implemented thanks to a similarity measure. Finally, the adaptation knowledge is used by adaptation. It is often represented by adaptation rules.

An important feature of CBR is that these knowledge containers are complementary, in the sense that the "weakness" of one of them can be compensated by "strength" of the other ones. For example, if the case base is large, then little adaptation knowledge is necessary. Conversely, with a lot of adaptation knowledge, fewer cases are needed.

The next section describes with more details the different steps of CBR with their use of the knowledge containers.

2.2 The CBR Steps

2.2.1 Elaboration

A CBR system is triggered by a query given by the user, that is treated by the elaboration step. Elaboration prepares case retrieval by enriching the problem description in order to obtain a target problem. This preliminary step points out in particular the problem features that may have an impact on the solution. These features can be inferred from domain knowledge in order to ease the problem-solving, in particular the retrieval and adaptation steps.

2.2.2 Retrieval

Retrieval consists in searching in the case base a case $(\texttt{srce}, \texttt{sol}(\texttt{srce}))$ whose reuse is useful to solve the target problem:

$$\texttt{retrieval}:(\texttt{CaseBase}, \texttt{tgt}) \mapsto (\texttt{srce}, \texttt{sol}(\texttt{srce})) \in \texttt{CaseBase}$$

It is based on the knowledge of the similarity between problems, according to the following principle: similar problems have (or may have) similar solutions.

Similarity Measure

A frequent way to represent similarity is to use a similarity measure $\mathscr{S}$: Problems × Problems → [0; 1] in which the features are weighted according to their estimated importance in the problem solving. This way, it can be expressed

- That two problems src e and tgt are similar: $\mathscr{S}(\texttt{src}, \texttt{tgt}) \geqslant \mathscr{S}_{\min}$, where $\mathscr{S}_{\min}$ is a predefined similarity threshold;
- That, given the target problem tgt, retrieval prefers a case $(\texttt{src}e_1, \texttt{sol}(\texttt{src}e_1))$ to a case $(\texttt{src}e_2, \texttt{sol}(\texttt{src}e_2))$: $\mathscr{S}(\texttt{src}e_1, \texttt{tgt}) > \mathscr{S}(\texttt{src}e_2, \texttt{tgt})$.

Sometimes, a measure of dissimilarity (e.g., a distance function) d : Problems × Problems ⟶ $[0; +\infty[$ is used instead of a similarity measure, knowing that $\mathscr{S}$ must be maximized when d must be minimized. A classical way to associate a dissimilarity measure d to a similarity measure $\mathscr{S}$ (and conversely) consists in writing $\mathscr{S}(\texttt{src}, \texttt{tgt}) = 1/(1 + d(\texttt{src}, \texttt{tgt}))$.

A frequent class of similarity measures is defined as follows. First, the features of src e and tgt are matched (e.g., if the case representation is a simple attribute-value representation, two features with the same attribute are matched). Then, a local similarity measure is computed between each of the matched descriptors. Then, the global similarity measure $\mathscr{S}(\texttt{src}, \texttt{tgt})$ is computed by an aggregation of the values of the local similarity measures, using weights according to the feature importance. One way to choose these weights is to use a machine learning technique: the best set of weights is the one that best fits a training set of preference relations.

In the approach developed by Hüllermeier (2007), gradual similarity relations are used. They are inspired from approximate reasoning based on fuzzy rules (cf. Sect. 4.1).

Classification and Indexing

In many CBR system, retrieval has profit of a structure on the case base. The idea is to organize the case base in classes along several features. In particular, the use of an *index* hierarchy is frequent, an index of a source case being considered as a kind of summary of this case (sometimes expressed in a less expressive formalism (Koehler 1996)). This hierarchy gathers cases having common features in a class. Let idx(tgt) be the index associated to the target problem and idx(src e) be the index associated to each (src e, sol(src e)) ∈ CaseBase. Then, the source cases whose indexes are the closest ones to idx(tgt) in the hierarchy (according to some distance function between nodes of a graph) are the first candidates (e.g., if idx(src e) shares with idx(tgt) a direct superclass, src e and tgt are considered to be close).

In Resyn/CBR, an application of CBR to synthesis in organic chemistry, the index idx(src e) of (src e, sol(src e)) is a generalization of src e and retrieval is performed by a classification process (Lieber and Napoli 1996). Retrieval returns a source case (src e, sol(src e)) associated with a *similarity path* S(src e, tgt) that ensures the adaptability of the source case to solve the target problem. A similarity path is a sequence of relations from src e to $\texttt{C}_0 = \texttt{tgt}$, with the index $\texttt{I}_0 = \texttt{idx}(\texttt{src}e)$ as intermediate of the hierarchy that generalizes the source case:

$$\texttt{srce} \sqsubseteq \texttt{I}_0 \xrightarrow{\ell_1} \texttt{I}_1 \xrightarrow{\ell_2} \ldots \xrightarrow{\ell_p} \texttt{I}_p \sqsupseteq \texttt{C}_q \xleftarrow{r_q} \cdots \xleftarrow{r_2} \texttt{C}_1 \xleftarrow{r_1} \texttt{C}_0 = \texttt{tgt}$$

Building a similarity path from srce to tgt is a matching process. A cost is associated to any similarity path. It is used to choose the source case for which a similarity path with the lowest cost can be built. Each relation r of a similarity path ($\texttt{r} \in \{\sqsubseteq, \xrightarrow{\ell_1}, \xrightarrow{\ell_2}, \ldots, \xrightarrow{\ell_p}, \sqsupseteq, \xleftarrow{r_q}, \ldots \xleftarrow{r_2}, \xleftarrow{r_1}\}$ where the ℓ_i's and the r_j's are transformation rules) is associated to an adaptation function $\mathscr{A}_{\texttt{r}}$: the pair $(\texttt{r}, \mathscr{A}_{\texttt{r}})$ constitutes an adaptation rule (also called reformulation in (Melis et al. 1998)). For example, the relation $\sqsubseteq$ ("is more specific than") is associated to a solution generalization function $\mathscr{A}_{\sqsubseteq}$ and the relation $\sqsupseteq$ ("is more general than") is associated to a solution specialization function $\mathscr{A}_{\sqsupseteq}$. Each of these relations are exploited in the adaptation step and retrieval ensures the adaptability of the retrieved source case. For this reason, this approach belongs to the family of adaptation-guided approaches to retrieval (Smyth and Keane 1996).

In (Koehler 1996), a case-based planner is described in which the plans are described in a complex temporal logic but retrieval is done in a tractable description logic: cases are indexed in this more abstract and more tractable formalism and the source cases whose index are the closest ones to the index of the target problem are retrieved.

2.2.3 Adaptation

After retrieval, the solution of the source case is proposed as a solution to the target problem. Usually, this solution needs to be adapted in order to take into account differences between source and target problems. The objective of adaptation is to solve tgt on the basis of the retrieved case (srce, sol(srce)):

$$\texttt{adaptation:}((\texttt{srce}, \texttt{sol(srce)}), \texttt{tgt}) \mapsto \texttt{sol(tgt)}$$

Note that only the adaptation of a single case is considered in this section.

Adaptation is essential when the solution of the source problem cannot be reused as such for solving the target problem. It consists in modifying the source case using domain knowledge and adaptation knowledge, taking into account differences between the source and target problems (which are frequently highlighted during retrieval).

Adaptation can be considered as an analogical problem solving that can be read in two different ways: "sol(tgt) is to sol(srce) as tgt is to srce" and "sol(tgt) is to tgt as sol(srce) is to srce". These two ways correspond to two general approaches to adaptation[1]:

[1] It is noteworthy that this differs from analogical proportions (presented in Sect. 3) for which these two ways to read the four terms of an analogy are equivalent, according to the "exchange of the means" property.

- Transformational adaptation (Carbonell 1983) consists in modifying directly the source solution. It aims at modifying either the values of some solution features (this is called adaptation by substitution) or complex parts of the solution (this is called structural adaptation);
- Derivational adaptation (Carbonell 1986) consists in building entirely the solution of the target problem by applying the method that was used to generate the source solution (which often requires a modification of this method to take into account specificities of the target problem).

This can be read on the schema of Fig. 1. Indeed, when the horizontal relations are considered (i.e., between problems and between solutions), this corresponds to transformational adaptation. The principle of adaptation is then to find the variations between solution features from variations between problem features. When vertical relations are considered (i.e., from a problem to a solution), this corresponds to derivational adaptation.

Transformational Adaptation

First, the solution of the source case is copied in order to constitute a first solution of the target problem. This "first draft" is then modified according to the differences between the source and target problems pointed out by the matching process.

The approaches to adaptation vary according to the types of operations. The adaptation by substitution simply replaces elements of the solution by other elements, while structural adaptation modifies with more depth the structure of the solution by deleting and adding elements.

In the case-based planner CHEF (Hammond 1986) dedicated to cooking recipes, the adaptation by substitution modifies some ingredients in order to satisfy constraints of the target problem. CHEF also makes structural modifications on the steps of the recipe. The system Déjà Vu (Smyth and Keane 1995) uses adaptation strategies and adaptation specialists. An adaptation specialist uses transformation operations to perform local adaptations, whereas adaptation strategies solve the conflicts between adaptation specialists. Model-based adaptation (such as the CASEY system (Koton 1988)) exploits transformations that are controlled by a causal reasoning.

Derivational Adaptation

Derivational adaptation wholly regenerates the solution of the target problem by *replaying* the reasoning having led to the solution of the source case (when an operator cannot be applied, some local search is generated). Its application usually requires that a strong domain knowledge is available (ideally, a complete domain knowledge in the sense that the relation "has for solution" between problems and solutions is completely known to the system).

Some Unifying Approaches to Adaptation

From the development of ad hoc approaches of adaptation, some general principles and approaches have been pointed out, proposing general models of adaptation.

In (Fuchs and Mille 1999), a general model of tasks has been introduced to characterize the operations realized in the framework of formal models of adaptations. Adaptation consists in choosing a difference, in applying the corresponding modification and then in checking the consistency of the result. A modification can be a substitution, a deletion or an addition of elements. A substitution or an addition requires the search of an adequate element and this is done thanks to the domain knowledge.

In (Fuchs et al. 2000), the authors propose an approach to adaptation based on the notion of influence of the problem descriptors to the solution descriptors which, combined with the matchings performed during retrieval, makes possible to highlight differences of solution descriptors that can be applied to the source solution in order to obtain a target solution. This approach makes a strong connection between retrieval knowledge (based on problem differences) and adaptation knowledge (based on solution differences). It has been applied to numerical problems in the so-called differential adaptation approach (see (Fuchs et al. 2014)).

Adaptation and Belief Revision

The issue of adaptation and the issue of belief revision (see chapter "Belief Revision, Belief Merging and Information Fusion" of this volume) are both based on the notion of modification and change, hence the idea to exploit a revision operator for performing adaptation.

An agent having beliefs ψ on a static world can be confronted to new beliefs μ in conflict with ψ: $\psi \wedge \mu$ is inconsistent ($\wedge$ being the operator of conjunction of belief bases in the considered formalism). If μ are assumed to have priority over ψ, then the problem of incorporating μ to ψ is the one of the *revision* of ψ by μ. The result $\psi \dot{+} \mu$ depends on the revision operator $\dot{+}$. In (Alchourrón et al. 1985) are defined postulates that $\dot{+}$ must (or should) satisfy, in particular, predicates expressing that $\psi \dot{+} \mu$ has to be computed with a minimal change of ψ into ψ' such that $\psi' \wedge \mu$ is consistent. In (Katsuno and Mendelzon 1991), revision has been studied in a propositional framework and it has been studied more recently is other formalisms, such as the qualitative algebras (for these algebras, see chapter "Qualitative Reasoning about Time and Space" of this volume).

Revision-based adaptation can be defined as follows. Let $\mathscr{L}$ be a formalism in which can be expressed the domain knowledge DK, the source case to be adapted Source (i.e., the problem srce and its solution sol(srce)) and the target case Target (i.e., Target is given by the target problem tgt, the solution being initially unknown). Let $\dot{+}$ be a revision operator $\dot{+}$ on $\mathscr{L}$. $\dot{+}$-adaptation consists in modifying minimally the source case (this minimality being the one of the chosen revision operator $\dot{+}$) in order to make it consistent with the target case, keeping in mind the fact that these cases have to be considered with the integrity constraints given by the domain knowledge:

$$(\mathtt{DK} \wedge \mathtt{Source}) \dot{+} (\mathtt{DK} \wedge \mathtt{Target})$$

This general approach to adaptation constitutes a general framework including different approaches to adaptation including the adaptation by similarity paths. The idea is that the adaptation knowledge AK associated with this type of adaptation enables to define an operator $\dot{+}_{\mathtt{AK}}$. Therefore, the $\dot{+}_{\mathtt{AK}}$-adaptation adapts cases using both adaptation knowledge and domain knowledge.

Revision-based adaptation has been studied in propositional logic (Lieber 2007) then in a larger framework (Cojan and Lieber 2008). A similar adaptation has also been studied in the framework of the expressive description logic $\mathscr{ALC}$ (Cojan and Lieber 2011) and in the tractable description logic $\mathscr{EL}_{\bot}$ (Chang et al. 2014) (for description logics, see chapter "Reasoning with Ontologies" of this volume).

2.2.4 Validation and Case Storage

Once the target problem solved, the new case (tgt, sol(tgt)) has to be tested and evaluated. This evaluation is generally done by a human, in particular when the CBR system has incomplete problem-solving knowledge, and aims at answering the question "Is sol(tgt) a correct solution of tgt?" If the result of this evaluation is positive, then the new case can be stored in the case base. Else, the solution sol(tgt) has to be repaired and an explanation of this failure may be pointed out to avoid such a failure in the future. This is the role of the validation step (sometimes called revision) to question the system knowledge that has led to this failure, hence its relation with knowledge acquisition issues, presented in the next section.

2.3 Knowledge Acquisition for a CBR System

In order to implement a CBR system (or any knowledge-based system, denoted by KBS in the following (Stefik 1995)), its knowledge base has to be acquired and to evolve over time. In this section, "knowledge acquisition" (KA) is used as a general term for getting knowledge: from experts, from a machine learning process, or from both, and constitutes a field of knowledge engineering (see chapter "Knowledge Engineering" of this volume). A CBR system knowledge base consists of four containers, the issue of KA for such a system can be described by four interrelated issues.

Case Base KA

The case acquisition, or case authoring, consists mainly in the representation of informal problem-solving episodes. A classical way to do it consists in interviewing an expert about the way he/she solved a problem in the past and then in formalizing it. Sometimes, there are many available data that are stored informally on machines, but requires to be automatically transformed into actual cases, handable by a CBR process. For example, if problem-solving episodes are available in a textual form, natural language processing techniques can be used to interpret them into a formal representation (Dufour-Lussier et al. 2014).

Acquisition of the Domain Knowledge (or Domain Ontology)

The issue of KA of ontologies has been studied a lot in the KA community (see chapter "Knowledge Engineering" of this volume) and CBR systems benefit from it. The specificity of the acquisition of domain ontology is its close links with the other containers, as detailed hereafter (actually, this can be argued for each pairs of knowledge containers).

The case acquisition involves the need to define a vocabulary for representing cases. This vocabulary constitutes an important part of the domain knowledge, or domain ontology.

As mentioned above, the adaptation process uses both adaptation knowledge and domain knowledge (see, e.g., revision-based adaptation). In particular, it is frequent to substitute a class with another one that is close (e.g., an apple by a pear in a recipe), this closeness being often related to the ontology (e.g., apples and pears are both fruits).

In a similar way, the retrieval process often uses an ontology (e.g., to compare two values of the same attribute).

Acquisition of Similarity (Retrieval Knowledge)

Retrieval knowledge is often represented thanks to a similarity measure, acquisition of this case container frequently amounts to the acquisition of such a measure, based on known preferences between cases, given target problems. In (Stahl 2005), a learning of similarity measure procedure is defined for this purpose.

Adaptation Knowledge Acquisition

A knowledge-light approach uses mainly the case base for generating adaptation knowledge (Wilke et al. 1996).

In (Hanney 1996), the case base is exploited to generate inductively adaptation rules in the condition-action form. The training set is given by pairs of cases from the case base: such a case pair $((\texttt{srce}_1, \texttt{sol}(\texttt{srce}_1)), (\texttt{srce}_2, \texttt{sol}(\texttt{srce}_2)))$ is read as an adaptation $\texttt{adaptation}((\texttt{srce}_1, \texttt{sol}(\texttt{srce}_1)), \texttt{srce}_2) = \texttt{sol}(\texttt{srce}_2)$. The conditions express differences between problems that are related to differences between solutions. In (Craw et al. 2006), the same principle has been applied using decision tree induction algorithms. In (McSherry 1999), adaptation is performed by searching in the case base case pairs whose differences are similar to the differences between the retrieved case and the target problem. The adaptation consists in applying this difference between solutions in order to obtain a solution to $\texttt{tgt}$. In (d'Aquin et al. 2007), a knowledge discovery process using an algorithm of closed frequent itemset extraction (see chapter "Formal Concept Analysis: From Knowledge Discovery to Knowledge Processing" of Volume 2) is used in order to acquire adaptation knowledge on all the pairs of source cases. The adaptation rules are based on the differences between cases, represented by descriptors labelled with the type of variations (constant, added or removed) from the source to the target.

The approaches presented above are off-line, but, as can be seen in the following, some on-line approaches have been developed that exploit the steps of the CBR cycle to extract new pieces of knowledge.

Opportunistic Knowledge Acquisition for CBR

This form of knowledge acquisition consists in having profit of failures during the building of a solution. The approach relies on interactions between the domain expert and the system in order to acquire missing information that would have prevented the failure. It is an online approach that takes place during the validation step and is only triggered in case of failure, when the output of the adaptation process is not a valid solution of the target problem, hence the adjective "opportunistic".

The system CHEF was probably the first system to apply an opportunistic knowledge acquisition process from failures (Hammond 1990). DIAL was another early system using this principle (Leake et al. 1996). In (Hall 1986), a previous work on learning by failure, outside CBR, was presented.

The FIKA (Failure-driven Interactive Knowledge Acquisition) approach defines general principles for interactive and opportunistic knowledge acquisition in CBR that has been applied to the systems FRAKAS and IAKA. The FRAKAS system (Cordier et al. 2007) is a decision support system that exploits failures of revision-based adaptations in order to highlight gaps in the domain knowledge of the system (with respect to the expert knowledge). The knowledge acquisition process is triggered during which the interactive analysis of the failure leads to new units of knowledge that are in accordance with the expert knowledge. In IAKA, these principes have been applied to adaptation knowledge acquisition (Cordier et al. 2008). The goal is to exploit the corrections performed by the expert on the solution during the validation phase in order to trigger an interactive knowledge acquisition process. This process consists in identifying and correcting the adaptation knowledge at the origin of the failure.

2.4 Some CBR Systems

This section describes some CBR systems for the purpose of illustration. First, some generic tools useful for CBR are presented. Then, several application-dependent CBR systems are presented according to the categories they belong to.

Some Generic Tools for CBR

The system jColibri is a logical framework for developing CBR systems (Recio-Garcia 2008). In order to build a CBR application in jColibri, a task model have to be configurated and the methods associated to each task have to be implemented. This system uses an ontology of tasks and methods that defines an extendable base of the framework design. For a particular application, it is sufficient to instanciate this base and to determine the necessary extensions.

MyCBR (Stahl and Roth-Berghofer 2008) is another tool for building CBR systems that is focused on various way of modeling similarity.

Tuuurbine (Gaillard et al. 2014) is a tool for case retrieval when cases and domain knowledge are represented within the semantic web standard RDFS: the target problem is translated into one or several SPARQL queries (a SPARQL query can be used

to query an RDFS base) whose execution returns an exact match (for semantic web, see chapter "Semantic Web" of Volume 3). If no exact match is found, then the query is modified minimally in new queries for which an exact match is found.

Revisor (`revisor.loria.fr`) is a tool for revision-based adaptation in various formalisms (propositional logic, linear constraints and qualitative algebras).

Case-Based Planning

A CBR system solving planning problems (usually given by an initial state, a goal statement and a set of operators on states) and thus, building plans, is a case-based planning system. A case-based planner relying only on the search in the state space does what is sometimes called planning from first principles or planning from scratch. By contrast, some authors qualify the action of a case-based planner as planning from second principles (Koehler 1996).

The system CHEF, already mentioned above, is such a system: for CHEF, a recipe is represented by a preparation plan (Hammond 1986).

Prodigy/Analogy is a case-based planner working on a classical representation of operators (condition, del-list, add-list) working with the assumption of completeness of the problem-solving relation (the system can check whether a plan `sol` is a correct solution of a planning problem `pb`, without help from a human) (Veloso 1994). This planner is based on derivational adaptation (Carbonell 1986), on retrieval/adaptation of multiple cases, on the use of a planner from first principles for replaying the retrieved plans, and on the notion of footprint. The footprint of the initial state e_0 of a plan P is an abstraction of e_0 obtained by removing pieces of information that are not necessary for the execution of P. This notion of footprint has been reused, in particular, for the indexing process of the Resyn/CBR system mentioned above.

Many case-based planning approaches have been developed in the CBR community using different principles. Let us mention the use of plan abstraction for case-based planning (Bergmann and Wilke 1995): plans are described at several levels of abstraction, and this approach uses abstraction and refinement processes to travel from one level to another one. Finally, let us mention (Cox et al. 2005) and (Spalazzi 2001) that are syntheses on case-based planning.

Process-Oriented CBR (PO-CBR)

A PO-CBR system is a CBR system in which cases represent processes. PO-CBR has some similarities with case-based planning but differs in the same way as processes differ from plans: the latters are usually strongly related with formal operators (defined by conditions and actions), whereas a process is a structured set of *tasks* which are in general atomic objects (names). The most classical representation of cases in PO-CBR is the one of workflows. A selection of papers on PO-CBR has been published in (Minor et al. 2014).

Conversational CBR (CCBR)

Classically, in a CBR system, the target problem is given entirely to the system and then solved by the CBR process. By contrast, in conversational CBR, the target problem is interactively built through a human-machine dialog (Aha et al. 2001),

using the case base: based on the initial query, the case base is searched and specific questions are posed to the user. Then, the process repeats itself until a sufficiently detailed target problem is given. This approach to CBR is used in particular for help-desk applications.

Textual CBR (TCBR)

In many applications, cases are, at the start of the development, available in an informal way, for instance in the form of texts in natural language. The issue of TCBR is to apply CBR to cases encoded as texts (Weber et al. 2005). One way to do this consists in translating (semi-)automatically these texts into formal cases using natural language processing techniques (see, e.g., (Dufour-Lussier et al. 2014)), such as information extraction (as in (Brüninghaus and Ashley 2001)). Another way consists in manipulating directly textual cases. For this purpose, similarity measures between texts are used, for example, compression-based similarity measures (Cunningham 2009).

Trace-Based Reasoning (TBR)

TBR is a reasoning type similar to CBR, with some differences. If CBR considers so-called problem-solving *episodes*, CBR systems exploiting the temporal facets of an episode are rare, just as the descriptors involved are not necessarily temporally located in relation to one another. Moreover, in CBR, a problem-solving episode is considered independently of the different "contexts" in which the episodes were held.

Human experience, when it is considered as temporal by essence, can be represented by a temporal trace revealing elements of an underlying implicit process. For instance, the trace of use of a computer device or program captures some of the user knowledge needed by his/her task. The trace theory gives a definition of this notion of trace, how it can be represented together with the way the retrieval of episodes of use are computed. When the traces are exploited on the basis of retrieval and adaptation, TBR can be seen as a variation on CBR (Georgeon et al. 2011; Mille 2006; Zarka et al. 2011) and is based on a cycle similar to the one of Fig. 2.

CBR Applied to Particular Fields

There has been many applications of CBR to medical domains as well as to other fields of health science, for various tasks such as assisting diagnosis or treatment, for tasks in medical engineering, etc. This can be explained in part by the fact that the knowledge of physicians combine theoretical knowledge (comparable to the domain knowledge in CBR) and experience (that is represented by cases in CBR). The papers (Bichindaritz and Marling 2006) and (Begum et al. 2011) present syntheses of work on CBR to health science.

In a similar way, CBR has been applied to the legal domain (see, e.g., (Brüninghaus and Ashley 2001)), in which laws correspond to domain knowledge and legal precedents to cases.

In fact, CBR has been widely applied to many domains in which an important part of the knowledge consists in specific experience, such as architecture

(Dave et al. 1995), cooking (Cordier et al. 2014), design (Goel 1989), forest fires (Rougegrez 1994), games (Woolford and Watson 2017), music (de Mántaras 1998), running (Smyth and Cunningham 2017), theorem prooving (Melis 1995) (to cite only a few of such domains with particular examples).

3 Reasoning by Analogy and Analogical Proportions

The role of analogy in human reasoning has been acknowledged for a long time. Analogical reasoning exploits parallels between situations. It refers to the reasoning with which the human mind infers from an observed similarity another similarity that is not known. While induction goes from several specific situations to a general rule, analogy goes from one similarity between specific situations to another one. It enables us to state analogies for explanation purposes, for drawing plausible conclusions, or for creating new devices by transposing old ones in new contexts. For this reason, analogical reasoning has been studied for a long time, in philosophy, e.g., (Hesse 1966), in cognitive psychology, e.g., (Gentner et al. 2001; Hofstadter and Sander 2013; Holyoak 2005), and in artificial intelligence, e.g., (Helman 1988; Hofstadter and Mitchell 1995; Melis and Veloso 1998a), under various approaches (French 2002; Prade and Richard 2014a; McGreggor et al. 2014). Thus, since the beginnings of artificial intelligence, researchers have been interested in analogical reasoning as a basis for efficient heuristics for solving puzzles where a series has to be completed (Evans 1964), or for speeding up automatic deduction processes (Becker 1969; Kling 1972). This latter idea has then been resumed and systematically explored in studies such as the ones of (Melis and Veloso 1998b) or (Sowa and Majumdar 2003). At the modeling level, analogy can be envisaged at least in two different manners, either (i) as a matter of mapping two situations, one considered as a source of information, the other as a target about which one would like to draw some inference, or (ii) in terms of analogical proportions, i.e., statements of the form "A is to B as C is to D". In the two following subsections, we consider these two views in sequence.

It should be also pointed out that case-based reasoning, as presented above, can be viewed as a form of analogical reasoning. As explained in the first part of this chapter, CBR uses a base of known cases often stored as (problem, solution) pairs. When confronted to a new problem B, the problems A similar to B such that A appears in a problem-solution pair (A, C) are retrieved from the case base. Using a so-called adaptation technique, the solution C of the problem A is transformed into a candidate solution D of B (see, e.g., (Aamodt and Plaza 1994)). Thus, one can say that the target pair (B, D) parallels pairs (A, C) retrieved from the information source, but we may also state that "solution D is to solution C as problem B is to problem A", which corresponds to the two above-mentioned views of analogy.

3.1 Analogy in Terms of Mappings

The classical view of analogy relies on the establishment of a parallel between two situations (or universes), which are described in terms of objects, properties of the objects, and relations linking the objects. It amounts to identifying one-to-one correspondences between objects in situation 1 and objects in situation 2, on the basis of similar properties and relations that hold both for the objects in situation 1 and for the objects in situation 2. This is the basis of approaches proposed in cognitive psychology. Usual illustrations of this view are Rutherford's analogy between the atom structure and the solar system, or the similarity between electricity and hydraulics equations.

At the forefront of the proposals coming from cognitive science in the last three decades, three leading approaches should be especially mentioned: the structure mapping theory (SMT) proposed by (Gentner 1983, Gentner 1989), the analogical constraint mapping approach proposed by (Holyoak and Thagard 1989), (Thagard et al. 1990), and the model of analogy making based on the idea of the parallel terraced scan developed by (Hofstadter and Mitchell 1995), (Mitchell 1993, Mitchell 2001).

Structure mapping theory views an analogy as a mapping between a source and a target domain. The associated structure-mapping engine (SME) (Falkenhainer et al. 1989) returns the correspondences between the constituents of the base and target descriptions (expressed in terms of relations, properties, and functions), a set of candidate inferences about the target according to the mapping, and a structural evaluation score. Such a view is closely related to the idea of structural similarity (Syrovatka 2000), and has been also advocated early in artificial intelligence (Winston 1980); see also (Gust et al. 2006) for a presentation of the HDTP model based on a second order logical modeling of SMT, and (Weitzenfeld 1984) for a discussion about the interest of isomorphic structures when comparing situations. Besides, the view of analogy as a constraint satisfaction process, also defended in (Indurkhya 1987; Van Dormael 1990), is at work in the analogical constraint mapping engine (ACME) (Holyoak and Thagard 1989; Holyoak et al. 1994), which represents constraints by means of a network of supporting and competing hypotheses regarding what elements to map, and where an algorithm identifies mapping hypotheses that collectively represent the overall mapping that best fits the interacting constraints.

Roughly speaking, following (French 2002), one may distinguish between three broad groups: (i) the symbolic models that establish a structural similarity between the source and target descriptions generally expressed in formal logic terms, as SME; (ii) the connectionist models well suited for representing relational structures with nodes and links between nodes, as in ACME from using a kind of neuron network-like structure, or in LISA (Hummel and Holyoak 1997) the strong constraint of pair-wise connection between objects is relaxed to partial and dynamic connections (see (French 2002) for other references); (iii) the hybrid models relying on a combination of the previous approaches. These latter models generally use constraint satisfaction networks, explore competing hypotheses and are stochastic in nature. They rather focus on the optimization process at work to extract the most plausible solution.

Moreover, this type of approach naturally embeds a graded view of similarity, while symbolic approaches have generally difficulties to handle similarity beyond mere identity. The COPYCAT project (Hofstadter and Mitchell 1995; Mitchell 1993) is probably one of the most well-known attempt of analogy-making program falling in the hybrid category. Based on a similar paradigm, let us also mention Tabletop (French and Hofstadter 1991; French 1995), and NARS (Wang 2009).

In the recent years, SMT (structure-mapping theory) has proved to be effective for handling several AI problems (Forbus et al. 2017), for instance for solving IQ tests. They have dealt with the Raven Progressive Matrices test (Raven 2000), which are non-verbal tests supposedly measuring general intelligence: A 3×3 Raven matrix exhibits 8 geometric pictures displayed as its 8 first cells: the remaining 9th cell is empty. In these tests, a set of 8 candidate pictures is also given among which the subject is asked to identify the solution. The approach uses a sketch understanding system named CogSketch (Forbus et al. 2011). It takes a sketch drawn by the user as input, which has to be segmented into objects, and generates a qualitative representation of those objects (or their edges and groups of objects), and their relations (relative position, topology, etc.). For instance, CogSketch can tell which objects are placed side by side, whether two objects intersect, or whether one is located inside another. At the end of the process, each picture is represented as an entity with attributes and relations with other entities. At this stage, we have obtained a representation of the relative position of the objects.

CogSketch uses this edge level representation (which identifies the corresponding edges in two distinct objects) to compare two objects in a sketch, with the aim of determining if there is a transformation (rotation, size modification) or even a deformation (total shape change) between these two objects. With this information, the objects with equivalent or strict shapes in common, are grouped together. At this stage, we have a representation of the modification between objects.

In order to select the correct answer for the target test, the system described in (Lovett et al. 2010) proceeds as follows:

1. The first two rows of the current matrix are evaluated via SME in order to generate some rules for both of them, which are called *pattern of variance* and are a representation of how the objects change across the row of images. There are four different strategies available to build up these patterns of variances.
2. SME is then used again, but now for comparing the two patterns of variance previously found for the top two rows, and obtaining a *similarity score*. This comparison is called *second-order* comparison as it operates on patterns instead of object representations.
3. This similarity score is compared to a threshold to determine its validity.
4. If the patterns of variance are considered similar enough, an *analogical generalization* (which is a new pattern) is built describing what is common to both rows.
5. Each one of the 8 candidate answers is scored by inserting that answer into the bottom row, computing a pattern of variance, and then using SME to compare this pattern to the generalization pattern for the top two rows. The final answer is the one with the highest score.

6. In the case where the two patterns of variance corresponding to the top rows are not similar enough, another strategy is applied.

3.2 Analogy in Terms of Proportions

The word *analogy* is also associated with the notion of *analogical proportions*, i.e., statements of the form "A is to B as C is to D". The idea of this type of statement goes back (at least) to Aristotle, and was inspired by geometric proportions ($\frac{A}{B} = \frac{C}{D}$) and arithmetic proportions ($A - B = C - D$) between numbers. As can be seen, such proportion involve four elements, considered by pairs. Here are examples of such proportions: "calf is to bull as foal is to stallion", "colibri is to birds as mouse is to mammals", "beer is to Germany as wine is to France". In the first example, the four items involved are animals, which are thus pairwise comparable using the same features. In the second example, we have still animals, but species and orders. In the last example, the four items clearly belong to two different categories: here A and C are drinks while B and D are countries. In that latter case, the 'is to' refers to some relationship(s) existing between two items belonging to two distinct categories respectively, A and B on the one hand, C and D on the other hand, and the 'as' expresses the identity of this/these relationship(s). In the first example, 'is to' may be understood as referring to a mere comparison, moreover B and C commute leading to a new acceptable proportion, which is much more debatable in the last two examples, and especially the last one. In the following, we mainly address the first kind of proportion where the four items belong to the same category. Regarding the second kind of proportion, one may mention a preliminary work that bridges formal concept analysis with analogical proportions and looks for metaphors in a formal context (an example of metaphor is "Dugléré is the Mozart of (French) cooking" (in the XIXth century!), which is clearly related to the proportion "Dugléré is to (French) cooking as Mozart is to music") (Miclet et al. 2014).

Some of the artificial intelligence studies on analogical reasoning have focused on analogical proportions. This is the case for two already mentioned works. The ANALOGY program (Evans 1964) which was able – in an empirical way not directly applicable to other domains – to properly select a figure composed of geometrical elements, among different proposed choices, in order to give an "analogical" solution to three figures of the same nature. Some 30 years later, the COPYCAT system (Hofstadter and Mitchell 1995) was able to make a similar solving for triples of character strings to be completed by a fourth string, using a different approach based on artificial neural nets (see (French 2002) for a detailed discussion).

An attempt to formalize analogical reasoning started from the idea that $Q(t)$ can be inferred from $(P(s), Q(s))$ and $P(t)$ (where P and Q are predicates). This can be read as the proportion "$P(s)$ is to $Q(s)$ as $P(t)$ is to $Q(t)$", and indeed the analogical jump from $(P(s), Q(s))$ and $P(t)$ to $Q(t)$ can be seen as a form of analogical proportion-based inference (Bounhas et al. 2017a). However, the idea developed in (Davies and Russell 1987; Russell 1989) was to add additional information in

order to make the inference pattern valid by requiring the implicit hypothesis that P determines Q inasmuch as $\nexists x\, P(x) \wedge \neg Q(x)$. This may be ensured if there exists an underlying functional dependency, or more generally, if it is known for instance that when something is true for an object of a certain type, then it is true for all objects of that type. Besides, the statement "P determines Q" which can be possibly translated into $\forall x(P(x) \Rightarrow Q(x))$. If this functional dependency is considered too strong, it may be weakened, for instance into "The more similar $P(s)$ and $P(t)$ are, the more it is guaranteed as possible that $Q(s)$ and $Q(t)$ are similar" (where P and Q are now gradual predicates) (Dubois et al. 2002). This leads to a potential formalization of case-based reasoning. More recently, it has been presented in (Weller and Schmid 2007) an approach based on anti-resolution w.r.t. an equational theory for solving analogical proportions of the form "A is to B as C is to D" where D is unknown, by applying the same transformation to B as the one that enables us to go from A to C.

For about two decades, a series of European studies (Federici et al. 1996; Lepage 2001; Yvon et al. 2004; Stroppa and Yvon 2005b), summarized below, has aimed at developing formal models of analogical proportions and at showing their interest, in particular in computational linguistics (see (Stroppa and Yvon 2005a; Lepage et al. 2009 and Langlais and Patry 2007)). These studies start from the fact that analogical proportions obey postulates. Indeed, it has been observed for a long time that an analogical proportion "A is to B as C is to D", denoted by $A : B :: C : D$ in the following, should satisfy the following remarkable properties:

$$\begin{aligned} \textit{Symmetry of the relation "as"}: \quad & A : B :: C : D \Leftrightarrow C : D :: A : B \\ \textit{Exchange of the means}: \quad & A : B :: C : D \Leftrightarrow A : C :: B : D \end{aligned}$$

Furthermore, every expression of the form $A : A :: B : B$ or $A : B :: A : B$ is assumed to be a (trivial) analogical proportion. Besides, the two properties of symmetry and exchange, also satisfied by mathematical proportions, are at the origin of the term "analogical proportion". In particular, it has been noticed on the basis of the two properties introduced above, that the proportion $A : B :: C : D$ can be rewritten on the form of 8 equivalent proportions (including itself). It can be shown that the 24 possibilities of permutation of 4 objects can be partitioned in 3 equivalence classes of 8 proportions each, with an example of each class below:

$$A : B :: C : D \qquad A : B :: D : C \qquad A : C :: D : B$$

In addition, (Lepage 2001) has contributed to a model based on set theory of proportional analogies, where A, B, C and D are considered as situations characterized by sets of binary features. This model has been somewhat simplified in (Miclet and Prade 2009) and has led to the following definition:

$$A : B :: C : D \Leftrightarrow A \setminus B = C \setminus D \text{ and } B \setminus A = D \setminus C$$

where $\setminus$ denotes the set difference. This means that A differs from B as C differs from D and that B differs from A as D differs from C. This has a direct counterpart in a propositional logic modeling.

3.3 Proportional Analogy in Boolean Logic

When the terms of an analogical proportion take their values in {0, 1} (i.e., the focus is on whether a description feature is true or false), the proportion becomes a relation between 4 truth values, and can be expressed by the Boolean logic formula.

$$a : b :: c : d \text{ if and only if } ((a \wedge \neg b \equiv c \wedge \neg d) \wedge (b \wedge \neg a \equiv d \wedge \neg c))$$

which obviously fits with the above reading in terms of difference ($x \wedge \neg y$ is the logical difference between x and y). The 6 truth assignments of (a, b, c, d) making the proportional analogy $a : b :: c : d$ true appear in bold font in the table below. The truth values obey the logical expression given above (Miclet and Prade 2009; Prade and Richard 2013).

a	**0**	0	0	**0**	0	**0**	0	0	1	1	**1**	1	**1**	1	1	**1**
b	**0**	0	0	**0**	1	**1**	1	1	0	0	**0**	0	**1**	1	1	**1**
c	**0**	0	1	**1**	0	**0**	1	1	0	0	**1**	1	**0**	0	1	**1**
d	**0**	1	0	**1**	0	**1**	0	1	0	1	**0**	1	**0**	1	0	**1**
$a : b :: c : d$	**1**	0	0	**1**	0	**1**	0	0	0	0	**1**	0	**1**	0	0	**1**

The Boolean analogical proportion is a particular case of so-called *logical proportions* that are built from similarity and dissimilarity indicators (Prade and Richard 2013). When comparing two Boolean variables a and b there are two similarity indicators, namely a positive one $a \wedge b$ and a negative one $\neg a \wedge \neg b$, and two dissimilarity indicators $\neg a \wedge b$ and $a \wedge \neg b$.[2] Logical proportions connect four Boolean variables through a conjunction of two equivalences between similarity or dissimilarity indicators pertaining respectively to two pairs (a, b) and (c, d). More precisely a logical proportion is the conjunction of two equivalences between indicators for (a, b) on one side and indicators for (c, d) on the other side. In the case of analogical proportion only dissimilarity operators are used. There are 120 syntactically and semantically distinct logical proportions. All these proportions share a remarkable property: they are true for exactly 6 patterns of values of $abcd$ among 2^4 possible values. This is only a small subset of the $\binom{16}{6} = 8008$ quaternary Boolean operators true for only 6 patterns. For instance, $((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((a \wedge b) \equiv (c \wedge d))$ is a logical proportion, expressing that "a differs from b as c differs from d" and that "a is similar to b as c is similar to d", which is true for the 6 patterns 0000, 1111, 1010, 0101, 0001, and 0100. The reader is referred to (Prade and Richard 2013) for a thorough study of the different types of logical proportions.

Among logical proportions $LP(a, b, c, d)$ those satisfying the *code independence* property are of particular interest. This property expresses that there should be no distinction when encoding information positively or negatively. In other words, encoding truth (resp. falsity) with 1 or with 0 (resp. with 0 and 1) is just a matter of convention, and should not impact the final result. Thus we should have the following

[2] These indicators are also the building blocks of the view of similarity proposed by (Tversky 1977).

entailment between the two logical expressions: $LP(a, b, c, d) \Rightarrow LP(\neg a, \neg b, \neg c, \neg d)$. There only exist eight logical proportions that satisfy the above property (Prade and Richard 2013). The code independent proportions split into 4 *homogeneous* proportions that are symmetrical (one can exchange (a, b) with (c, d)) and 4 *heterogeneous* ones that are not symmetrical. Homogeneity here refers to the fact that in the expression of the proportions, both equivalences link indicators of the same kind (similarity or dissimilarity), while in the case of heterogeneous proportions they link indicators of opposite kinds. Homogeneous logical proportions include analogical proportion and two other closely related proportions:

- *reverse analogy*: $\text{Rev}(a, b, c, d) \triangleq ((\neg a \wedge b) \equiv (c \wedge \neg d)) \wedge ((a \wedge \neg b) \equiv (\neg c \wedge d))$
  It reverses analogy into "b is to a as c is to d". Indeed $\text{Rev}(a, b, c, d) = b : a :: c : d$.
- *paralogy*: $\text{Par}(a, b, c, d) \triangleq ((a \wedge b) \equiv (c \wedge d)) \wedge ((\neg a \wedge \neg b) \equiv (\neg c \wedge \neg d))$. It expresses that what a and b have in common (positively or negatively), c and d have it also, and conversely. It can be shown that $\text{Par}(a, b, c, d) = c : b :: a : d$, which provides an expression of analogical proportion in terms of *similarity* indicators.

Switching the positive and the negative similarity indicators pertaining to the pair (c, d) in $\text{Par}(a, b, c, d)$, we obtain the fourth homogeneous logical proportion called *inverse paralogy*: $\text{Inv}(a, b, c, d) \triangleq ((a \wedge b) \equiv (\neg c \wedge \neg d)) \wedge ((\neg a \wedge \neg b) \equiv (c \wedge d))$. $\text{Inv}(a, b, c, d)$ states that "what a and b have in common, c and d do not have it and conversely". It expresses a kind of "orthogonality" between the pairs (a, b) and (c, d). Inv is the unique logical proportion (among the 120's!) which remains unchanged under any permutation of two terms among the four (Prade and Richard 2013).

The four *heterogeneous* logical proportions have a quite different semantics. They express that there is an intruder among $\{a, b, c, d\}$, which is not a, which is not b, which is not c, and which is not d respectively (Prade and Richard 2014b). They are at the basis of an "oddness" measure, which may be used in classification, following the straightforward idea of classifying a new item in the class where it appears to be the least at odds (Bounhas et al. 2017b).

Besides, the equation $a : b :: c : x$ where x is the unknown may have no solution (this is the case, e.g., for $1 : 0 :: 0 : x$). In the Boolean case the solution exists only if $a = b$ or $a = c$. When this solution exists, it is unique and given by $x = c \equiv (a \equiv b)$ (that is also the solution, when it exists, of $Rev(a, b, c, x)$ and of $Par(a, b, c, x)$. This result was first noticed in (Klein 1982) in an empirical approach based on semiotic observations, which made no distinction between $a : b :: c : d$, $Rev(a, b, c, d)$, and $Par(a, b, c, d)$ (Prade and Richard 2013).

Let us now consider objects described by means of a set of Boolean features (binary attributes). In this setting, logical reasoning by analogy consists in identifying the analogical proportions that hold on a subset of attributes between four objects and to infer the value of the remaining attributes, or of the class attribute for the fourth object, knowing the value for the three others. This idea has been successfully used for building the solution of Raven Progressive Matrices IQ tests, *without* the help of any candidate solutions (Correa Beltran et al. 2016).

In terms of machine learning (see chapter "Statistical Computational Learning" in this volume and chapter "Designing Algorithms for Machine Learning and Data Mining" in Volume 2), the objective is to learn the value $u(x)$ of a function u for an object x. Let us consider classification: in this framework, $u(x)$ is the label of a class chosen in a finite set of classes. A training set $\mathscr{S}$ composed of examples of objects a_i, for which the supervision $u(a_i)$ is known, is available:

$$\mathscr{S} = \{(a_1, u(a_1)), \ldots, (a_m, u(a_m))\}$$

The idea is to find 3 objects a, b and c of $\mathscr{S}$ such that $a : b :: c : x$.[3] It must be noticed that the object x to be classified is compared to a *triple* of objects (a, b, c), which differs from the classification based on the k nearest neighbors for which x is compared to its neighbors taken *individually*. Then, the value of u on x can be computed by solving the equation $u(a) : u(b) :: u(c) : u(x)$.

This technique is based on the hypothesis that to the analogical relation between the object descriptors corresponds an analogical relation between the values of the supervision function u. This hypothesis has been verified with success for classification rule learning with objects described by binary and nominal attributes (noting that a nominal attribute can be replaced by a set of binary attributes) on classical databases (Bayoudh et al. 2007a).

An interesting feature of such analogical classifiers is that the size of the learning set can be drastically reduced without decreasing the success rate on a test set. This property can be explained in the following way. Call the *analogical extension* $AE(S)$ of a set S of m vectors (binary, nominal or numerical) the multiset composed of the m^3 solutions to the equations $a : b :: c : x$, where a, b and c are elements of S. When the vectors are numerical and the arithmetic proportion is used, $AE(S)$ has same mean and covariance matrix as S. Analogical classification with S as a learning set is indeed very similar in that case to a k-nearest neighbours method using $AE(S)$, but requires m instead of m^3 learning patterns. The price to pay is in classification time of a new pattern, but it can be managed with preprocessing methods of S.

Classification based on analogical proportions has also been generalized to numerical features thanks to a multiple-valued extension of the logical definition of analogical proportion (Bounhas et al. 2017a).

Recent formal studies have shown that analogical classifiers always give exact predictions in the special cases where the classification process is governed by an affine Boolean function (which includes x-or functions) and only in this case, which does not prevent to get good results in other cases (as observed in practice), but which is still to be better understood (Couceiro et al. 2017). This suggests that analogical proportions enforce a form of linearity, just as numerical proportions fit with linear interpolation.

[3] Or to find all the triples (a, b, c) realizing that and then to make a vote, as in the k-nearest neighbor method. Empirical studies suggest that if we restrict ourselves to triples where c is a k-nearest neighbor (a, b being generally quite far) this does not really harm the results (Bounhas et al. 2017a).

3.4 Analogical Proportions Between Sequences

In order to obtain a general notion of analogical proportion and to apply it to various spaces, Yvon and Stroppa have proposed a definition that satisfies the symmetry and exchange postulates and that is helpful to solve analogical equations (Stroppa and Yvon 2005c). They take lessons from geometric proportions in $\mathbb{R}$, where the rule of three applies: $\frac{u}{v} = \frac{w}{x} \Leftrightarrow u \times x = v \times w$. In order to analyse the second relation, it is natural to decompose the four numbers in prime factors. For example $\frac{6}{10} = \frac{21}{35}$ can be written $\frac{2\times3}{2\times5} = \frac{7\times3}{7\times5}$. In other words, we can say that the numbers $u = 6$, $v = 10$, $w = 21$ and $x = 35$ are in analogical proportion because there exist four factors $f_1 = 2$, $f_2 = 7$, $f_3 = 3$ and $f_4 = 5$ such that $u = f_1 \times f_3$, $v = f_1 \times f_4$, $w = f_2 \times f_3$, $x = f_2 \times f_4$.

Is it possible to transfer this cross factorization in another universe? Let $\Sigma^\star$ be the set of sequences on an alphabet Σ with the non commutative concatenation operation (explicitely denoted by "."). For instance, let us consider the numerical analogy $18 : 63 :: 30 : 105$ and an analogy on sequences, here made of French words: déridés : ridons :: démarchés : marchons. They can be factorized in the following way:

18	= 2	×	3	× 2	×	1	×	3		
63	= 1	×	3	× 1	×	7	×	3		
30	= 2	×	5	× 2	×	1	×	3		
105	= 1	×	5	× 1	×	7	×	3		
déridés	= dé	.	rid	. é	.	ε	.	s		
ridons	= ε	.	rid	. ε	.	on	.	s		
démarchés	= dé	.	march	. é	.	ε	.	s		
marchons	= ε	.	march	. ε	.	on	.	s		

It can be noted that, in both cases, each quadruple of factors of rank i read in a column is either (f_i, f_i, g_i, g_i) or (f_i, g_i, f_i, g_i). A factor may be the neutral element of the considered universe (1 for multiplication in $\mathbb{R}$ and ε for concatenation in $\Sigma^\star$).

This idea of factorizing in elementary analogical proportions has been used by Yvon and Stroppa for defining algorithms for checking proportions and for solving analogical equation between sequences, using systems with finite states. This idea was addressed in a different way in (Miclet et al. 2008) where an extension of the edit distance is used that defines an *analogical dissimilarity* between four sequences and leads to an approximate solving of analogical equations.

Another application of analogical equation solving on sequences is the generation of plausible patterns. In this framework, the study of (Stroppa and Yvon 2006) was about applications to phonetics and morphology. In (Bayoudh et al. 2007b), it has been shown how to generate plausible training examples for the recognition of handwritten characters.

4 Interpolative Reasoning

Case-based reasoning relates *two* similar problems and transfers the solution of one of them to the other one. An analogical proportion states particular similarity and dissimilarity relations between *four* terms. Thus, case-based reasoning and analogical reasoning are two forms of similarity-based reasoning. But they are not the only ones. In this last section of the chapter we present a brief overview of studies based on another similarity-based reasoning: the interpolative (and extrapolative) reasoning. Interpolation allows us, when the current situation is intermediate between known situations, to conclude in an intermediate way with respect to the conclusions of these situations. When the conclusion of only one situation, close to the current situation, is known, a solution can be extrapolated for the current situation, provided that some available information about the variations around this close situation can be exploited. Therefore, interpolation and extrapolation need variables with ordered referentials and some notions of similarity. These forms of reasoning, though they are important in commonsense reasoning, have got very little attention in AI outside the community working on fuzzy sets and approximate reasoning. First, some recalls about fuzzy sets and approximate reasoning are given. Then, interpolation and extrapolation in this framework are discussed. Finally, some studies on this subject that are not based on fuzzy sets are briefly presented.

4.1 Fuzzy Sets and Approximate Reasoning

In addition to the representation of uncertainty (see chapters "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" and "Representations of Uncertainty in Artificial Intelligence: Beyond Probability and Possibility" in this volume) and preferences (see chapter "Compact Representation of Preferences" in this volume), the semantics of fuzzy sets can be based on *similarity*. In fact, this corresponds to the first interpretation pointed out for fuzzy sets (Bellman et al. 1966): the higher the membership degree of an element is, the closest to the core of the fuzzy set it is (the core of a fuzzy set being the set of elements with a membership degree equal to 1). For instance, a fuzzy set A with a triangular membership degree μ_A such that a is the only value verifying $\mu_A(a) = 1$ represents the set of values more or less close to a (the closeness linearly decreases when the element goes away from a if μ_A is triangular). More generally, a fuzzy rule of the form "if x is A then y is B" can be intuitively understood as "if x is close to a then y is close to b" when A and B are two fuzzy sets of respective cores $\{a\}$ and $\{b\}$. This idea can be extended to rules with several conditions. Deduction based on these rules can be done thanks to the approximate reasoning method that is presented now.

The principle of approximate reasoning introduced in (Zadeh 1979) (see (Bouchon-Meunier et al. 1999) for a detailed overview) is based on a mechanism of combination/projection of the representation of the available pieces of information. These pieces of information are represented by possibility distributions from which a new

possibility distribution, representing the conclusion, can be deduced. So, let X and Y be two variables having their values respectively in referentials U and V. If it is known that "X is A'" and that "if X is A then Y is B", represented respectively by $\pi_X = \mu_{A'}$ and $\pi_{(X,Y)} = \mu_A \rightarrow \mu_B$, it can be concluded that

$$\mu_{B'}(v) = \pi_Y(v) = \sup \min(\pi_X(u), \pi_{(X,Y)}(u, v))$$

where A, A' (resp., B, B') are the fuzzy subsets of U (resp., V) that restrict the more or less possible values of X and Y, and $\rightarrow$ is a logical connector that defines here a fuzzy relation on $U \times V$ modeling the relation between X and Y expressed by the "if …then …" rule linking them. The above expression is nothing but the computation of the marginal possibility distribution of Y from the joint distribution of (X, Y) obtained by the conjunctive combination of available pieces of information. The pattern of reasoning corresponding to the schema, from "if X is A'" and "if X is A then Y is B" it entails that "Y is B'", corresponds to the idea of "generalized modus ponens"[4] (Zadeh 1979). According to the meaning given to the rule "if …then …", different operators can be chosen for $\rightarrow$: they are multivalued conjunctions or implications (Dubois and Prade 1996) depending on the interpretation of the rule as specifying that all the elements of the (fuzzy) Cartesian product $A \times B$ are values that are *all* possible for (X, Y) or, on the contrary, that the elements of $A \times \overline{B}$ are impossible (where $\overline{B}$ denotes the complement of B).

This type of approximate reasoning has been applied to case-based reasoning by using fuzzy rules expressing that "the more two situations are similar from some viewpoint, the more it is guaranteed possible that they are according to other viewpoints" (Hüllermeier et al. 2002) (see Sect. 4.1.3 in chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" of this volume for a brief presentation of this kind of rules "with guaranteed possibility") and they can then be related to methods of the k-nearest neighbors type.

4.2 Graduality and Interpolation

The choice of a particular implication connective, the so-called Gödel implication ($s \rightarrow t = t$ if $s \leq t$ and $s \rightarrow t = 0$ if $s > t$) or, simply its binary restriction called Rescher-Gaines implication ($s \rightarrow t = 1$ if $s \leq t$ and $s \rightarrow t = 0$ if $s > t$) allows us to give a *gradual* semantics (Dubois and Prade 1992) to the rule under the form "the more X is A, the more Y is B", which can be also read as "the closer X is to a the closer Y is to b". This is equivalent to a set of non fuzzy rules "if $X \in A_\alpha$ then

[4]Rather then seeing a fuzzy set as a set of elements close to its core value, similarity measures between fuzzy sets themselves can be defined, and then it is possible to give some meaning to the analogical proportion of the form $A : A' :: B : B'$, but B' obtained this way does not have, in general, a reason to be compatible with the result of the generalized modus ponens as defined above. However, some choice of resemblance relations and of operators allows us to reconcile these two viewpoints; see for example (Bouchon-Meunier and Valverde 1999).

$Y \in B_\alpha$ for $\alpha \in (0, 1]$ that express well the fact that the closer X is to a, i.e., in a cut $A_\alpha = \{u \in U | \mu_A(u) \geq \alpha\}$ of high degree α, the more Y is in a cut of B of high degree (the more the cut is of high degree α, the closer to a the values in the cut). It can be shown that the approximate reasoning applied to a base of gradual rules[5] offering an appropriate and sufficient coverage of U allows us to model linear or non linear interpolations (Dubois and Prade 1992). The situation where the fuzzy subsets A_i correspond to the fuzzy rule base "if X is A_i then Y is B_i" for $i = 1, n$ does *not* constitute a coverage, even in an approximate way, of U has been also studied by several authors; see (Perfilieva et al. 2012) for an overview of generalized interpolation methods between "scattered" rules.

The semantics in terms of similarity of a fuzzy set is also a starting point of (Ruspini 1991) for defining a gradual consequence relation. The initial intuition is simple: the consequence relation $p \vdash q$ between two propositional statements p and q in classical logic corresponds to an inclusion relation $[p] \subseteq [q]$ between their respective sets of models. The inclusion can be weakened into an approximate inclusion in two very different ways (when $[p] \nsubseteq [q]$): either it is required only that *all the preferred* models of p are included in $[q]$, and this is the starting point (from a semantic viewpoint) of nonmonotonic reasoning (see chapters "Knowledge Representation: Modalities, Conditionals and Nonmonotonic Reasoning" and "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" of this volume), or it is required only that $[p]$ is included in the set of models of q extended to the counter-models of q that are *close* enough to its models. This leads to two different types of weakened consequence relations, of which the properties partly differ (Dubois and Prade 1998). According to this last view, a logical approach to interpolation has been proposed (Dubois et al. 1997b). Let us finally mention the formal framework of "extensional" fuzzy sets (Klawonn 2000) (i.e., fuzzy sets that are unions of fuzzy "clusters" of elements with respect to a fuzzy relation of similarity) that allows to formally define a partionning process of data that can afterwards be used to build fuzzy rules adapted to existing data.

4.3 Similarity-Based Qualitative Reasoning

A more qualitative approach to similarity-based reasoning, that does not require the definition of membership functions, has been more recently proposed. It consists in interpreting terms that are not a priori vague, in a flexible way. For instance, having the possibility to interpret "married" as "married or living as husband and wife" allows us to solve inconsistencies in information merging problems (Schockaert and Prade 2011). In the same spirit, it is possible to enrich sets of categorization rules using geometrical-like properties in conceptual spaces in the sense of (Gärdenfors 2000). The properties appearing in the conditions or conclusions of these rules are

[5]Gradual rules have been independently considered under the name of "topoi" in (Raccah 1996), from a cognitive perspective.

treated like abstract entities. By using as primitive the relation "to be between" for these entities, it is possible to obtain schemas of interpolative reasoning that can be characterized at the same time semantically and syntactically, as well as an extrapolative reasoning scheme, based on a "parallelism" relation between pairs of concepts, staying in both cases at a symbolic level that requires only the knowledge of relations between entities (Schockaert and Prade 2013).

There exist other forms of qualitative reasoning (see chapter "Qualitative Reasoning about Time and Space" in this volume). Let us also mention, in this perspective, an approach for reasoning on relative order of magnitude, based on the principles of combination and projection of approximate reasoning (recalled in Sect. 4.1 above), and using a representation of proximity and of negligibility in terms of fuzzy relations (Hadj Ali et al. 2003).

5 Conclusion

Human judgement and reasoning often use comparisons and rely on similarities, but also on the perception of differences. It is also at work in decision making; see (Gilboa and Schmeidler 1995; Dubois et al. 1997a) for similarity-based approaches, not reviewed here. As surveyed in this chapter, different AI approaches have tried to give substance to this idea, in particular in case-based reasoning and in analogical reasoning. In these two types of reasoning two operations of primary importance emerge: similarity-based search (e.g., for case retrieval) and adaptation. Assessing the similarity is always a delicate issue and can be considered in different ways. Even if the starting intuitions seem to be similar, the different approaches detailed here can be distinguished according to the way situations are related. The study of adaptation is not less rich and shows the importance that must be given to domain knowledge in the reasoning process. This is also an opportunity to establish a link with some aspects of knowledge discovery and, more generally with learning, which are also related to reasoning issues.

References

Aamodt A, Plaza E (1994) Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun 7(1):39–58

Aha DW, Breslow LA, Muñoz-Avila H (2001) Conversational case-based reasoning. Appl Intell 14(1):9–32

Alchourrón CE, Gärdenfors P, Makinson D (1985) On the logic of theory change: partial meet functions for contraction and revision. J Symb Log 50:510–530

Bayoudh S, Miclet L, Delhay A (2007a) Learning by analogy: a classification rule for binary and nominal data. In: Veloso MM (ed) Proceedings of the 20th international joint conference on artificial intelligence (IJCAI 2007), Hyderabad, 6–12 January 2007. AAAI Press, pp 678–683

Bayoudh S, Mouchère H, Miclet L, Anquetil É (2007b) Learning a classifier with very few examples: analogy-based and knowledge-based generation of new examples for character recognition. In: Kok JN, Koronacki J, de Mántaras RL, Matwin S, Mladenic D, Skowron A (eds) Proceedings of the 18th European conference on machine learning (ECML 2007) Warsaw, 17–21 September 2007. LNCS, vol 4701. Springer, Berlin, pp 527–534

Becker JD (1969) The modeling of simple analogic and inductive processes in a semantic memory system. In: Proceedings of the 1st international joint conference on artificial intelligence (IJCAI'69), pp 655–668

Begum S, Ahmed MU, Funk P, Xiong N, Folke M (2011) Case-based reasoning systems in the health sciences: a survey of recent trends and developments. IEEE Trans Syst Man Cybern Part C (Appl Rev) 41(4):421–434

Bellman RE, Kalaba R, Zadeh L (1966) Abstraction and pattern classification. J Math Anal Appl 13:1–7

Bergmann R, Wilke W (1995) Building and refining abstract planning cases by change of representation language. J Artif Intell Res 3:53–118

Bichindaritz I, Marling C (2006) Case-based reasoning in the health sciences: what's next? Artif Intell Med 36(2):127–135

Bouchon-Meunier B, Dubois D, Godo L, Prade H (1999) Fuzzy sets and possibility theory in approximate and plausible reasoning. In: Bezdek J, Dubois D, Prade H (eds) Fuzzy sets in approximate reasoning and information systems. The handbooks of fuzzy sets. Kluwer, Boston, pp 15–190

Bouchon-Meunier B, Valverde L (1999) A fuzzy approach to analogical reasoning. Soft Comput 3:141–147

Bounhas M, Prade H, Richard G (2017a) Analogy-based classifiers for nominal or numerical data. Int J Approx Reason 91:36–55

Bounhas M, Prade H, Richard G (2017b) Oddness/evenness-based classifiers for Boolean or numerical data. Int J Approx Reason 82:81–100

Brüninghaus S, Ashley K (2001) The role of information extraction for textual CBR. Case-based reasoning research and development, pp 74–89

Carbonell JG (1983) Learning by analogy: formulating and generalizing plans from past experience. In: Michalski RS, Carbonell JG, Mitchell TM (eds) Machine learning, an artificial intelligence approach. Morgan Kaufmann Inc., San Francisco, pp 137–161

Carbonell JG (1986) Derivational analogy: a theory of reconstructive problem solving and expertise acquisition. Machine learning, vol 2. Morgan Kaufmann Inc., San Francisco, pp 371–392

Chang L, Sattler U, Gu T (2014) Algorithm for adapting cases represented in a tractable description logic. In: Lamontagne L, Plaza E (eds) Case-based reasoning research and development, Proceedings of ICCBR-2014. Springer, Berlin, pp 63–78

Cojan J, Lieber J (2008) Conservative adaptation in metric spaces. In: Althoff K-D, Bergmann R, Minor M, Hanft A (eds) ECCBR. Lecture notes in computer science, vol 5239. Springer, Berlin, pp 135–149

Cojan J, Lieber J (2011) An algorithm for adapting cases represented in ALC. In: Walsh T (ed) IJCAI. IJCAI/AAAI, pp 2582–2589

Cordier A, Fuchs B, Lieber J, Mille A (2007) Failure analysis for domain knowledge acquisition in a knowledge-intensive CBR system. In: Michael Richter RW (ed) Proceedings of the 7th international conference on case-based reasoning. LNAI. Springer, Berlin, pp 463–477

Cordier A, Fuchs B, de Carvalho LL, Lieber J, Mille A (2008) Opportunistic acquisition of adaptation knowledge and cases - the IAKA approach. In: Althoff K-D, Bergmann R, Minor M, Hanft A (eds) Advances in case-based reasoning, Proceedings of the 9th European conference, ECCBR 2008, Trier, Germany, 1–4 September 2008. Lecture notes in computer science, vol 5239. Springer, Berlin, pp 150–164

Cordier A, Dufour-Lussier V, Lieber J, Nauer E, Badra F, Cojan J, Gaillard E, Infante-Blanco L, Molli P, Napoli A, Skaf-Molli H (2014) Taaable: a case-based system for personalized cook-

ing. In: Montani S, Jain LC (eds) Successful case-based reasoning applications-2. Studies in computational intelligence, vol 494. Springer, Berlin, pp 121–162
Correa Beltran W, Prade H, Richard G (2016) Constructive solving of Raven's IQ tests with analogical proportions. Int J Intell Syst 31(11):1072–1103
Couceiro M, Hug N, Prade H, Richard G (2017) Analogy-preserving functions: a way to extend Boolean samples. In: Sierra C (ed) Proceedings of the 26th international joint conference on artificial intelligence (IJCAI'17), Melbourne, 19–25 August 2017, pp 1575–1581
Cox MT, Muñoz-Avila H, Bergmann R (2005) Case-based planning. Knowl Eng Rev 20(3):283–287
Craw S, Wiratunga N, Rowe R (2006) Learning adaptation knowledge to improve case-based reasoning. Artif Intell 170(16–17):1175–1192
Cunningham P (2009) A taxonomy of similarity mechanisms for case-based reasoning. IEEE Trans Knowl Data Eng 21(11):1532–1543
d'Aquin M, Badra F, Lafrogne S, Lieber J, Napoli A, Szathmary L (2007) Case base mining for adaptation knowledge acquisition. In: Veloso MM (ed) IJCAI, pp 750–755
Dave B, Schmitt G, Shih S-G, Bendel L, Faltings B, Smith I, Hua K, Bailey S, Ducruet J-M, Jent K (1995) Case-based spatial design reasoning. In: Haton J-P, Keane M, Manago M (eds) Advances in case-based reasoning - Second European workshop, EWCBR'94. LNCS, vol 984. Springer, Berlin, pp 198–210
Davies TR, Russell SJ (1987) A logical approach to reasoning by analogy. In: Proceedings of the 10th international joint conference on artificial intelligence (IJCAI'87). Morgan Kaufmann, pp 264–270
de Mántaras RL (1998) It Don't Mean A Thing (If It Ain't Got That Swing). In: Prade H (ed) Proceedings of the 13th European conference on artificial intelligence (ECAI-98), Brighton, United Kingdom, pp 694–696
Dubois D, Esteva F, Garcia P, Godo L, de Mántaras RL, Prade H (1997a) Fuzzy modelling of case-based reasoning and decision. In: Leake DB, Plaza E (eds) Proceedings of the 2nd international conference on case-based reasoning research and development (ICCBR-97), Providence, RI, 25–27 July 1997. LNCS, vol 1266. Springer, Berlin, pp 599–610
Dubois D, Esteva F, Garcia P, Godo L, Prade H (1997b) A logical approach to interpolation based on similarity relations. Int J Approx Reason 17:1–36
Dubois D, Hüllermeier E, Prade H (2002) Fuzzy set-based methods in instance-based reasoning. IEEE Trans Fuzzy Syst 10:322–332
Dubois D, Prade H (1992) Gradual inference rules in approximate reasoning. Inf Sci 61:103–122
Dubois D, Prade H (1996) What are fuzzy rules and how to use them. Fuzzy Sets Syst 84:169–185
Dubois D, Prade H (1998) Similarity versus preference in fuzzy set-based logics. In: Orlowska E (ed) Modelling incomplete information: rough set analysis. Physica Verlag, Heidelberg, pp 441–461
Dufour-Lussier V, Le Ber F, Lieber J, Nauer E (2014) Automatic case acquisition from texts for process-oriented case-based reasoning. Inf Syst
Evans T (1964) A heuristic program to solve geometry-analogy problems. In: Proceedings of the A.F.I.P. spring joint computer conference, vol 25, pp 5–16
Falkenhainer B, Forbus KD, Gentner D (1989) The structure-mapping engine: algorithm and examples. Artif. Intell. 41(1):1–63
Federici S, Pirrelli V, Yvon F (1996) A dynamic approach to paradigm-driven analogy. In: Wermter S, Riloff E, Scheler G (eds) Connectionist, statistical, and symbolic approaches to learning for natural language processing. LNCS, vol 1040. Springer, Berlin, pp 385–398
Forbus K, Usher J, Lovett A, Lockwood K, Wetzel J (2011) CogSketch: sketch understanding for cognitive science research and for education. Top Cogn Sci 3(4):648–666
Forbus KD, Ferguson RW, Lovett AM, Gentner D (2017) Extending SME to handle large-scale cognitive modeling. Cogn Sci 41(5):1152–1201
French RM (1995) The subtlety of sameness: a theory and computer model of analogy-making. MIT Press, Cambridge
French RM (2002) The computational modeling of analogy-making. Trends Cogn Sci 6(5):200–205

French RM, Hofstadter D (1991) Tabletop: an emergent, stochastic model of analogy-making. In: Proceedings of the 13th annual conference of the cognitive science society. Lawrence Erlbaum, Hillsdale, NJ, pp 175–182
Fuchs B, Mille A (1999) A knowledge-level task model of adaptation in case-based reasoning. In: Branting K, Althoff K-D, Bergmann R (eds) Proceedings of the third international conference on case-based reasoning, ICCBR-99. Lecture notes in artificial intelligence, vol 1650. Springer, Berlin, pp 118–131
Fuchs B, Lieber J, Mille A, Napoli A (2000) An algorithm for adaptation in case-based reasoning. In: Horn W (ed) 14th European conference on artificial intelligence - ECAI'2000, Berlin. IOS Press, Amsterdam, pp 45–49
Fuchs B, Lieber J, Mille A, Napoli A (2014) Differential adaptation: an operational approach to adaptation for solving numerical problems with CBR. Knowl Based Syst 68:103–114
Gaillard E, Infante-Blanco L, Lieber J, Nauer E (2014) Tuuurbine: a generic CBR engine over RDFS. In: Case-based reasoning research and development, vol 8765. Cork, Ireland, pp 140–154
Gärdenfors P (2000) Conceptual spaces the geometry of thought. MIT Press, Cambridge
Gentner D (1983) Structure-mapping: a theoretical framework for analogy. Cogn Sci 7(2):155–170
Gentner D (1989) The mechanisms of analogical learning. In: Vosniadou S, Ortony A (eds) Similarity and analogical reasoning. Cambridge University Press, New York, pp 197–241
Gentner D, Holyoak K, Kokinov B (2001) The analogical mind: perspectives from cognitive science. MIT Press, Cambridge
Georgeon OL, Mille A, Bellet T, Mathern B, Ritter FE (2011) Supporting activity modelling from activity traces. Expert Syst
Gilboa I, Schmeidler D (1995) Case-based decision theory. Q J Econ 110:605–639
Goel AK (1989) Integration of case-based reasoning and model-based reasoning for adaptive design problem solving. Ph.D. thesis, Ohio State University
Gust H, Kühnberger K, Schmid U (2006) Metaphors and heuristic-driven theory projection (HDTP). Theor Comput Sci 354(1):98–117
Hadj Ali A, Dubois D, Prade H (2003) Qualitative reasoning based on fuzzy relative orders of magnitude. IEEE Trans Fuzzy Syst 11:9–23
Hall RJ (1986) Learning by failing to explain. In: Proceedings of the fifth national conference on artificial intelligence (AAAI 86), pp 568–572
Hall RP (1989) Computational approaches to analogical reasoning: a comparative analysis. Artif Intell 39:39–120
Hammond K (1986) CHEF: a model of case-based planning. In: Press A (ed) Fifth national conference on artificial intelligence, Menlo Park, CA, pp 267–271
Hammond K (1990) Explaining and repairing plans that fail. Artif Intell 45(1–2):173–228
Hanney K (1996) Learning adaptation rules from cases. MSc thesis, Trinity College Dublin, Ireland
Helman DH (ed) (1988) Analogical reasoning: perspectives of artificial intelligence. Cognitive science, and philosophy. Kluwer, Dordrecht
Hesse M (1966) Models and analogies in science, 1st edn. Sheed & Ward, London, 1963; 2nd augmented edn. University of Notre Dame Press
Hofstadter D, Mitchell M (1995) The Copycat project: a model of mental fluidity and analogy-making. In: Hofstadter D (ed) Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought. Basic Books Inc., New York, pp 205–267
Hofstadter D, Sander E (2013) Surfaces and essences: analogy as the fuel and fire of thinking. Basic Books, New York
Holyoak K (2005) Analogy. The cambridge handbook of thinking and reasoning. Cambridge University Press, Cambridge
Holyoak KJ, Thagard P (1989) Analogical mapping by constraint satisfaction. Cogn Sci 13:295–355
Holyoak KJ, Novick LR, Melz ER (1994) Component processes in analogical transfer: mapping, pattern completion, and adaptation. In: Holyoak KJ, Barnden JA (eds) Advances in connectionist and neural computation theory, vol. 2: analogical connections. Ablex Publishing, Westport, pp 113–180

Hüllermeier E (2007) Case-based approximate reasoning. Springer, Berlin
Hüllermeier E, Dubois D, Prade H (2002) Model adaptation in possibilistic instance-based reasoning. IEEE Trans Fuzzy Syst 10(3):333–339
Hummel JE, Holyoak KJ (1997) Distributed representations of structure: a theory of analogical access and mapping. Psychol Rev 104(3):427–466
Indurkhya B (1987) Approximate semantic transference: a computational theory of metaphors and analogies. Cogn Sci 11:445–480
Katsuno H, Mendelzon A (1991) Propositional knowledge base revision and minimal change. Artif Intell 52(3):263–294
Klawonn F (2000) Fuzzy points, fuzzy relations and fuzzy functions. In: Novak V, Perfilieva I (eds) Discovering the world with fuzzy logic. Physica-Verlag, Heidelberg, pp 431–453
Klein S (1982) Culture, mysticism and social structure and the calculation of behavior. In: Proceedings of the 5th European conference on artificial intelligence - ECAI, pp 141–146
Kling R (1972) A paradigm for reasoning by analogy. Artif Intell 2:147–178
Koehler J (1996) Planning from second principles. Artif Intell 87:145–186
Kolodner J (1993) Case-based reasoning. Morgan Kaufmann, San Francisco
Koton P (1988) Reasoning about evidence in causal explanations. In: Press A (ed) Seventh national conference on artificial intelligence, Menlo Park, CA, pp 256–261
Langlais P, Patry A (2007) Translating unknown words by analogical learning. In: Joint conference on empirical methods in natural language processing (EMNLP) and conference on computational natural language learning (CONLL). Prague, pp 877–886
Leake D, Kinley A, Wilson D (1996) Acquiring case adaptation knowledge: a hybrid approach. In: Proceedings of the 14th national conference on artificial intelligence (AAAI). AAAI Press, pp 684–689
Lepage Y (2001) Analogy and formal languages. Electron Not Theor Comput Sci 53
Lepage Y, Migeot J, Guillerm E (2009) A measure of the number of true analogies between chunks in Japanese. In: Vetulani Z, Uszkoreit H (eds) Human language technology. Challenges of the information society, third language and technology conference, LTC 2007, Poznan, 5–7 October 2007, Revised Selected Papers. LNCS, vol 5603. Springer, Berlin, pp 154–164
Lieber J (2007) Application of the revision theory to adaptation in case-based reasoning: the conservative adaptation. In: Proceedings of the 7th international conference on case-based reasoning (ICCBR-07). Lecture notes in artificial intelligence, vol 4626. Springer, Belfast, pp 239–253
Lieber J, Napoli A (1996) Using classification in case-based planning. In: Wahlster W (ed) European conference on artificial intelligence (ECAI'96). Wiley, Chichester, pp 132–136
Lovett A, Forbus K, Usher J (2010) A structure-mapping model of Raven's progressive matrices. In: Proceedings of the 32nd annual conference of the cognitive science society, Portland, OR
McGreggor K, Kunda M, Goel AK (2014) Fractals and ravens. Artif Intel 215:1–23
McSherry D (1999) Demand driven discovery of adaptation knowledge. In: Proceedings of the sixteenth international joint conference on artificial intelligence. Morgan Kaufmann, San Francisco, pp 222–227
Melis E (1995) A model of analogy-driven proof-plan construction. In: Proceedings of the 14th international joint conference on artificial intelligence (IJCAI'95). Montréal, pp 182–189
Melis E, Lieber J, Napoli A (1998) Reformulation in case-based reasoning. In: Smyth B, Cunningham P (eds) Fourth European workshop on case-based reasoning, EWCBR-98. Lecture notes in artificial intelligence, vol 1488. Springer, Berlin, pp 172–183
Melis E, Veloso M (1998a) Analogy in problem solving. Handbook of practical reasoning: computational and theoretical aspects. Oxford University Press, Oxford
Melis E, Veloso M (1998b) Analogy in problem solving. In: del Cerro LF, Gabbay D, Ohlbach HJ (eds) Handbook of practical reasoning: computational and theoretical aspects, vol 17(1). Oxford University Press, Oxford, pp 1–73
Miclet L, Prade H (2009) Handling analogical proportions in classical logic and fuzzy logics settings. In: Sossai C, Chemello G (eds) Proceedings 10th European conference on symbolic and

quantitative approaches to reasoning with uncertainty (ECSQARU'09), Verona, 1–3 July 2009. LNCS, vol 5590. Springer, Berlin, pp 638–650
Miclet L, Bayoudh S, Delhay A (2008) Analogical dissimilarity: definition, algorithms and two experiments in machine learning. J Artif Intell Res (JAIR) 32:793–824
Miclet L, Barbot N, Prade H (2014) From analogical proportions in lattices to proportional analogies in formal concepts. In: Schaub T, Friedrich G, O'Sullivan B (eds) Proceedings of the 21st European conference on artificial intelligence, 18–22 August 2014, Prague, pp 627–632
Mille A (2006) From case-based reasoning to traces-based reasoning. Ann Rev Control 30(2):223–232
Minor M, Montani S, Recio-Garcia JA (2014) Information systems, vol 40 (Special Section on Process-Oriented Case-based Reasoning)
Minsky M (1975) A framework for representing knowledge
Mitchell M (1993) Analogy-making as perception: a computer model. MIT Press, Cambridge
Mitchell M (2001) Analogy-making as a complex adaptive system. In: Segel L, Cohen I (eds) Design principles for the immune system and other distributed autonomous systems. Oxford University Press, Oxford
Peirce CS (1955) Philosophical writings. Selected and edited, with an introduction by J. Buchler. Dover Publication, New York
Perfilieva I, Dubois D, Prade H, Esteva F, Godo L, Hod'áková P (2012) Interpolation of fuzzy data: analytical approach and overview. Fuzzy Sets Syst 192:134–158
Prade H, Richard G (2013) From analogical proportion to logical proportions. Logica Universalis 7(4):441–505
Prade H, Richard G (eds) (2014a) Computational approaches to analogical reasoning - current trends. Springer, Berlin
Prade H, Richard G (2014b) Homogenous and heterogeneous logical proportions. IfCoLog J Log Appl 1(1):1–51
Raccah PY (1996) Topoi et Gestion des Connaissances. Masson
Raven J (2000) The Raven's progressive matrices: change and stability over culture and time. Cogn Psychol 41(1):1–48
Recio-Garcia JA (2008) jCOLIBRI: a multi-level platform for building and generating CBR systems. PhD thesis, University of Madrid
Richter MM (1998) Introduction. In: Lenz M, Bartsch-Spörl B, Burkhard H-D, Wess S (eds) Case-based reasoning technologies. From foundations to applications. LNCS, vol 1400. Springer, Berlin, pp 1–15
Richter MM, Weber RO (2013) Case-based reasoning. Springer, Berlin
Riesbeck CK, Schank RC (1989) Inside case-based reasoning. Lawrence Erlbaum Associates
Rougegrez S (1994) Similarity evaluation between observed behaviours for the prediction of processes. In: Wess S, Althoff K-D, Richter MM (eds) Topics in case-based reasoning - First European workshop (EWCBR'93), Kaiserslautern. LNCS, vol 837. Springer, Berlin, pp 155–166
Ruspini EH (1991) On the semantics of fuzzy logic. Int J Approx Reason 5(1):45–88
Russell SJ (1989) The use of knowledge in analogy and induction. Pitman, UK
Schank R (1982) Dynamic memory: a theory of reminding and learning in computer and people. Cambridge University Press, Cambridge
Schockaert S, Prade H (2011) Solving conflicts in information merging by a flexible interpretation of atomic propositions. Artif Intell 175:1815–1855
Schockaert S, Prade H (2013) Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces. Artif Intell 202:86–131
Smyth B, Cunningham P (2017) Running with cases: a CBR approach to running your best marathon. In: Aha DW, Lieber J (eds) Case-based reasoning research and development, Proceedings of ICCBR-2017. Springer, Berlin, pp 360–374
Smyth B, Keane MT (1995) Retrieval and adaptation in déjà vu, a case-based reasoning system for software design. In: Adaptation of knowledge for reuse: a 1995 AAAI fall symposium, Cambridge, Massachusetts. AAAI Press, pp 228–240

Smyth B, Keane MT (1996) Using adaptation knowledge to retrieve and adapt design cases. Knowl-Based Syst 9(2):127–135
Sowa JF, Majumdar AK (2003) Analogical reasoning. In: Proceedings of the international conference on conceptual structures. LNAI, vol 2746. Springer, Dresden, pp 16–36
Spalazzi L (2001) A survey on case-based planning. Artif Intell Rev 16(1):3–36
Stahl A (2005) Learning similarity measures: a formal view based on a generalized CBR model. In: Case-based reasoning research and development, Proceedings of ICCBR-2005. Springer, Berlin, pp 507–521
Stahl A, Roth-Berghofer T (2008) Rapid prototyping of CBR applications with the open source tool myCBR. In: Advances in case-based reasoning, 9th European conference, ICCBR-2008, Trier, Germany. Proceedings. LNAI, vol 5239. Springer, Berlin, pp 615–629
Stefik M (1995) Introduction to knowledge systems. Morgan Kaufmann Publishers Inc., San Francisco
Stroppa N, Yvon F (2005a) An analogical learner for morphological analysis. In: Online proceedings of the 9th conference on computer natural language learning (CoNLL-2005), pp 120–127
Stroppa N, Yvon F (2005b) Analogical learning and formal proportions: definitions and methodological issues. Technical report D004, ENST-Paris
Stroppa N, Yvon F (2005c) Analogical learning and formal proportions: definitions and methodological issues. Technical report, ENST-2005-D004 June 2005. http://www.tsi.enst.fr/publications/enst/techreport-2007-6830.pdf
Stroppa N, Yvon F (2006) Du quatrième de proportion comme principe inductif: une proposition et son application à l'apprentissage de la morphologie. Traitement Automatique des Langues 47(2):33–59
Syrovatka J (2000) Analogy and understanding. Theoria. Revista de Teoria, Historia y Fundamentos de la Ciencia 15(3):435–450
Thagard P, Holyoak KJ, Nelson G, Gochfeld D (1990) Analog retrieval by constraint satisfaction. Artif Intell 46(3):259–310
Tversky A (1977) Features of similarity. Psychol Rev 84:327–352
Van Dormael J (1990) The emergence of analogy. Analogical reasoning as a constraint satisfaction process. Philosophica 46:157–177
Veloso MM (1994) Planning and learning by analogical reasoning. LNAI, vol 886. Springer, Berlin
Wang P (2009) Analogy in a general-purpose reasoning system. Cogn Syst Res 10(3):286–296
Weber RO, Ashley KD, Brüninghaus S (2005) Textual case-based reasoning. Knowl Eng Rev 20(3):255–260
Weitzenfeld JS (1984) Valid reasoning by analogy. Philos Sci 51(1):137–149
Weller S, Schmid U (2007) Solving proportional analogies by E-generalization. In: Freksa C, Kohlhase M, Schill K (eds) KI 2006: Advances in artificial intelligence. LNCS, vol 4314. Springer, Berlin, pp 64–75
Wilke W, Vollrath I, Althoff K-D, Bergmann R (1996) A framework for learning adaptation knowledge based on knowledge light approaches. In: Adaptation in case based reasoning: a workshop at ECAI 1996, Budapest
Winston PH (1980) Learning and reasoning by analogy. Commun ACM 23:689–703
Woolford M, Watson I (2017) SCOUT: a case-based reasoning agent for playing race for the galaxy. In: Aha DW, Lieber J (eds) Case-based reasoning research and development, Proceedings of ICCBR-2017. Springer, Berlin, pp 390–402
Yvon F, Stroppa N, Delhay A, Miclet L (2004) Solving analogical equations on words. Technical report, Ecole Nationale Supérieure des Télécommunications
Zadeh LA (1979) A theory of approximate reasoning. In: Hayes J, Mitchie D, Mikulich L (eds) Machine intelligence, vol 9. Elsevier, Amsterdam, pp 149–194
Zarka R, Cordier A, Egyed-Zsigmond E, Mille A (2011) Rule-based impact propagation for trace replay. In: Ram A, Wiratunga N (eds) International case-based reasoning conference (ICCBR 2011), Greenwich, London, United Kingdom. Springer, Berlin, pp 482–495

Statistical Computational Learning

Antoine Cornuejols, Frédéric Koriche and Richard Nock

Abstract Statistical computational learning is the branch of Machine Learning that defines and analyzes the performance of learning algorithms using two metrics: sample complexity and runtime complexity. This chapter is a short introduction to this important area of research, geared toward the reader interested in developing learning algorithms for AI models. We first provide the formal background about statistical learning problems, captured by three basic ingredients: tasks, models and loss functions. We next examine the PAC learning framework and its generalizations, used to capture the concepts of statistical learnability and computational (or efficient) learnability. Based on this framework, the conditions of statistical learnability are investigated through the properties of uniform convergence and algorithmic stability. We also survey several theoretical results and algorithms in the topics of concept learning and convex learning, which take a central place in statistical computational learning. We then conclude this survey with some trends and open questions in learning AI models, by mainly focusing on sparse models, probabilistic models, preference models and deep neural models.

1 Introduction

The cognitive ability of *learning* has long fascinated philosophers, psychologists, statisticians, computer scientists and, of course, the parents of young children. In Computer Science, Turing already speculated in Turing (1950) that learning would be used to build machines that think. Since then, the field of Machine Learn-

A. Cornuejols (✉)
AgroParisTech, Paris, France
e-mail: antoine.cornuejols@agroparistech.fr

F. Koriche
CRIL-CNRS and Université d'Artois, Lens, France
e-mail: frederic.koriche@cril.univ-artois.fr

R. Nock
NICTA, ANU College of Engineering and Computer Science, Canberra, Australia
e-mail: richard.nock@nicta.com.au

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_11

ing has flourished with the development of various learning frameworks, theories, algorithms, and practical applications. In fact, we are nowadays surrounded by learning-based computer technologies: our smartphones learn to recognize voice commands, our digital cameras learn to identify faces, antispam softwares learn to filter our email messages, and recommender systems learn our preferences about daily consumer objects. Learning algorithms are also widely used in scientific applications such as astronomy, bioinformatics, medicine, economy and robotics.

Broadly speaking, the main concern of Machine Learning is to study *how computer algorithms can improve automatically through experience*. Virtually all machine learning activities involve a *task* we wish to solve, a set of candidate prediction *models* for solving this task, and an *objective function* for measuring the performance of a model at solving the task. In this setting, the term "experience" refers to the information provided to the learning algorithm for assessing the quality of candidate models, and ultimately, choosing the right one.

To illustrate these aspects with a concrete example, consider the common task of classifying incoming email messages as either SPAM or non-SPAM. As electronic messages usually contain a text in natural language, possibly coupled with graphical elements and URL links, the problem of recognizing whether an incoming email is a spam, or not, is far from easy. So, in order to facilitate the learning process, each electronic message is associated with a set (or vector) of *features*, which capture informative properties of the message, such as its size, its text-to-image ratio, the presence of some domain names in the header, or the occurrence of certain regular expressions in the content. Based on this feature representation, the task of spam filtering is essentially to map email messages, described by their features, to the set of labels {SPAM, non−SPAM}. Any such mapping is called *hypothesis* or *model*, and the set of candidate models available to the learner is called the *hypothesis class*. Since spam filtering is a binary classification task, various models can be used, such as decision trees, separating hyperplanes, or Bayesian classifiers. Finally, we need to assess the performance of the chosen model at filtering incoming messages. Here, a natural objective function is the "zero-one" loss function, which simply counts the number of mistakes made by the model in labeling messages.

Based on the three ingredients, tasks, models and objective functions, the goal of a learning algorithm is essentially to find, in its hypothesis class, a model that optimizes some given objective function for the task at hand. To achieve this goal, the learner has usually access to a *training set*, that is, a sequence of data instances upon which the quality of candidate models can be measured. In spam filtering, the training set is a pool of email messages, each described by its features, and labeled by SPAM or by non-SPAM. Importantly, this training set captures only a small fragment of emails we are expected to receive. So, the learning problem is not to find a model that makes few mistakes on the training set, but to extrapolate from observed instances a model that accurately classifies *new*, incoming messages. In a nutshell, the key characteristic of learning algorithms lies in their ability to *generalize*, that is, to predict from observed data, the outcome of future data.

This chapter focuses on *statistical computational learning*, the branch of Machine Learning that lies at the intersection of statistical modeling and computational

learning theory. In this setting, the generalization ability of learning algorithms is defined and analyzed through two key metrics: sample complexity and runtime complexity. Because statistical computational learning has long been recognized as the mainstream theoretical framework for analyzing the performance of learning algorithms, a detailed survey of this research field and its applications would require a whole book! In fact, there are already excellent printed works on statistical and computational learning, targeted to various audiences (Natarajan 1991; Kearns and Vazirani 1994; Anthony and Biggs 1997; Vapnik 1998; Engel and Broeck 2001; Hastie et al. 2009; DasGupta 2011; Kulkarni and Harman 2011; Webb and Copsey 2011; Devroye et al. 2013; James et al. 2013; Vapnik 2013; Sugiyama 2015). Furthermore, many introductory books in Machine Learning are devoting a significant part to statistical and/or computational learning theory (Mitchell 1997; Bishop 2006; Alpaydin 2009; Flach 2012; Mohri et al. 2012; Murphy 2012; Shalev-Shwartz and Ben-David 2014; Theodoridis 2015). So, this chapter is an elementary introduction to statistical computational learning, geared toward readers who have familiar with AI models, such as logical representations, geometric descriptions, and graphical models.

We introduce in Sect. 2 the formal background about statistical learning problems. The central notions of *statistical learnability* and *computational learnability* are defined in Sect. 3. The related optimization principles and conditions of learnability are examined in Sect. 4. With these theoretical notions in hand, the important topics of *concept learning* and *convex learning* are surveyed in Sects. 5 and 6, respectively. Finally, we conclude this chapter by discussing about some trends and open questions in statistical learning with sparse models, probabilistic models, preference models, and neural networks.

Notation. For the sake of clarity we shall use as much as possible the standard notation in Machine Learning. Scalars and vectors are denoted by lowercase letters. Sets, matrices, sequences, and distributions are denoted by uppercase letters. We use boldface letters for vectors and matrices. For an integer n, we use $[n]$ as an abbreviation of $\{1, \ldots, n\}$. Given a sequence S of m vectors $(\boldsymbol{x}_1, \ldots, \mathbf{x}_m)$, we use $x_{i,j}$ to denote the jth element of $\boldsymbol{x}_i$. The inner product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is denoted $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, and for any $p \in [1, \infty]$, the ℓ_p norm of $\boldsymbol{x}$ is denoted $\|\boldsymbol{x}\|_p$. In other words,

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \quad \text{and in particular} \quad \|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| \text{ and } \|\boldsymbol{x}\|_\infty = \max_{i \in [n]} |x_i|$$

We omit the subscript from the standard ℓ_2 (Euclidean) norm when it is clear from the context. The number of nonzero coordinates in $\boldsymbol{x}$, often called ℓ_0 *pseudo-norm* of $\boldsymbol{x}$, is denoted $\|\boldsymbol{x}\|_0$. For a set of scalars $X \subseteq \mathbb{R}$, the greatest lower bound of X and the least upper bound of X are denoted inf X and sup X, respectively. Finally, we shall assume throughout this chapter that the sample space of any probability distribution is equipped with an implicit σ-algebra upon which the distribution is

defined. Given a probability distribution $\mathscr{D}$ over a sample space $\mathscr{X} \subseteq \mathbb{R}^n$, we use $\boldsymbol{x} \sim \mathscr{X}$ to indicate that $\boldsymbol{x}$ is sampled according to $\mathscr{D}$. Probabilities and expectations over $\mathscr{D}$ are denoted $\mathbb{P}$ and $\mathbb{E}$, respectively.

2 Statistical Learning Problems

In order to provide a clear definition of "statistical computational learning", we need to capture in a formal way the three aforementioned ingredients: *tasks*, *models*, and *objective functions*. We start this section by discussing about these notions, and then describe the statistical learning framework upon which the rest of chapter is built.

2.1 Tasks

As Machine Learning can be considered as a data-driven approach to problem solving, the notion of "task" is described through its data instances. Specifically, an *instance space* is a (possibly infinite) subset $\mathscr{Z}$ of $\mathbb{R}^d$. Each coordinate $i \in [d]$ represents a distinct feature, and each instance $\boldsymbol{z} \in \mathscr{Z}$ is a vector of d feature values.

Learning algorithms can solve a wide variety of tasks and, for this reason, it may be useful to separate them into categories. A first separation, commonly advocated in the Machine Learning literature, is to distinguish *supervised* learning tasks from *unsupervised* ones.

Basically, supervised learning tasks capture applications for which we need to predict the dependence of an outcome $\boldsymbol{y} \in \mathscr{Y}$ on an observed information $\boldsymbol{x} \in \mathscr{X}$. Here, $\mathscr{Z}$ is the Cartesian product $\mathscr{X} \times \mathscr{Y}$ of a *domain set* $\mathscr{X}$ and a *target set* $\mathscr{Y}$. Pairs of the form $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ are often referred to as *labeled instances* or *examples*. The dimensions of $\mathscr{X}$ and $\mathscr{Y}$ are denoted n and p, respectively. A supervised learning task is *uni-dimensional* if $p = 1$, and *multi-dimensional* if $p > 1$. Some of the most common supervised learning tasks include the following:

- *Classification:* $\mathscr{Y}$ is a finite subset of $\mathbb{Z}$, encoding a collection of labels. The spam filtering task mentioned in the introduction of this chapter is an example of *binary* classification problem, where $\mathscr{Y}$ is usually defined by $\{0, 1\}$ or $\{-1, +1\}$. Classification problems with more than two labels are often referred to as *multi-class* or *multi-nominal* classification tasks.
- *Regression:* $\mathscr{Y}$ is a (typically bounded) subset of $\mathbb{R}$, capturing the domain of some real-valued variable. A common example of regression task is to estimate the revenue of a company, using historical accounting data.
- *Multi-label classification:* $\mathscr{Y}$ is a subset of $\{0, 1\}^p$ or $\{-1, +1\}^p$ for $p > 1$. Here, the learner as access to p distinct labels, and the goal is to map each input vector to a *subset* of these labels. A common example of multi-label classification in document analysis is to "tag" incoming news according to their most relevant topics (e.g. sports, entertainment, politics, science).

- *Multi-variate regression:* By analogy with multi-label classification, $\mathcal{Y}$ is a subset of $\mathbb{R}^p$ for $p > 1$. A well-studied example in ecological modeling is to simultaneously predict multiple target variables describing the condition or quality or plant species.
- *Structured prediction:* This setting covers multi-dimensional prediction tasks in which target variables are organized into some *structure*, such as a permutation, a tree, or a bipartite graph. One example is *parsing*, the task of mapping a natural language sentence into a tree that predicts its grammatical structure. Another example is *label ranking*, the task of mapping a feature vector (e.g. a user profile) into a permutation of items (e.g. movies).

In contrast with supervised tasks, there is *no* target set $\mathcal{Y}$ in unsupervised tasks. Here, $\mathcal{Z}$ is a set $\mathcal{X}$ of unlabeled instances. The overall goal of unsupervised learning is to extract from observed data some regularities or patterns which are likely to be found in future data. Two of the most popular unsupervised learning tasks are:

- *k-Means clustering*: The goal is to partition the instance space $\mathcal{X}$ into k clusters, each identified by a centroid $\boldsymbol{c}$ in $\mathcal{Z}$. Any incoming instance $\boldsymbol{x}$ is mapped to the centroid $\boldsymbol{c}$ that minimizes the squared distance $\|\boldsymbol{x} - \boldsymbol{c}\|^2$.
- *Density estimation*: Here, the task is to find a probability distribution over $\mathcal{X}$ that estimates the likeliness of incoming instances. This distribution can be viewed as a maximum likelihood estimator of the data instances supplied to the learner.

2.2 Models

In order to solve a given task, the learner has access to a set of candidate hypotheses, called the *hypothesis class*, and denoted $\mathcal{H}$. From a general viewpoint, any hypothesis in $\mathcal{H}$ can be viewed as a mapping $h : \mathcal{X} \to \mathcal{Y}^\dagger$, where $\mathcal{X}$ is the set of input observations, and $\mathcal{Y}^\dagger$ is a set of decisions. By analogy with learning tasks, hypotheses can be separated into *discriminative* models and *descriptive* models. Basically, discriminative models are dedicated to supervised learning tasks. Here, the decision set $\mathcal{Y}^\dagger$ coincides with the target set $\mathcal{Y}$, and hence, any class $\mathcal{H}$ of discriminative models is a subset of the function space $\mathcal{Y}^\mathcal{X}$. By contrast, descriptive models are used to explain observations by extracting regularities or patterns. For those models, the choice of $\mathcal{Y}^\dagger$ depends on how observations are explained. An important subclass of descriptive models is the family of *generative* models, where $\mathcal{Y}^\dagger = [0, 1]$, and $\mathcal{H}$ is a set of probability distributions over $\mathcal{X}$. While generative models are mainly devoted to unsupervised learning tasks, they may be applied to supervised learning problems by first extracting from examples a probabilistic model that estimates the underlying distribution, and then using this model for solving various tasks.

Some of the most common families of hypothesis classes which have been examined in Machine Learning include:

- *Logical models:* $\mathcal{H}$ is typically a set of functions of the form $h : \{0, 1\}^n \to \{0, 1\}$. In other words, the domain of a logical model is a set of Boolean features, and

its range is a Boolean variable. Simple logical models are constructed using a single logical operator; they include *monomials* (conjunctions of literals), *clauses* (disjunctions of literals), and *XOR clauses* (exclusive-or of literals). More complex functions are built using at least two logical operators. They include, among others, *DNF formulas* (disjunctions of monomials), *decision trees* (disjunctions of monomials organized into a tree), and *decision lists* (sets of monomials organized into a preference list). The main learning task considered for logical models is binary classification; this problem had long been considered as a central topic in computational learning theory (Natarajan 1991; Anthony 2010; Kearns et al. 1994a; Kearns and Vazirani 1994). Besides Boolean functions, logical models investigated in Machine Learning include *relational models*, defined over structured domain spaces (Getoor and Taskar 2007; De Raedt 2008).

- *Geometric models:* $\mathscr{H}$ is a set of geometric objects or functions over $\mathscr{X} \subseteq \mathbb{R}^n$. Arguably, the simplest hypothesis class in the family of geometric models is the class of *separating hyperplanes*, also known as *linear threshold functions*, which has been studied since the very start of Machine Learning (Rosenblatt 1958). Here, each hypothesis $h : \mathbb{R}^n \to \{-1, +1\}$ is represented by a pair $(\boldsymbol{w}, b)$, where $\boldsymbol{w} \in \mathbb{R}^n$ is a weight vector, and $b \in \mathbb{R}$ is a threshold value. The label assigned to any input object $\boldsymbol{x} \in \mathbb{R}^n$ is given by

$$h(\boldsymbol{x}) = \begin{cases} +1 & \text{if } \langle \boldsymbol{w}, \boldsymbol{x} \rangle > b \\ -1 & \text{otherwise.} \end{cases} \tag{1}$$

  For *zero-threshold* or *homogeneous* linear functions, h is simply described by its weight vector $\boldsymbol{w}$, and defined by $h(\boldsymbol{x}) = \text{sign}\, \langle \boldsymbol{w}, \boldsymbol{x} \rangle$. More complex geometric objects may be defined using a weight vector $\boldsymbol{w} \in \mathbb{R}^p$, together with a *feature expansion* mapping: $\boldsymbol{\phi} : \mathscr{X} \to \mathscr{X}^\dagger$, where $\mathscr{X}^\dagger$ is an Euclidean or Hilbert space. In this general setting,

$$h(\boldsymbol{x}) = \begin{cases} +1 & \text{if } \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle > b \\ -1 & \text{otherwise.} \end{cases} \tag{2}$$

  Linear functions and their feature expansions can be extended, in a natural way, to regression tasks, multi-nominal classification tasks, and even multi-dimensional prediction tasks. Besides hyperplanes, *manifolds* and *distances* take also an important place in geometric learning. Namely, manifolds are used for extracting a low-dimensional structure from a high-dimensional domain (Ma and Fu 2011), and distance functions are commonly used in classification, regression, and clustering (Aggarwal and Reddy 2013).
- *Probabilistic models:* $\mathscr{H}$ is a set of probability distributions over an instance space $\mathscr{Z} \subseteq \mathbb{R}^d$. Of particular importance are probabilistic *graphical* models, which encode high-dimensional probability distributions in a compact and intuitive way (Koller and Friedman 2009; Murphy 2012). Here, each hypothesis is represented by a pair $(G, \boldsymbol{\theta})$, where G is a graph over $[d]$ nodes, and $\boldsymbol{\theta}$ is a vector of parameters

which together determine a probability distribution over $\mathscr{Z}$. In *directed* graphical models, also known as *Bayesian networks* (Pearl 1988; Darwiche 2009), G is a directed acyclic graph, and $\boldsymbol{\theta}$ is a set of conditional probability tables associated with the nodes of G. In *undirected* graphical models (Wainwright and Jordan 2008), such as *factor graphs* and *Markov networks*, G is an undirected graph and $\boldsymbol{\theta}$ is a vector of energy functions defined on the edges (for factor graphs) or the cliques (for Markov networks) of the graph. Probabilistic graphical models can be applied to a wide variety of learning tasks, including density estimation and structured prediction.

- *Preference models:* $\mathscr{H}$ is a set of functions from $\mathscr{X}$ to $\mathscr{Y}$, where $\mathscr{X}$ is a set of objects, possibly coupled with user profiles, and $\mathscr{Y}$ is a partial or total ordering over some reference set $\mathscr{I}$. In preference learning (Fürnkranz and Hüllermeier 2010), the family of models may be organized into different subclasses, depending on the type of reference set $\mathscr{I}$, and the type of preference relation $\mathscr{Y}$. In *object ranking* (Cohen et al. 1999), $\mathscr{I}$ is a set of objects in $\mathscr{X}$, while in *label ranking* (Vembu and Gärtner 2010), $\mathscr{I}$ is a set of labels associated with objects in $\mathscr{X}$. Orthogonally, *total rankings* are permutations over $\mathscr{I}$, while *partial rankings* are pre-orderings on $\mathscr{I}$. For example, in the task of *top-k object ranking* commonly used in information retrieval, the goal is to find a total ordering over the k best objects in $\mathscr{X}$, while others objects are considered indifferent. Similarly, the task of *bipartite ranking* is to separate objects in $\mathscr{X}$ in two categories: the most preferred objects, and the less preferred ones (Clémençon and Vayatis 2007). Common preference models advocated in the Machine Learning literature include the *Placket-Luce model* (Plackett 1975), the *Mallows model* (Mallows 1957), and their extensions (Fligner and Verducci 1986; Lebanon and Lafferty 2002; Meila and Chen 2010; Lu and Boutilier 2014; Zhao et al. 2016).
- *Neural models:* $\mathscr{H}$ is a class of artificial neural networks, inspired from the structure of neural networks in the brain. A *feedforward neural netwok* is defined by a labeled and weighted directed acyclic graph. Each node in the graph a simple model of neuron, labeled by an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Common scalar functions include the sign function $\sigma(a) = \text{sign}(a)$, the threshold function given by (1), and the sigmoid function $\sigma(a) = {}^{1}/_{1+\exp(-a)}$. Each edge in the graph, linking the output of some neuron to the input of another neuron, is associated with a weight that reflects the strength of the signal joining both neurons. The input of a neuron is obtained by taking the weighted sum of the outputs of its incident neurons. It is often assumed that neurons are organized in *layers*. Namely, the set of nodes in the graph is partitioned into $d+1$ subsets $\{V_0, V_1, \ldots, V_d\}$, where V_0 is the input layer, V_d is the output layer, and $\{V_1, \ldots, V_{d-1}\}$ are the *hidden* layers. The *depth* and *width* of the network are given by d and $\max_i |V_i|$, respectively. Based on this layer structure, the output of the layer V_t is given by:

$$\boldsymbol{x}_t = \boldsymbol{\sigma}\left(\boldsymbol{W}_t^\top \boldsymbol{x}_{t-1} + \mathbf{b}_t\right)$$

where $\boldsymbol{x}_{t-1}$ is the input of the tth layer, $\boldsymbol{\sigma}$ is the (possibly rectified) activation function for this layer, $\boldsymbol{W}_t$ is the weight matrix capturing weighted edges between

the layers V_{t-1} and V_t, and $\mathbf{b}_t$ is a bias vector. The family of hypothesis classes of artificial neural networks is very expressive: notably, Boolean functions of polynomial circuit complexity can be represented by neural networks of polynomial size (Parberry 1994). For this reason, neural networks have been a subject of extensive research in statistical computational learning (Anthony and Barlett 1999; Anthony 2001; Du and Swamy 2013). *Deep networks*, characterized by more than two layers, have recently shown very impressive practical performance on a wide variety of learning tasks (Goodfellow et al. 2016).

2.3 Objective Functions

In Machine Learning, the connexion between tasks and models is established through *objective* or *loss* functions. Formally, a loss function is a map ℓ from $\mathscr{H} \times \mathscr{Z}$ to $\mathbb{R}$ that penalizes a model $h \in \mathscr{H}$ picked by the learner when it observes the instance $z \in \mathscr{Z}$. In other words, $\ell(h, z)$ is the cost incurred by h on z. Some of the most common loss functions include the following:

- *Zero-one loss:* Applied to binary classification, this function measures whether a binary hypothesis is misclassifying a labeled instance. Formally, $\ell(h, (\boldsymbol{x}, y)) = 1$ if $h(\boldsymbol{x}) \neq y$, and $\ell(h, (\boldsymbol{x}, y)) = 0$ otherwise.
- *Quadratic loss:* This function, commonly used in regression tasks, measures the squared distance between a predicted value and the target value. Namely, $\ell(h, (\boldsymbol{x}, y)) = (h(\boldsymbol{x}) - y)^2$.
- *Hinge loss:* This function is a convex surrogate of the zero-one loss in linear classification. For a zero-threshold separating hyperplane h, represented by its weight vector $\boldsymbol{w}$, the hinge loss of h on some example $(\boldsymbol{x}, y)$ is given by

$$\ell(h, (\boldsymbol{x}, y)) = \max\{0, 1 - y \langle \boldsymbol{w}, \boldsymbol{x} \rangle\}$$

- *Log-loss:* Used in density estimation, this function measures the negative log-likelihood of a probabilistic model $h : \mathscr{X} \to [0, 1]$ given an incoming instance $\boldsymbol{x}$. Formally, $\ell(h, \boldsymbol{x}) = -\ln[h(\boldsymbol{x})]$.
- *Conditional log-loss:* As a direct extension of the log-loss, this function is often used in structured prediction. Given a conditional probabilistic model h that maps each input object $\boldsymbol{x}$ to a probability distribution $h(\cdot \mid \boldsymbol{x})$ over $\mathscr{Y}$, the conditional log-loss of h with respect to an example $(\boldsymbol{x}, \boldsymbol{y})$ is $\ell(h, (\boldsymbol{x}, \boldsymbol{y})) = -\ln[h(\boldsymbol{y} \mid \boldsymbol{x})]$.

2.4 The Framework

The three components - tasks, models, and objective functions - are common to many machine learning frameworks. The specificity of statistical learning lies in a

fourth component that captures how data instances are generated. Here, it is assumed that instances are independently and identically distributed (i.i.d.) according to some probability distribution $\mathscr{D}$ over $\mathscr{Z}$. Importantly, $\mathscr{D}$ is an *arbitrary* but *hidden* distribution: the incoming data can be generated according to any possible distribution over the sample space $\mathscr{Z}$, and the learning algorithm has no prior information about this distribution. Instead, the learner has access to $\mathscr{D}$ through a procedure $\text{EX}(\mathscr{D})$, that runs in unit time, and on each call returns an instance $\boldsymbol{z} \in \mathscr{Z}$ drawn randomly and independently according to $\mathscr{D}$. This procedure, referred to as *example oracle*, is used to generate a *training set*, that is, a sequence $S = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$ of instances which are i.i.d. according to $\mathscr{D}$.

Recall that in supervised learning, the instance space $\mathscr{Z}$ is the Cartesian product of a domain set $\mathscr{X}$ and a target set $\mathscr{Y}$. So, in this setting, $\mathscr{D}$ is a joint distribution over $\mathscr{X} \times \mathscr{Y}$. Equivalently, this distribution can be viewed as the conditional probability of observing the labeled object $(\boldsymbol{x}, \boldsymbol{y})$ given an unlabeled object $\boldsymbol{x}$. For instance, in the spam filtering task, $\mathscr{D}$ specifies the probability of encountering a spam message, given a feature description of this message. In unsupervised learning, the learner has only access to unlabeled observations, and its goal is essentially to predict the data-generation model $\mathscr{D}$ using a limited number of calls to $\text{EX}(\mathscr{D})$.

With these components in hand, we are now in position to describe the learning framework upon which the remaining sections are built. Formally, a *statistical learning problem* is defined as follows:

Given:

- A task described by its instance space $\mathscr{Z}$
- A hypothesis class $\mathscr{H}$ for $\mathscr{Z}$
- A loss function $\ell : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$
- A distribution $\mathscr{D}$ accessible through the example oracle $\text{EX}(\mathscr{D})$

Find a hypothesis $h \in \mathscr{H}$ that minimizes

$$L_{\mathscr{D}}(h) = \mathbb{E}_{\mathbf{z} \sim \mathscr{D}}[\ell(h, \mathbf{z})] \tag{3}$$

The objective function $L_{\mathscr{D}} : \mathscr{H} \to \mathbb{R}$ in (3) is called the *true risk*, or *risk* for short. It measures the expected loss of a hypothesis $h \in \mathscr{H}$ with respect to the probability distribution $\mathscr{D}$ over $\mathscr{Z}$. Since the learner has only access to a sample of data instances picked randomly according to $\mathscr{D}$, we define the *empirical risk* of a hypothesis h with respect to a training set $S = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$ as:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, \boldsymbol{z}_i) \tag{4}$$

The main difficulty of statistical learning is to estimate the unknown true risk according to the known empirical risk. The intimate relation between $L_{\mathscr{D}}$ and L_S will be discussed in Sect. 4.

In practice, how do we analyze the performance of learning algorithms? There is no simple answer to this question, since the instance space $\mathscr{Z}$ of most learning problems is immense, or infinite. In practice, we typically have a limited "dataset" for the task we wish to solve. If the dataset is already separated into a training sample S and a test sample S', then we just have to train our algorithm on S, and to measure the empirical risk of its output model on S'. Yet, if the dataset does not include a predefined test sample, we need to resort on a statistical validation technique for assessing the performance of the learner. The following *k-fold cross-validation* procedure is often applied: randomly partition the dataset S in k parts or "folds" $S_1, \ldots, S_k$, pick one fold S_j for testing, train the algorithm on the complementary set $S \backslash S_j$, and evaluate the resulting hypothesis h_j on the test fold S_j. This process is repeated k times, until each fold has been picked for testing once. The *cross-validation risk* of the k hypotheses $(h_1, \ldots, h_k)$ returned by the algorithm is given by

$$L_{\mathrm{CV}}(h_1, \ldots, h_k) = \frac{1}{k} \sum_{j=1}^{k} L_{S_j}(h_j)$$

3 Complexity Measures

As mentioned above, a statistical learning problem involves a task, described by its instance space $\mathscr{Z}$, a hypothesis class $\mathscr{H}$ for $\mathscr{Z}$, a loss function ℓ, and a hidden distribution $\mathscr{D}$ over $\mathscr{Z}$ which is only accessible through an example oracle $\mathrm{EX}(\mathscr{D})$. The goal is to to find a model with good generalization performance, that is, a hypothesis $h \in \mathscr{H}$ for which the true risk $L_{\mathscr{D}}(h)$ is as small as possible. Based on this formulation, there are two main sources of complexity in the computational approach to statistical learning. The first, *sample complexity*, measures the inherent difficulty of generalizing from examples: it is the number of calls to $\mathrm{EX}(\mathscr{D})$ which are required to find a good hypothesis. The second, *runtime complexity*, measures the amount of computational steps required to find such a model. This section explores in more detail both sources of complexity which are related to the concept of *learnability*.

3.1 Sample Complexity

As indicated above, sample complexity is the amount of information learning requires. to find a "good" hypothesis. In order to capture this metric in a more rigorous way, we need a formal model of *learnability*, that explains the ability of algorithms to predict with respect to a hypothesis class, given access to training samples.

We begin with a conceptually simple notion of learnability, introduced by (Valiant 1984) and thoughtfully detailed in various textbooks about computational learning theory (Natarajan 1991; Kearns and Vazirani 1994; Anthony and Biggs 1997). In *Probably Approximately Correct* (PAC) learning, we are concerned with supervised learning tasks, where the instance space $\mathscr{Z}$ is the product of a domain set $\mathscr{X}$ and a target set $\mathscr{Y}$. Originally, the PAC learning framework was defined for binary classification tasks, but we can easily extend the framework to other discriminative tasks, using an appropriate loss function. The key assumption in PAC learning, often called *realizability condition*, is to consider that the hypothesis class $\mathscr{H}$ includes at least one model, say h^*, which correctly solves the task at hand. In other words, the outcome $\boldsymbol{y}$ of any input object $\boldsymbol{x}$ is given by $\boldsymbol{y} = h^*(\boldsymbol{x})$. The realizability assumption can be captured using a restricted example oracle $\text{EX}(h^*, \mathscr{D})$ which returns, on each call, a labeled example $(\boldsymbol{x}, h^*(\boldsymbol{x}))$, where $\boldsymbol{x}$ is drawn at random according to a hidden distribution $\mathscr{D}$ over $\mathscr{X}$.

A PAC learning algorithm takes as input a *confidence* parameter and a *accuracy* parameter, denoted δ and ε, respectively. These parameters are use to control two types failures which are inherent to learn from samples drawn at random according to an unknown distribution $\mathscr{D}$. The confidence parameter is necessary since there is always a chance that the training set picked by the learner is not representative of $\mathscr{D}$. For example, the learner might be very unlucky by picking a sample consisting of repeated draws of the same object in $\mathscr{X}$, despite the fact that the distribution is spread evenly over all the domain set $\mathscr{X}$. The accuracy parameter is also necessary since, even with a training set that is representative of $\mathscr{D}$, some objects in $\mathscr{X}$ may have a very low probability under $\mathscr{D}$, and hence, the learning algorithm will not see the target function's behavior on those objects. So, the best we can hope is that the likeliness of both types of failure can be made arbitrary small, at the cost of increasing the size of the training set.

Definition 1 (*PAC Learning*) Let $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$ be an instance space, $\mathscr{H}$ be a hypothesis class over $\mathscr{Z}$, and $\ell : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$ be a loss function. Then, $\mathscr{H}$ is PAC *learnable* with respect to ℓ if there exist an algorithm LEARN with the following property: for any hypothesis $h^* \in \mathscr{H}$, any distribution $\mathscr{D}$ over $\mathscr{X}$, and any pair $(\delta, \varepsilon) \in (0, 1)^2$, if LEARN is given inputs δ and ε, and access to $\text{EX}(h^*, \mathscr{D})$, then LEARN returns a hypothesis $h \in \mathscr{H}$ that satisfies $L_{\mathscr{D}}(h) \leq \varepsilon$ with probability $1 - \delta$.

In essence, PAC learning is a *distribution-free* model of statistical learning: for any possible distribution $\mathscr{D}$ over the domain set $\mathscr{X}$, the algorithm LEARN must be "approximately correct" with high probability. The sample complexity of LEARN is the number of calls to the example oracle $\text{EX}(h^*, \mathscr{D})$, that is, the number m of training examples required to output with confidence $1 - \delta$, an ε-accurate hypothesis. If m is polynomial in $1/\delta$ and $1/\varepsilon$, then the hypothesis class $\mathscr{H}$ is called PAC *learnable with polynomial sample complexity*.

Though conceptually elegant, the PAC learning framework relies on some realizability condition which is unrealistic in practice. Indeed in many, if not most, statistical learning problems, there is no well-defined target model that perfectly labels

incoming instances. For example, if we choose the class $\mathscr{H}$ of decision trees for learning to filter spam messages, we are not guaranteed that $\mathscr{H}$ will always include a decision tree that accurately filters any possible electronic message. This realizability assumption is relaxed in the *agnostic* PAC learning framework, which is general enough to cover both supervised and unsupervised learning tasks involving arbitrary distributions over their instance space (Haussler 1992). In essence, the agnostic PAC learning framework follows the general setting of statistical learning, investigated by (Vapnik, 1998, 2013).

Definition 2 (*Agnostic PAC Learning*) Let $\mathscr{Z}$ be an instance space, $\mathscr{H}$ be a hypothesis class over $\mathscr{Z}$, and $\ell : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$ be a loss function. Then, $\mathscr{H}$ is *agnostic* PAC *learnable* with respect to ℓ if there exist an algorithm LEARN with the following property: for any distribution $\mathscr{D}$ over $\mathscr{Z}$, and any $(\delta, \varepsilon) \in (0, 1)^2$, if LEARN is given inputs δ and ε, and access to $\text{EX}(\mathscr{D})$, then LEARN returns a hypothesis $h \in \mathscr{H}$ that satisfies, with probability $1 - \delta$,

$$L_{\mathscr{D}}(h) - \inf_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') \leq \varepsilon \tag{5}$$

Again, agnostic PAC learning is a distribution-free model: for every distribution over the instance space, the learner is ask to find with high probability, a model whose performance is near to that of the best model in its hypothesis class. It is important to emphasize that, in the agnostic case, $\mathscr{D}$ is a arbitrary distribution over the *whole* instance space $\mathscr{Z}$, and $\text{EX}(\mathscr{D})$ is a procedure that returns on each call an instance $\boldsymbol{z} \in \mathscr{Z}$ drawn independently at random according to $\mathscr{D}$. In particular, if $\mathscr{Z}$ is the instance space $\mathscr{X} \times \mathscr{Y}$ of a supervised learning task, then $\mathscr{D}$ is an arbitrary joint distribution over $\mathscr{X} \times \mathscr{Y}$, and $\text{EX}(\mathscr{D})$ generates a sample $(\boldsymbol{x}, \mathbf{y})$ where $\boldsymbol{y}$ is not determined by some hypothetical target function, but drawn at random with probability $\mathscr{D}(\boldsymbol{y} \mid \boldsymbol{x})$, whenever $\boldsymbol{x}$ is drawn at random with probability $\mathscr{D}(\boldsymbol{x})$.

3.2 Runtime Complexity

Based on the definition of agnostic PAC learnability, we might be tempted to characterize the runtime complexity of a PAC algorithm LEARN as the amount of computation it performs for returning with probability $1 - \delta$ a hypothesis whose risk is ε-close to the best possible risk. Yet, this measure is not really satisfactory, because we have swept under the rug two key issues.

The first issue is related to the input of the learning algorithm A. Typically, the runtime complexity of A does not only depend on the accuracy (ε) and confidence (δ) parameters, but also on the *dimension* d of the learning task. A natural approach for incorporating this parameter in the input of a statistical learning problem is consider *stratified* classes parameterized by d. Formally, a stratified instance space is a set $\mathscr{Z} = \bigcup_{d \in \mathbb{N}} \mathscr{Z}_d$, where each $\mathscr{Z}_d$ is a subset of $\mathbb{R}^d$. A stratified hypothesis class is defined in a similar way using $\mathscr{H} = \bigcup_{d \in \mathbb{N}} \mathscr{H}_d$. For example, if $\mathscr{H}$ is the class of

hyperplanes of arbitrary dimension, then $\mathscr{H}_1$ is the set of points in the line $\mathbb{R}$, $\mathscr{H}_2$ is the set of lines in the plane $\mathbb{R}^2$, $\mathscr{H}_3$ is the set of planes in the space $\mathbb{R}^3$, and so on. By extension, a stratified class of loss functions is a set $\mathscr{L} = \{\ell_d\}_{d\in\mathbb{N}}$, where each ℓ_d is a mapping $\mathscr{H}_d \times \mathscr{Z}_d \to \mathbb{R}$. Based on these stratified classes, the generalization ability of the algorithm LEARN is analyzed for every dimension d, every confidence δ and accuracy ε, and every distribution $\mathscr{D}$ over $\mathscr{Z}_d$.

The second issue is related to the output of the learner. Specifically, the model returned by a computer algorithm is not an abstract function $h \in \mathscr{H}$, but a symbolic *representation* of this mathematical object. From a computational viewpoint, this representation would be of little use if an exponential amount of computational resources was needed for inferring $h(\boldsymbol{x})$ given some incoming instance $\boldsymbol{x}$. So, to alleviate this issue, each candidate hypothesis $h \in \mathscr{H}$ should be associated with a representation for which the inference task is tractable. To this end, let $\mathscr{R}$ be a set of finite strings defined over some alphabet Σ. Then, $\mathscr{R}$ is called a *representation class* for $\mathscr{H}$ if there exist a surjective function from $\mathscr{R}$ to $\mathscr{H}$: each representation $\boldsymbol{r} \in \mathscr{R}$ is associated with exactly one hypothesis in $\mathscr{H}$ denoted $h_{\boldsymbol{r}}$, and each hypothesis $h \in \mathscr{H}$ is associated with at least one representation $\boldsymbol{r} \in \mathscr{R}$ such that $h_{\boldsymbol{r}} = h$. By extension, $\mathscr{R} = \bigcup_{d\in\mathbb{N}} \mathscr{R}_d$ is called a *stratified representation class* for $\mathscr{H} = \bigcup_{d\in\mathbb{N}} \mathscr{H}_d$ if for each dimension d, $\mathscr{R}_d$ is a representation class of $\mathscr{H}_d$.

With these notions in hand, we are now in position to provide a formal model of computationally efficient learnability. The next definition is essentially a variant of the computational learning models presented in Kearns et al. (1994b), Shalev-Shwartz and Ben-David (2014).

Definition 3 (*Efficient Agnostic PAC Learning*) Let $\mathscr{Z}$ be a stratified instance space, $\mathscr{H}$ be a stratified hypothesis class, and $\mathscr{L}$ be a stratified class of loss functions over $\mathscr{H}$ and $\mathscr{Z}$. In addition, let $\mathscr{R}$ be a stratified representation class for $\mathscr{H}$. Then, $\mathscr{H}$ is *efficiently agnostic* PAC *learnable* with respect to ℓ and $\mathscr{R}$ if both the following conditions hold:

- **Polynomial inference**: There exist an algorithm EVAL such that for every positive integer d, every representation $\boldsymbol{r} \in \mathscr{R}_d$, and every instance $\boldsymbol{x} \in \mathscr{Z}_d$, if EVAL is given inputs $\boldsymbol{r}$ and $\boldsymbol{x}$, then EVAL returns $h_{\boldsymbol{r}}(\boldsymbol{x})$ in time polynomial in d.
- **Polynomial convergence**: There exist an algorithm LEARN such that for every positive integer d, every distribution $\mathscr{D}$ over $\mathscr{Z}_d$, and every $(\delta, \varepsilon) \in (0, 1)^2$, if LEARN is given inputs d, δ and ε, and access to $\text{EX}(\mathscr{D})$, then LEARN returns in time polynomial in d, $1/\delta$ and $1/\varepsilon$ a representation $\boldsymbol{r} \in \mathscr{R}_d$ that satisfies, with probability $1 - \delta$,

$$L_{\mathscr{D}}(h_{\boldsymbol{r}}) - \inf_{h'\in\mathscr{H}_d} L_{\mathscr{D}}(h') \leq \varepsilon$$

Conceptually, there is a fundamental difference between "statistical learnability" specified in Definition 2, and "computational learnability" characterized by Definition 3. On the one hand, a hypothesis class $\mathscr{H}$ is statistically learnable if we can find an algorithm that converges with high probability to the best hypothesis in $\mathscr{H}$,

using a *finite* number of calls to the example oracle. On the other hand, computational learnability imposes much stronger conditions. In order to establish that $\mathscr{H}$ is efficiently learnable, we must not only prove that $\mathscr{H}$ is tractable for inference, but also find an algorithm that converges in probability to the best model, using a polynomial amount of operations. Since each call to the example oracle takes unit time, this directly implies that $\mathscr{H}$ must be learnable with polynomial sample complexity.

This crucial difference will be illustrated in the forthcoming sections. Many hypothesis classes of interest in the AI literature are learnable, even in the agnostic case, if computational considerations are not taken into account. By contrast, very few of them are *efficiently* learnable. An important class of statistical learning problems satisfying the property of efficient learnability is the family of convex learning problems, examined in Sect. 6.

4 Learning as Optimization

Arguably, statistical learning shares strong similarities with optimization problems. Based on the framework presented in Sect. 2.4, any statistical learning problem can be viewed as a *stochastic optimization problem*, involving a decision variable h defined over $\mathscr{H}$, a random variable z specified by a probability distribution $\mathscr{D}$ over $\mathscr{Z}$, and a loss function $\ell : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$. The problem is to

$$\begin{aligned} &\text{minimize} \quad \mathbb{E}_{z\sim\mathscr{D}}[\ell(h, z)] \\ &\text{subject to} \quad h \in \mathscr{H} \end{aligned} \tag{6}$$

Recall that the expression $\mathbb{E}_{z\sim\mathscr{D}}[\ell(h, z)]$ is the true risk of h, denoted $L_{\mathscr{D}}(h)$. The key specificity - and difficulty - of statistical learning lies in the fact that this objective function cannot be directly evaluated, since the underlying distribution $\mathscr{D}$ is unknown. In other words, statistical learning is a *black-box* stochastic optimization problem, for which the objective function can only be approximated using a limited number of calls to an example oracle $\text{Ex}(\mathscr{D})$. In the statistical learning literature, various optimization principles have been proposed for replacing the unknown risk function (6) with a known, evaluable objective function. In this section, we begin to review several optimization principles, and next, we examine some general conditions for learnability which justify the use of these principles, and open the door to new optimization strategies.

4.1 Optimization Principles

In statistical learning, the data generation process $\mathscr{D}$ is unknown, but we still do have access to a training sample S, given explicitly by a dataset, or implicitly through

an example oracle EX($\mathscr{D}$). Let $\mathscr{S}$ denote the set of all finite training sets over $\mathscr{Z}$, that is, $\mathscr{S} = \bigcup_{m \in \mathbb{N}} \mathscr{Z}^m$. The main idea behind most optimization principles in statistical learning is to replace the unknown objective function (6) defined over $\mathscr{D}$, with an evaluable objective function f_S defined for every training set $S \in \mathscr{S}$. The corresponding optimization problem is to

$$\begin{aligned} &\text{minimize} \quad f_S(h) \\ &\text{subject to} \quad h \in \mathscr{H} \end{aligned} \tag{7}$$

A *learning rule* is a map $A : \mathscr{S} \to \mathscr{H}$ that takes as input a training set $S \in \mathscr{S}$, and returns as output a hypothesis $A(S) \in \mathscr{H}$. We note in passing that any agnostic PAC learning algorithm LEARN can be unambiguously specified by a learning rule A and an integer-valued function $m : (0, 1)^2 \to \mathbb{N}$. Namely, given as input a desired confidence δ and a desired accuracy ε, the algorithm LEARN starts by picking a training set $S \in \mathscr{S}$ by calling $m(\delta, \varepsilon)$ times the example oracle EX($\mathscr{D}$), and then uses the learning rule A with S in order to produce a model $A(S) \in \mathscr{H}$. Here, $m(\delta, \varepsilon)$ captures the sample complexity of LEARN.

Based on these considerations, we say that a learning rule $A : \mathscr{S} \to \mathscr{H}$ *solves* the optimization task (7) if for every input $S \in \mathscr{S}$, the algorithm A returns as output a hypothesis $A(S) \in \mathscr{H}$ satisfying $f_S(A(S)) = \inf_{h \in \mathscr{H}} f_S(h)$. If in addition A runs in time polynomial in the dimension d of the training instances, and the size m of the training set S, then we say that A *efficiently solves* the optimization task (7).

4.1.1 Empirical Risk Minimization

Perhaps the most common approach for handling statistical learning problems is to replace the true risk function $L_{\mathscr{D}}$ by the empirical risk function L_S that measures the average loss of a model on the observed instances (Vapnik 1998; Zhang 2010). Based on this principle, called *Empirical Risk Minimization* (ERM), the objective function f_S, defined for a training set $S = (z_1, \ldots, z_m)$, is given by

$$f_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)] = L_S(h) \tag{8}$$

Correspondingly, any learning rule $A : \mathscr{S} \to \mathscr{H}$ that solves the optimization task (7), using (8) as objective function, is called an *empirical risk minimizer*.

Borrowing the terminology of stochastic optimization, the ERM principle is equivalent to the paradigm of *sample average approximation*, which aims at approximating the expected value function by a sample average function (Birge and Louveaux 2011). Though this idea is conceptually simple, and statistically justified by the law of large numbers, we must keep in mind that L_S is only an estimator of $L_{\mathscr{D}}$, In practice, the divergence between these objective functions depend on the choice of the hypothesis class $\mathscr{H}$ and the available training set S. More precisely, the true risk of the

hypothesis $A(S)$ returned by an empirical risk minimizer A can be decomposed as the sum of two terms:

$$L_{\mathscr{D}}(A(S)) = \underbrace{\inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h)}_{\text{approximation}} + \underbrace{\left[L_{\mathscr{D}}(A(S)) - \inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h) \right]}_{\text{estimation}}$$

The approximation term measures the minimum risk achievable by any possible model in the hypothesis class $\mathscr{H}$. The estimation term evaluates the performance of the hypothesis $A(S)$ chosen by the learning rule A, relatively to the best model in $\mathscr{H}$. By minimizing the sum of both terms, we are faced with a dilemma between approximation and estimation, called *bias-complexity trade-off* (Shalev-Shwartz and Ben-David 2014). On the one hand, if we choose a very rich hypothesis class $\mathscr{H}$, then we decrease the approximation error by covering good models for the task at hand but, at the same time, we increase the sample complexity required to guarantee that, with high probability, training sets are representative of the underlying distribution $\mathscr{D}$. Thus, if the available training set S is too small for achieving this guarantee, the objective function f_S is likely to be a poor estimator of $L_{\mathscr{D}}$, and hence, the hypothesis $A(S)$ is prone to *overfitting*, by having an optimal performance on training data, but a poor performance on test data. On the other hand, if we choose a very small hypothesis class $\mathscr{H}$, then we increase the odds that the available training set is representative, but we also increase the approximation error by missing good models for the learning task. So here, $A(S)$ is prone to *underfitting*, by exhibiting a relatively stable, but low performance, on both training data and test data.

4.1.2 Structural Risk Minimization

A natural idea to prevent overfitting situations is to penalize complex hypotheses, in favor of simpler ones, whenever they share the same empirical risk. This idea follows the well-known law of parsimony, according to which *plurality should not be posited without necessity*. This law, called *Occam's razor* after the philosopher William of Ockham, gives precedence to simplicity: of two competing theories, the simpler explanation of an entity is to be preferred.

In the paradigm of *Structural Risk Minimization* (SRM), due to Vapnik and Chervonenkis (Vapnik and Chervonenkis 1974), it is assumed that the hypothesis class $\mathscr{H}$ is associated with a stratified representation class $\mathscr{R} = \bigcup_{k \in \mathbb{N}} \mathscr{R}_k$, where k is a structural parameter. For instance, if $\mathscr{H}$ is the class of all (zero-threshold) separating hyperplanes over the domain set $\mathscr{X} \subseteq \mathbb{R}^n$, then its representation class $\mathscr{R} \subseteq \mathbb{R}^n$ can be stratified by the number k of nonzero weights. Namely, each stratum $\mathscr{R}_k$ is the set of all weight vectors $\boldsymbol{w} \in \mathbb{R}^n$ such that $\|\boldsymbol{w}\|_0 \leq k$. Given a model $h \in \mathscr{H}$, we use k_h to denote the smallest integer k such that $h = h_{\boldsymbol{r}}$ for at least one representation $\boldsymbol{r} \in \mathscr{R}_k$. Based on these notions, the objective function is a mapping of the form

$$f_S(h) = L_S(h) + \text{pen}_{m,\delta}(k_h) \tag{9}$$

where $\text{pen}_{m,\delta}$ is a penalty function that depends on the size m of the training set, and a confidence parameter $\delta \in (0, 1)$. Ideally, the penalty function should satisfy the condition that for every confidence $\delta \in (0, 1)$ and every distribution $\mathscr{D}$ over the instance space, with probability $1 - \delta$ over the choice of $S \sim \mathscr{D}^m$, the following bound holds for any hypothesis $h \in \mathscr{H}$:

$$|L_{\mathscr{D}}(h) - L_S(h)| \leq \text{pen}_{m,\delta}(k_h) \tag{10}$$

If this condition is indeed satisfied, then the estimation error of h is bounded by $L_S(h) + \text{pen}_{m,\delta}(k_h)$. In other words, the SRM principle handles the bias-complexity trade-off by giving preference to simple hypotheses (with small penalty value) which behave well on the training set.

A closely related paradigm is the *Minimum Description Length* (MDL) principle, due to Rissanen (1983, 1985), and surveyed in detail by Grünwald (2007). Here, it is assumed that the hypothesis class $\mathscr{H}$ is associated with a *prefix-free* representation class $\mathscr{R}$. Namely, $\mathscr{R}$ is a prefix-free language if no representation $\boldsymbol{r} \in \mathscr{R}$ is the prefix of a distinct representation $\boldsymbol{r}' \in \mathscr{R}$. Notice that $\mathscr{R}$ can be viewed as a stratified representation class $\bigcup_{k\in\mathbb{N}} \mathscr{R}_k$, where $\mathscr{R}_k$ is the set of all representations, or "codewords", of length k. Based on this observation, h_k measures the length of the smallest codeword $\boldsymbol{r}$ such that $h = h_{\boldsymbol{r}}$, and it is simply denoted $|h|$. In the MDL principle, the objective function is given by

$$f_S(h) = L_S(h) + \text{pen}_{m,\delta}(|h|) \;\text{ where }\; \text{pen}_{m,\delta}(|h|) = \sqrt{\frac{|h| + \ln \frac{2}{\delta}}{2m}} \tag{11}$$

Notably, using the well-known Kraft's inequality property of prefix-free languages, it can be shown that the penalty function $\text{pen}_{m,\delta}(|h|)$ satisfies the condition (10). A detailed proof is given in Shalev-Shwartz and Ben-David (2014).

To sum up, the MDL paradigm provides an elegant way to circumvent the pitfall of overfitting in rich hypothesis classes, by penalizing models with their code length. However, the MDL principle does not come without practical issues: a key difficulty in the design of MDL-based learning algorithms is to find an appropriate prefix free representation language for the hypothesis class at hand. Another important issue is the runtime complexity of the optimization task. Notably, if the loss function ℓ is convex, then ERM objective (8) remains convex, but the MDL objective (11) is generally not convex due to the additional, non-convex, penalty term. Similar computational issues arise for the more general SRM principle, for which penalty functions in (9) are typically not convex.

4.1.3 Regularized Risk Minimization

For hypothesis classes $\mathscr{H}$ represented by linear functions, the predominant approach to penalize complex models is through "regularizing" their representation. In this setting, called *Regularized Risk Minimization* (RRM), the representation class of $\mathscr{H}$ is a set of weight vectors, denoted here $\mathscr{W}$. The objective function f_S takes the following form:

$$f_S(h) = L_S(h) + \text{reg}(\boldsymbol{w}) \tag{12}$$

where $\boldsymbol{w}$ is the vector representation of h, and $\text{reg} : \mathscr{W} \to \mathbb{R}$ is a regularization term that penalizes hypotheses according to the "complexity" of their vector representation. The complexity of vectors is typically measured using some norm over $\mathscr{W}$. For example, the regularizer $\text{reg}(\boldsymbol{w}) = \lambda \|\boldsymbol{w}\|_2^2$ due to Tikhonov (1943), penalizes weights with large magnitudes. Alternatively, the regularizer $\text{reg}(\boldsymbol{w}) = \lambda \|\boldsymbol{w}\|_1$ gives preference to parsimonious models involving few nonzero weights. In both expressions, the parameter λ is a positive scalar that controls the regularization effect. We emphasize that regularization functions are not always defined through norms. For instance, the entropic regularizer $\text{reg}(\boldsymbol{w}) = \lambda \sum_i w_i \ln {}^1/_{w_i}$ is often used when the representation class $\mathscr{W}$ is a probability simplex.

Obviously, the RRM paradigm shares strong similarities with the SRM principle: both approaches aim at preventing overfitting issues by penalizing models which are excessively complex for the task at hand. From a pragmatic viewpoint, there are, yet, important differences related to the formulation of the statistical learning problem as an optimization task, and the resolution of this optimization task. The regularization term in RRM is often specified by a simple analytic form, while the penalty term in SRM is typically much more difficult to characterize. For example, the penalty term in (11) is defined using the code length $|h|$ of a model $h \in \mathscr{H}$, which requires a prefix-free representation language for $\mathscr{H}$. Furthermore, most regularization terms in the Machine Learning literature are convex functions. If, in addition, the representation class $\mathscr{W}$ is convex, and the loss function is convex for $\mathscr{W}$, then the optimization task (7) using (12) as objective function is a convex optimization problem, which can be efficiently solved by a wide variety of algorithms. As mentioned above, objective functions for SRM and MDL principles typically lead to intractable optimization tasks, due to the non-convex nature of penalty terms.

4.2 Conditions for Learnability

The overall goal of optimization principles in statistical learning is to reformulate the black-box stochastic optimization task (6) as a standard, well-formed, optimization task (7). If we put aside computational considerations, there is still an important question that emerges from those principles: under which conditions an optimization algorithm for (7) is guaranteed, with high probability, to solve (6)?

In the statistical learning literature, various conditions for learnability have been proposed, in order to characterize the key relationships between learning and optimization (Vapnik 1998; Bousquet and Elisseeff 2002; Poffio et al. 2004; Mohammadi and van de Geer 2005; Mukherjee et al. 2006; Watanabe 2009; Wibisono et al. 2009; Shalev-Shwartz et al. 2010; Liu et al. 2017). We shall concentrate on two of them, namely, *uniform convergence* and *stability*, which play a central role in statistical learning theory.

To this end, we need some additional definitions. A *rate function* is a monotone decreasing mapping ϵ: $\mathbb{N} \to \mathbb{R}$ that converges to 0 as m tends to infinite. With these notions in hand, a learning rule A is called *(universally) consistent* with rate ϵ_{cons} if for any $m \in \mathbb{N}$ and any distribution $\mathscr{D}$ over $\mathscr{Z}$,

$$\mathbb{E}_{S\sim\mathscr{D}^m}[L_{\mathscr{D}}(A(S))] - \inf_{h\in\mathscr{H}} L_{\mathscr{D}}(h) \leq \epsilon_{\text{cons}}\ (m)$$

The next result, derived from Shalev-Shwartz et al. (2010), Sridharan (2012), states that consistency is a necessary and sufficient condition for achieving learnability in the setting of *bounded* loss functions, that is, cost functions of the form $\ell : \mathscr{H} \times \mathscr{Z} \to [0, b]$, where b is a positive scalar.

Theorem 1 (Learnability as Consistency) *Let $\mathscr{Z}$ be an instance space, $\mathscr{H}$ be a hypothesis class over $\mathscr{Z}$, and $\ell : \mathscr{H} \times \mathscr{Z} \to [0, b]$ be a bounded loss function. Then, $\mathscr{H}$ is (agnostic PAC) learnable with respect to ℓ if and only if there is a learning rule A for $\mathscr{H}$ and a rate function ϵ_{cons} such that A is consistent with rate ϵ_{cons}.*

4.2.1 Uniform Convergence

For bounded loss functions, the statistical learning problem is to find a learning rule that achieves a uniform rate for all distributions. To this point, it is well-known that the empirical risk minimizer is consistent, provided that its hypothesis class satisfies the *uniform convergence* property (Vapnik 1998, 2013). This key condition for learnability can be formalized in the following way.

Definition 4 (*Uniform Convergence*) Let $\mathscr{Z}$ be an instance space, $\mathscr{H}$ be a hypothesis class over $\mathscr{Z}$, and $\ell : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$ be a loss function. Then, $\mathscr{H}$ has the *uniform convergence property* with respect to ℓ if for every distribution $\mathscr{D}$ over $\mathscr{Z}$,

$$\lim_{m\to\infty} \mathbb{E}_{S\sim\mathscr{D}^m}\left[\sup_{h\in\mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)|\right] = 0$$

Intuitively, the quantity $\sup_{h\in\mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)|$ measures the ability of a training set S to adequately represent the underlying distribution $\mathscr{D}$ for the task at hand. Given an accuracy parameter ε, the training set S is called ε *-representative* if for all hypotheses $h \in \mathscr{H}$, we have $|L_{\mathscr{D}}(h) - L_S(h)| \leq \varepsilon$. Based on this notion, a hypothesis class $\mathscr{H}$ has the uniform convergence property if there exist an integer-valued

function $m_{\mathscr{H}} : (0,1)^2 \to \mathbb{N}$ such that, for every pair $(\delta, \varepsilon) \in (0,1)^2$, and every distribution $\mathscr{D}$ over $\mathscr{Z}$, if the example oracle $\text{EX}(\mathscr{D})$ is called $m \geq m_{\mathscr{H}}(\delta, \varepsilon)$ times, then the resulting sample $S \in \mathscr{Z}^m$ is ε-representative with probability $1 - \delta$. Based on this reformulation of uniform convergence, the metric $m_{\mathscr{H}}$ shares similarities with the sample complexity of learning. Specifically, $m_{\mathscr{H}}(\delta, \varepsilon)$ is the amount of information needed to ensure that, with probability $1 - \delta$, the training set S supplied to the learner is ε-representative. Thus, if S is sufficiently large, then the empirical risk of hypotheses is a faithful approximation of their true risk. The ERM principle (8) can therefore be used without the need of penalty or regularization terms.

Theorem 2 (Learnability via Uniform Convergence) *Let $\mathscr{Z}$ be an instance space, $\mathscr{H}$ be a hypothesis class over $\mathscr{Z}$, and $\ell : \mathscr{H} \times \mathscr{Z} \to [0, b]$ be a bounded loss function. If $\mathscr{H}$ has the uniform convergence property with sample complexity $m_{\mathscr{H}}(\delta, \varepsilon)$, then $\mathscr{H}$ is (agnostic PAC) learnable with sample complexity $m_{\mathscr{H}}(\delta, \varepsilon/2)$, and the empirical risk minimizer is consistent.*

Interestingly, for supervised classification and regression tasks, a converse result also holds; namely, $\mathscr{H}$ is learnable *if and only if* it enjoys the uniform convergence property (Blumer et al. 1989; Alon et al. 1997).

For rich hypothesis classes $\mathscr{H}$, the sample complexity $m_{\mathscr{H}}(\delta, \varepsilon)$ required to ensure uniform convergence can be much larger than the size of training sets available in practice, and hence, the ERM rule is prone to overfitting. So, we need here a weaker form of uniform convergence that justifies the use of alternative principles, such as SRM. To this end, assume that $\mathscr{H}$ is associated with a stratified representation class $\mathscr{R} = \bigcup_{k \in \mathbb{N}} \mathscr{R}_k$, and let $\mathscr{H}_k$ be the set of models represented by $\mathscr{R}_k$. Then, $\mathscr{H}$ is said to have the *locally uniform convergence* property if each $\mathscr{H}_k$ enjoys the uniform convergence condition with sample complexity $m_{\mathscr{H}_k}$. Intuitively, the quantity $m_{\mathscr{H}_k}(\delta, \varepsilon)$ is small for simple hypothesis classes $\mathscr{H}_k$, and increases with the structural parameter k. Given a sample size m, let $\varepsilon_k(m, \delta)$ be the minimum value of $\varepsilon \in (0, 1)$ for which $m_{\mathscr{H}_k}(\delta, \varepsilon) \leq k$. Since $\mathscr{H}_k$ has the uniform convergence property, it follows that any training sample $S \sim \mathscr{D}^m$ is $\varepsilon_k(m, \delta)$-representative with probability $1 - \delta$. Thus, the penalty rule

$$\text{pen}_{m,\delta}(k_h) = \varepsilon\left(m, \frac{\delta}{2^{k_h}}\right)$$

satisfies the condition (10), which in turn implies that any structural risk minimizer defined on this penalty rule is consistent. In a nutshell, the locally uniform convergence property is a sufficient condition for learnability using the SRM paradigm.

4.2.2 Stability

In contrast with uniform convergence, a condition defined for hypothesis classes, stability is a property related to learning rules. Intuitively, a learning algorithm is characterized by an overfitting behavior when it overreacts to small fluctuations in

the training data. Put another way, a learning rule $A : \mathcal{S} \rightarrow \mathcal{H}$ is stable if a small change of the input $S \in \mathcal{S}$ will only induce a small change of the output $h \in \mathcal{H}$.

The next definition of stability, often referred to as *average replace-one stability*, is based on replacing one instance in the training set, with a new instance drawn at random according to the underlying distribution. Given a sample $S = (z_1, \ldots, z_m)$ and an instance $z' \in \mathcal{Z}$, let $S_{z_i \leftarrow z'} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$ be the sequence obtained by replacing the ith observation of S with the instance z'.

Definition 5 (*Stability*) Let $\mathcal{Z}$ be an instance space, $\mathcal{H}$ be a hypothesis class over $\mathcal{Z}$, and $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function. Then, a learning rule A for $\mathcal{H}$ is *(on average replace-one) stable* with rate ϵ_{stable} if for any distribution $\mathcal{D}$ over $\mathcal{Z}$,

$$\frac{1}{m}\left|\sum_{i=1}^{m} \mathbb{E}_{S\sim\mathcal{D}^m,(z'_1,\ldots,z'_m)\sim\mathcal{D}^m}\left[\ell\left(A(S_{z_i\leftarrow z'_i}); z'_i\right) - \ell\left(A(S); z'_i\right)\right]\right| \leq \epsilon_{\text{stable}}\ (m)$$

For stable learning rules, the ERM principle is *not* a necessary condition for ensuring learnability. Instead, the learner is only required to converge toward the ERM minimizer when the number m of training instances tends to infinite. Formally, a learning rule A is an *Asymptotic Empirical Risk Minimizer* (AERM) with rate ϵ_{erm} if for any distribution $\mathcal{D}$ over $\mathcal{Z}$,

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_S(A(S)) - \inf_{h\in\mathcal{H}} L_S(h)\right] \leq \epsilon_{\text{erm}}\ (m)$$

The next result, demonstrated in Shalev-Shwartz et al. (2010), establishes an equivalence between statistical learnability and stable AERM rules.

Theorem 3 (Learnability via Stability) *Let $\mathcal{Z}$ be an instance space, $\mathcal{H}$ be a hypothesis class over $\mathcal{Z}$, and $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, b]$ be a bounded loss function. Then $\mathcal{H}$ is (agnostic PAC) learnable if and only if there exists a stable* AERM *for $\mathcal{H}$. In particular, if a learning rule A is stable with rate ϵ_{stable} and* AERM *with rate ϵ_{erm}, then A is consistent with rate*

$$\epsilon_{\text{cons}}\ (m) \leq \epsilon_{\text{stable}}\ (m) + \epsilon_{\text{erm}}\ (m)$$

In a nutshell, uniform convergence and stability provide different mathematical tools for building learning algorithms. If the hypothesis class $\mathcal{H}$ is endowed with uniform convergence, then Empirical Risk Minimization is the paradigm of choice for designing a learning rule with a good generalization ability. Yet, $\mathcal{H}$ may be learnable even if it does not satisfy the uniform convergence property: in this case, stable asymptotic empirical risk minimizers are guaranteed to work. For convex learning problems described in Sect. 6, such learning rules can be constructed in a simple and intuitive manner using the Regularized Risk Minimization principle.

5 Concept Learning

Basically, the problem of concept learning is to extrapolate, from a series of positive and negative examples, a model that accurately separate future, unseen instances. In other words, concept learning problems are binary classification tasks whose objective function is the zero-one loss. The instance space $\mathscr{Z}$ is a set $\mathscr{X} \times \{0, 1\}$ of instances labeled as negative (0) or positive (1). A concept is a subset of $\mathscr{X}$, or equivalently, an indicator function h mapping $\mathscr{X}$ to $\{0, 1\}$. By extension, a concept class is a subset $\mathscr{H} \subseteq \{0, 1\}^{\mathscr{X}}$. Recall that the zero-one loss function ℓ over $\mathscr{H}$ and $\mathscr{Z}$ is given by:

$$\ell(h; (\boldsymbol{x}, y)) = \begin{cases} 0 & \text{if } h(\boldsymbol{x}) = y \\ 1 & \text{otherwise} \end{cases}$$

Based on this objective function, the true risk and the empirical risk of a concept can be viewed as error measures. Namely, $L_{\mathscr{D}}(h)$ captures the probability that the concept h is making a mistake on a labeled instance $(\boldsymbol{x}, y)$ drawn at random according to $\mathscr{D}$. $L_S(h)$ is the proportions of mistakes made by h on the training set S.

In this section, we begin to examine the *Vapnik-Chervonenkis dimension* of concept classes, an important notion related to their sample complexity. We next survey some theoretical results related to learning concepts in the realizable case and the agnostic case. We close this section by briefly discussing about *bagging* and *boosting*, two efficient techniques for learning combinations of models.

5.1 VC-Dimension

As explained in Sect. 4.2, the uniform convergence property is a sufficient condition for establishing the learnability of hypothesis classes. In concept learning, this property is intrinsically related to the classification power of the concept class, called *Vapnik-Chervonenkis* (VC) *dimension* (Vapnik and Chervonenkis 1974). Intuitively, the VC-dimension of $\mathscr{H}$ is the maximum size of any set of input objects which can be labeled in any possible way using concepts taken from $\mathscr{H}$. More formally, let $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \subseteq \mathscr{X}$ be a set of m input objects, and let

$$\mathscr{H}_S = \{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_m) : h \in \mathscr{H}\}$$

be the restriction of $\mathscr{H}$ to S, that is, the set of functions from S to $\{0, 1\}$ which can be derived from $\mathscr{H}$. Then, S is called *shattered* by $\mathscr{H}$ if $\mathscr{H}_S$ is the set of all possible Boolean functions from S to $\{0, 1\}$, that is, $|\mathscr{H}_S| = 2^{|S|}$.

Definition 6 (*VC-dimension*) Let $\mathscr{X}$ be a set, and $\mathscr{H}$ be a set of functions from $\mathscr{X}$ to $\{0, 1\}$. Then, the VC-*dimension* of $\mathscr{H}$, denoted VCdim($\mathscr{H}$), is the maximal size

Table 1 VC-dimension of some concept classes. A k-term DNF formula is a disjunction of at most k monomials, and a k-DNF formula is a disjunction of monomials, with at most k literals per term

Concept class	VCdim
Monotone monomials on $\{0,1\}^n$	n
Homogeneous Linear functions on $\mathbb{R}^n$	n
Linear threshold functions on $\mathbb{R}^n$	$n+1$
Feedforward linear threshold neural networks with E edges on $\mathbb{R}^n$	$6E \log_2 E$
k-term DNF formulas on $\{0,1\}^n$	$\Theta(kn)$
k-DNF formulas on $\{0,1\}^n$	$\Theta(n^k)$
Polynomial threshold functions of degree k on $\mathbb{R}^n$	$\binom{n+k}{k}$
Arbitrary DNF formulas on $\{0,1\}^n$	2^n
Arbitrary functions from $\mathbb{R}^n$ to $\{0,1\}$	∞

of any set $S \subseteq \mathscr{X}$ that is shattered by $\mathscr{H}$. If $\mathscr{H}$ can shatter sets of arbitrary large size, then $\text{VCdim}(\mathscr{H}) = \infty$.

We mention in passing that for a finite class $\mathscr{H}$, a set S of instances cannot by shattered by $\mathscr{H}$ if $|\mathscr{H}| < 2^{|S|}$. It follows that

$$\text{VCdim}(\mathscr{H}) \leq \log_2 |\mathscr{H}|$$

Actually, the VC-dimension of finite concept classes $\mathscr{H}$ can be much smaller than the logarithm of their size. Consider for example the class $\mathscr{H} = \{h_1, \ldots, h_n\}$ of Boolean functions from $\{0,1\}^n \to \{0,1\}$, defined as follows: $h_i(\boldsymbol{x}) = 1$ if and only if all features in $\boldsymbol{x}$ ranging from i to n are set to 1. Clearly, $\mathscr{H}$ can shatter a singleton set $S = \{\boldsymbol{x}\}$ using $x_1 = 0$ and $x_2 = 1$. Yet, $\mathscr{H}$ cannot shatter any pair of input objects $S = \{\boldsymbol{x}, \boldsymbol{x}'\}$, because there is no pair of hypotheses $\mathscr{H}$ for which the first gives the labeling $(0,1)$ and the second gives the opposite labeling $(1,0)$. So, the VC-dimension of $\mathscr{H}$ is 1, and since n can be arbitrary large, the gap between $\text{VCdim}(\mathscr{H})$ and $\log_2 |\mathscr{H}|$ may be arbitrary large.

The VC-dimension of several concept classes is reported on Table 1; the proofs may be found in Anthony (2001, 2010). It is important to keep in mind that some infinite classes, such as linear threshold functions and feedforward neural networks, have a finite (and sometimes low) VC-dimension. The next theorem is a standard result in statistical learning theory, and its proof can be found in various textbooks (Anthony and Barlett 1999; Vapnik 2013; Shalev-Shwartz and Ben-David 2014).

Theorem 4 (Learnability of Concept Classes) *Let $\mathscr{H}$ be a hypothesis class from a domain $\mathscr{X}$ to $\{0,1\}$, and let ℓ be the zero-one loss function. Then, then following are equivalent:*

- *$\mathscr{H}$ has a finite VC-dimension.*
- *$\mathscr{H}$ has the uniform convergence property.*
- *$\mathscr{H}$ is agnostic PAC learnable.*

In particular, if $\mathrm{VCdim}(\mathscr{H}) \leq d$, *then the sample complexity of* $\mathscr{H}$ *is in* $\mathscr{O}(\frac{d+\ln(1/\delta)}{\varepsilon^2})$.

An important notion related to the VC-dimension is the *growth function* of a hypothesis class, which measures the number of different functions from a set S of size m to $\{0, 1\}$ that can be obtained by restricting $\mathscr{H}$ to S. Formally, the growth function of $\mathscr{H}$, is the mapping $\Pi_{\mathscr{H}} : \mathbb{N} \to \mathbb{N}$ given by

$$\Pi_{\mathscr{H}}(m) = \max_{S \subseteq \mathscr{X} : |S| = m} |\mathscr{H}_S|$$

Clearly, if the VC-dimension of $\mathscr{H}$ is d, then $\Pi_{\mathscr{H}}(m) = 2^d$ for all $m \leq d$. More precisely, by Sauer's Lemma (1972), the growth function of a concept class $\mathscr{H}$ for which the VC-dimension is upper-bounded by d satisfies $\Pi_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$ for all $m \in \mathbb{N}$. In particular, when m is becoming larger than d, the growth function is bounded by $(em/d)^d$, that is, $\Pi_{\mathscr{H}}$ increases polynomially with m. As a direct corollary of Theorem 4, if $\mathscr{H}$ has a finite VC-dimension, then $\mathscr{H}$ is agnostic PAC learnable with a sample complexity that is logarithmic in $\Pi_{\mathscr{H}}$.

5.2 Realizable Concept Learning

We first explore the PAC learnability of concept classes in the *realizable* setting, where a target function in the concept class is labeling the instances supplied to the learner. A useful algebraic tool in realizable PAC learning is the notion of *version space*, due to Mitchell (1982). Given a concept class $\mathscr{H}$ and a training sample $S \subseteq \mathscr{X} \times \{0, 1\}$, the version space of $\mathscr{H}$ with respect to S is given by

$$\mathrm{VS}(\mathscr{H}, S) = \{h \in \mathscr{H} \mid h(\boldsymbol{x}) = y \text{ for all } (\boldsymbol{x}, y) \in S\}$$

Let $\mathscr{D}$ denote the hidden distribution over $\mathscr{X}$, and $h^* \in \mathscr{H}$ denote the hidden target concept. Given a desired accuracy $\varepsilon \in (0, 1)$, the version space of $\mathscr{H}$ with respect to S is called *ε-exhausted* if $L_{\mathscr{D}}(h) \leq \varepsilon$ for any hypothesis in $\mathrm{VS}(\mathscr{H}, S)$. In other words, all candidate concepts in an ε-exhausted version space have error at most ε with respect to h^*. The following result, established in Blumer et al. (1989), Haussler (1988), provides a relation between ε-exhausted version spaces and the growth function of the concept class.

Theorem 5 *Let* $\mathscr{H}$ *be a hypothesis class from a domain* $\mathscr{X}$ *to* $\{0, 1\}$, *and let* ℓ *be the zero-one loss function. In addition, let* $\mathscr{D}$ *be a arbitrary distribution over* $\mathscr{X}$, *and* $h^* \in \mathscr{H}$ *be a target concept. Then for any* $\varepsilon \in (0, 1)$ *and any training sample* S *of size* m *drawn from* $\mathscr{D}$ *and labeled by* h^*, *the probability that* $\mathrm{VS}(\mathscr{H}, S)$ *is not* ε*-exhausted is at most*

$$2\Pi_{\mathscr{H}}(2m)2^{-\varepsilon m/2}$$

As a corollary, if the size m of the training sample S is at least

$$\frac{4}{\varepsilon}\left[\mathrm{VCdim}(\mathscr{H})\log_2\left(\frac{12}{\varepsilon}\right)+\log_2\left(\frac{2}{\delta}\right)\right] \tag{13}$$

the version space is ε-exhausted with probability $1-\delta$. Consequently, the concept class $\mathscr{H}$ is PAC learnable with a sample complexity which is linear in the VC-dimension of $\mathscr{H}$. So, in order to show that logical concept classes of polynomial VC-dimension are *efficiently* PAC learnable in the realizable case, we simply need to devise an algorithm that returns in polynomial time an element in the version space $\mathrm{VS}(\mathscr{H}, S)$, given as input a training sample S of size at least (13). In other words, realizable PAC learning is essentially a consistency (or feasibility) problem: given a set of labeled instances, find a concept that correctly labels all instances.

For simple concept classes, the consistency problem is relatively straightforward. For example, monomials and clauses may be learned using a standard variable elimination algorithm (Mitchell 1982; Kearns et al. 1987). Parity functions represented by XOR clauses can be learned using a closure algorithm (Helmbold et al. 1992). For linear threshold functions, the feasibility problem can be cast as a standard Linear Programming (LP) task, and hence, may be solved in polynomial time using an LP method. Here, the incremental Perceptron algorithm (Rosenblatt 1958) is more attractive in practice, but it is not generally efficient, because the number of its iterations depends on the margin of the training set, which can be exponential in the input dimension n (Anthony and Shawe-Taylor 1993).

Much less obvious is the consistency issue of expressive concept classes. On the one hand, k-DNF are efficiently PAC learnable using a simple extension of the variable elimination algorithm, and decision lists with clauses of size at most k can be efficiently learned using Rivest's algorithm (1987). On the other hand, for k-term DNF formulas, the consistency problem is NP-hard (even for $k = 3$) (Pitt and Valiant 1988). Similar hardness results have been found for expressive classes of geometric models: the consistency problem is NP-hard for k intersections of halfspaces (even for $k = 2$) (Megiddo 1988; Blum and Rivest 1992).

The above negative results hold for realizable and *proper* PAC learning; the concept returned by the learner must be a representation of a model in the hypothesis class $\mathscr{H}$. What about relaxing this condition? Namely, the computational issue of finding a representation of a model in $\mathscr{H}$ that is consistent with the data may be circumvented by allowing the learner to output in polynomial time a representation of a model in some larger concept class $\mathscr{H}'$ that includes $\mathscr{H}$. In this relaxed setting, often referred to as *improper* or *representation independent* PAC learning, the aforementioned class of k-term DNF formulas is efficiently learnable using k-CNF formulas, simply because any disjunction of k monomials can be encoded into a CNF expression, involving at most k literals per clause. Based on a similar encoding, the class of decision trees with at most s leaves is efficiently learnable using $\log_2 s$-decision lists (Blum 1992). In this representation independent setting, various *sub-exponential* time algorithms have been found for learning expressive concepts, such as DNF formulas or intersections of halfspaces, using polynomial threshold

representations (Klivans et al. 2004; Klivans and Servedio 2004). Yet, *polynomial* time learning algorithms seem to be unachievable, under the standard assumption that NP $\neq$ RP. Notably, several negative results indicate that DNF formulas are not efficiently learnable in the representation independent setting (Alekhnovich et al. 2008; Daniely and Shalev-Shwartz 2016). Analogous results have been obtained for intersections of halfspaces (Klivans and Sherstov 2009).

5.3 Agnostic Concept Learning

In the agnostic PAC learning setting, which does not make any assumption about the labels of incoming instances, the growth function of a hypothesis class, and hence its VC-dimension, may be used for assessing the sample complexity of binary classification under the zero-one loss. The proof of the next result, related to the sample complexity of uniform convergence, can be found in several textbooks (Mohri et al. 2012; Shalev-Shwartz and Ben-David 2014).

Theorem 6 *Let $\mathscr{H}$ be a hypothesis class from a domain $\mathscr{X}$ to $\{0, 1\}$, and let ℓ be the zero-one loss function. Then, for every distribution $\mathscr{D}$ over $\mathscr{X} \times \{0, 1\}$, every $\delta \in (0, 1)$ and every $h \in \mathscr{H}$, with probability $1 - \delta$ over the choice of $S \sim \mathscr{D}^m$,*

$$|L_D(h) - L_S(h)| \leq \sqrt{\frac{2 \ln \prod_{\mathscr{H}}(m)}{m}} + \sqrt{\frac{\ln 1/\delta}{2m}}$$

Thus, by combining the above result with Theorem 2, it follows that if $\mathscr{H}$ has a finite VC-dimension, then $\mathscr{H}$ is agnostic PAC learnable with sample complexity

$$O\left(\frac{\mathrm{VCdim}(\mathscr{H}) + \ln 1/\delta}{\varepsilon^2}\right)$$

As shown in Anthony and Barlett (1999), this asymptotic bound is tight: the O function can be replaced with the Θ function. Thus, the increase of sample complexity is mainly related to the accuracy parameter: the dependence on $1/\varepsilon$ is nearly linear in the realizable case, while it is quadratic in the agnostic case.

From a computational viewpoint, a sufficient condition for achieving efficient agnostic PAC learnablity is a polynomial time empirical risk minimizer. Indeed, as established in Theorem 2, the ERM learning rule is statistically consistent whenever $\mathscr{H}$ is endowed with the uniform convergence property. To this point, recall that realizable concept learning is a feasibility problem: find $h \in \mathscr{H}$ such that $L_S(h) = 0$. By contrast, agnostic concept learning is an optimization problem: minimize $L_S(h)$ subject to $h \in \mathscr{H}$. This crucial difference has drastic consequences on the computational learnability of concept classes. Notably, for simple classes such as monotone monomials and linear threshold functions, the problem of finding a concept that minimizes is empirical error on a training sample is NP-hard (Johnson and Preparata

1978; Angluin and Laird 1987; Höffgen and Simon 1992; Kearns and Li 1993; Kearns et al. 1994b). Consequently, monotone monomials and linear threshold functions are *not* efficiently agnostic PAC learnable, unless NP = RP.

In order to alleviate this computational barrier, a natural approach is to consider approximation schemes: for a given approximation parameter $\alpha \geq 1$, an α-approximation algorithm for $\mathscr{H}$ is a polynomial-time algorithm that takes as input an arbitrary sample S, and returns as output a hypothesis $h \in \mathscr{H}$, such that $L_S(h) \leq \alpha \inf_{h' \in \mathscr{H}} L_S(h')$. In other words, the learner must find a concept for which the empirical error is at most α times the empirical error of the ERM rule. Unfortunately, even under this relaxed setting, the problem of approximately minimizing the empirical error of monotone monomials and linear threshold functions remain NP-hard (Arora et al. 1997; Ben-David et al. 2003; Feldman et al. 2009).

Another approach, already suggested for realizable concept learning, is to allow the learner to return hypotheses in some class $\mathscr{H}'$ that covers $\mathscr{H}$. Yet, even in this representation-independent setting, simple concept classes are hard to learn (under the usual assumption that NP $\neq$ RP). For instance, monomials are not efficiently agnostic PAC learnable using arbitrary disjunctions of conjunctions (Kearns et al. 1994b), or halfspaces (Feldman et al. 2012).

In a nutshell, concept learning is an area of stark contrast from the viewpoint of runtime complexity. On the one hand, realizable concept learning is computationally easy for relatively simple classes, but remains difficult for more expressive hypothesis classes. On the other hand, the more "realistic" problem of agnostic concept learning proves to be very hard, even for simple hypothesis classes.

5.4 Bagging and Boosting

As expressive models are difficult to learn, what about learning simple models and combining them together, in order to produce more accurate predictors? Bagging and boosting are two techniques which grew out of this pragmatic question and became very practical tools for solving complex learning problems. The basic idea underlying these techniques is to amplify the accuracy of *weak learners*. One can think of a weak learner as an algorithm that uses a simple heuristic or "rule of thumb" in order to produce a hypothesis whose performance is just slightly better than a pure random guess. If such a weak learner can be implemented efficiently, then bagging and boosting may be used to iteratively combine weak hypotheses in order to produce a gradually better predictor. In what follows, we assume that $\mathscr{H}$ is closed under linear combinations, in order to produce model ensembles.

Introduced by Breiman (1996), the boostrap aggregating technique, abbreviated as *bagging*, aims at creating diverse weak hypotheses on different random samples of the training set S. As explained in Algorithm 1, These samples are taken uniformly with replacement, and a simple averaging of weak hypotheses is used to produce the final predictor. Bagging is particularly useful for learning combinations of decision trees, trained with weak learners such as ID3 (Quinlan 1986) or C4.5 (Quinlan

1993, 1996). When applied to tree models, bagging is often coupled with another idea, referred to as *subspace sampling*: at each iteration $t \in [T]$, select uniformly at random $n' \leq n$ features from $\mathscr{X}$ and train the weak learner A (without pruning) on the sample S'_t formed by the projection of S_t onto $[n']$. This encourages the diversity in the ensemble of weak hypotheses, and contributes to reduce the runtime of learning. The resulting method, called *random forests* (Breiman 2001), is easily parallelizable, and its performance in binary classification is comparable to that of Support Vector Machines (Caruana et al. 2008).

The algorithmic paradigm of *boosting*, studied by Schapire (1990), consists in gradually training diverse weak hypotheses by increasing the weight of previously misclassified examples. This paradigm gave rise to a practically useful algorithm, called AdaBoost (Freund and Schapire 1997), which is described in Algorithm 2. For convenience, the set of labels is here given by $\mathscr{Y} = \{-1, +1\}$. The AdaBoost algorithm maintains a probability distribution $\boldsymbol{p}_t$ over the training instances in S. Namely, on each round t, AdaBoost starts by training the weak learner A on the weighted dataset $(S_t, \boldsymbol{p}_t) = \{(\boldsymbol{x}_1, y_1, p_{t,1}), \ldots, (\boldsymbol{x}_m, y_m, p_{t,m})\}$. Next, the ensemble learner chooses a weight w_t for the weak hypothesis h_t, and then, updates the distribution $\boldsymbol{p}_t$ in a multiplicative way, where Z_t is the partition constant. A common choice for w_t is

$$w_t = \frac{1}{2} \ln\left(\frac{1}{\varepsilon_t} - 1\right) \quad \text{where} \quad \varepsilon_t = \sum_{i \in [m], h_t(\boldsymbol{x}_i) = y_i} p_{i,t}$$

Algorithm 1: Bagging

input: data set $S \in \mathscr{Z}^m$, number of rounds T, weak learner $A : \mathscr{Z}^m \to \mathscr{H}$
for $t = 1$ to T **do**
  build a sample S_t from S by drawing m instances with replacement
  run A on S_t to find a concept h_t in $\mathscr{H}$
end
output: $h(\boldsymbol{x}) = \frac{1}{T} \sum_{t=1}^{T} h_t(\boldsymbol{x})$

Algorithm 2: Boosting (AdaBoost)

input: data set $S \in \mathscr{Z}^m$, number of rounds T, weak learner $A : \mathscr{Z}^m \to \mathscr{H}$
initialize: set $p_{t,i} = \frac{1}{m}$ for each $i \in [m]$
for $t = 1$ to T **do**
  run A on $(S_t, \boldsymbol{p}_t)$ to find a concept h_t in $\mathscr{H}$
  choose w_t
  set $p_{t+1,i} = \frac{1}{Z_t} p_{t,i} \exp(-w_t h_t(\boldsymbol{x}_i))$
end
output: $h(\boldsymbol{x}) = \text{sign} \sum_{t=1}^{T} w_t h_t(\boldsymbol{x})$

The AdaBoost algorithm benefits from a solid theoretical analysis, surveyed in Schapire and Freund (2012). As a well-known result, let $\gamma \in (0, 1)$, and suppose that at each iteration of AdaBoost, the weak learner returns a hypothesis for which $\varepsilon_t \leq 1/2 - \gamma$. Then, the training error of the final hypothesis h returned by AdaBoost after T iterations is at most:

$$L_S(h) \leq \exp(-2\gamma^2 T)$$

From a practical viewpoint, the AdaBoost algorithm has been successfully applied to face recognition tasks, using axis-aligned rectangles for weak hypotheses (Viola and Jones 2001). Moreover, the boosting technique is particularly suited for learning linear combinations of decision rules (Cohen and Singer 1999; Schapire and Singer 1999), and alternating decision trees (Freund and Mason 1999).

Finally, it is important to emphasize that bagging and boosting are not limited to binary classification tasks. Notably, bagging and random forests have been applied to regression, density estimation, and manifold learning; a detailed survey can be found in Criminisi et al. (2012). The boosting technique has been extended to multi-class learning and ranking; see again (Schapire and Freund 2012) for a comprehensive survey about this paradigm.

6 Convex Learning

Convex learning problems cover a wide variety of learning tasks, where the hypothesis class is a convex set and the loss function is convex. Many, if not most, statistical learning problems which are easy to solve fall into this category. In this section, we first introduce some mathematical background about convex learning problems, next we examine several well-known algorithms for solving these problems, and then, we briefly survey the topic of Support Vector Machines which heavily relies on convex learning techniques.

6.1 Convex Learning Problems

Let $\mathscr{W}$ be a subset of an Euclidean space or, more generally, a Hilbert space. The, $\mathscr{W}$ is convex if for any two points $\boldsymbol{u}, \boldsymbol{w} \in \mathscr{W}$, and any scalar $\lambda \in (0, 1)$, the point formed by the convex combination $\lambda \boldsymbol{u} + (1 - \lambda)\boldsymbol{w}$ belongs to $\mathscr{W}$. By extension, a function $f : \mathscr{W} \to \mathbb{R}$ is convex if its epigraph $\{(\boldsymbol{w}, v) \mid v \geq f(\boldsymbol{w})\}$ is a convex set. For the sake of simplicity, we shall consider in this section that every convex function is differentiable, but most results can be extended to non-differentiable functions, using the notion of sub-differential (Hiriart-Urrut and Lemaréchal 2004; Rockafellar 1970). A real-valued, differentiable function f on $\mathscr{W}$ is convex if and only if, for any $\boldsymbol{u}, \boldsymbol{w} \in \mathscr{W}$,

$$f(\boldsymbol{u}) - f(\boldsymbol{w}) \geq \langle \nabla f(\boldsymbol{w}), \boldsymbol{u} - \boldsymbol{w} \rangle$$

Families of convex learning problems are typically characterized in terms of three basic properties about convex objectives. Namely, given a convex set $\mathscr{W}$ and three positive scalars ρ, α, and β, a convex function $f : \mathscr{W} \to \mathbb{R}$ is

- ρ-*Lipschitz* if for any $\boldsymbol{u}, \boldsymbol{w} \in \mathscr{W}$,

$$|f(\boldsymbol{u}) - f(\boldsymbol{w})| \leq \rho \, \|\boldsymbol{u} - \boldsymbol{w}\|$$

- α-*strongly convex* if for any $\boldsymbol{u}, \boldsymbol{w} \in \mathscr{W}$,

$$|f(\boldsymbol{u}) - f(\boldsymbol{w})| \geq \langle \nabla f(\boldsymbol{w}), \boldsymbol{u} - \boldsymbol{w} \rangle + \frac{\alpha}{2} \|\boldsymbol{u} - \boldsymbol{w}\|^2$$

- β-*smooth* if for any $\boldsymbol{u}, \boldsymbol{w} \in \mathscr{W}$,

$$|f(\boldsymbol{u}) - f(\boldsymbol{w})| \leq \langle \nabla f(\boldsymbol{w}), \boldsymbol{u} - \boldsymbol{w} \rangle + \frac{\beta}{2} \|\boldsymbol{u} - \boldsymbol{w}\|^2$$

Furthermore, given a positive scalar $B > 0$, we say that a convex set $\mathscr{W}$ is B-*bounded* if $\|\boldsymbol{w}\| \leq B$ for all $\boldsymbol{w} \in \mathscr{W}$.

Informally, the Lipschitzness property indicates that f cannot change too fast. A sufficient condition for this condition is that $\|\nabla f(\boldsymbol{w})\| \leq \rho$ for every $\boldsymbol{w} \in \mathscr{W}$. The properties of smoothness and strong convexity are related to the curvature of f. Notably, if f is twice-differentiable, then f is β-smooth and α-strongly convex if and only if, for every $\boldsymbol{w} \in \mathscr{W}$, we have:

$$\alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{w}) \preceq \beta \boldsymbol{I}$$

where $\boldsymbol{A} \preceq \boldsymbol{B}$ denotes the fact that $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite. In other words, the scalars α and β can be viewed as bounds on the eigenvalues of f. The ratio α/β is often referred to as the *condition number* of f.

Definition 7 (*Convex Learning*) Let $\mathscr{Z}$ be an instance space, $\mathscr{H}$ be a hypothesis class over $\mathscr{Z}$, and $\ell : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$ be a loss function. Then, $(\mathscr{Z}, \mathscr{H}, \ell)$ is a *convex learning problem* if $\mathscr{H}$ is representable by a convex set $\mathscr{W}$, and for every $z \in \mathscr{Z}$, the function $f : \mathscr{W} \to \mathbb{R}$ given by $f(\boldsymbol{w}) = \ell(h_{\boldsymbol{w}}, z)$ is convex.

For convex learning problems we shall replace the hypothesis class $\mathscr{H}$ by its convex representation class $\mathscr{W}$, and rewrite the loss function ℓ as a mapping from $\mathscr{W} \times \mathscr{Z}$. Based on the aforementioned properties about convex objectives, convex learning problems may be declined into several categories, depending on whether the loss function is Lipschitz, smooth, or strongly convex on its first argument. For example, consider the binary classification task defined over an instance space $\mathscr{Z} = \mathscr{X} \times \{-1, +1\}$, a convex representation class $\mathscr{W}$, and the hinge loss function:

$$\ell(\boldsymbol{w}, (\boldsymbol{x}, y)) = \max(0, 1 - y \langle \boldsymbol{w}, \boldsymbol{x} \rangle)$$

If the domain set is the ball $\mathscr{X} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq \rho\}$, then ℓ is both convex and ρ-Lipschitz. Now, if we use the same domain set, but replace the above loss function with the regularized hinge loss function:

$$\ell(\boldsymbol{w}, (\boldsymbol{x}, y)) = \max(0, 1 - y \langle \boldsymbol{w}, \boldsymbol{x} \rangle) + \frac{\alpha}{2} \|\boldsymbol{w}\|^2$$

it follows that ℓ is both α-strongly convex and ρ-Lipschitz. As another example, consider the regression task defined over an instance space $\mathscr{Z} = \mathscr{X} \times \mathbb{R}$, a convex representation class $\mathscr{W}$, and the square loss function

$$\ell(\boldsymbol{w}, (\boldsymbol{x}, y)) = (y - \langle \boldsymbol{w}, \boldsymbol{x} \rangle)^2$$

If the domain set is the ball $\mathscr{X} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq \beta/2\}$, then ℓ is both convex and β-smooth.

In general, a convex learning problem can be formulated as a *stochastic convex optimization* task of the form:

$$\begin{aligned} &\text{minimize} \quad L_D(\boldsymbol{w}) = \mathbb{E}_{z \sim \mathscr{D}}[\ell(\boldsymbol{w}, \boldsymbol{z})] \\ &\text{subject to} \quad \boldsymbol{w} \in \mathscr{W} \end{aligned} \tag{14}$$

where $\mathscr{W}$ is a convex set, and ℓ is convex on its first argument. We may attempt to solve this problem in a direct way, using a stochastic convex optimization algorithm that calls the example oracle $\text{EX}(\mathscr{D})$ for approximating the unknown objective L_D. Alternatively, we may rely on an indirect approach, by using learning rules defined over the empirical risk L_S, and described in Sect. 4. In the convex setting, Regularized Risk Minimization RRM is the paradigm of choice. Recall here that the RRM rule finds a minimizer of

$$\frac{1}{m} \sum_{i=1}^{m} \ell(\boldsymbol{w}, \boldsymbol{z}_i) + \text{reg}(\boldsymbol{w})$$

subject to $\boldsymbol{w} \in \mathscr{W}$, where $\text{reg} : \mathscr{W} \to \mathbb{R}_+$ is a regularization function. Namely, the next result established in Shalev-Shwartz et al. (2010), Shalev-Shwartz and Ben-David (2014) indicates that the RRM rule is stable for various families of convex learning problems.

Theorem 7 (Stability of RRM) *Let* $(\mathscr{Z}, \mathscr{W}, \ell)$ *be a convex learning problem. Then,*

- *the* RRM *rule with the Tikhonov regularizer* $\text{reg}(\boldsymbol{w}) = \lambda \|\boldsymbol{w}\|^2$ *is stable with rate* $O(1/m)$*, whenever* ℓ *is* ρ*-Lipschitz or* β*-smooth;*
- *the* ERM *rule (i.e.* RRM *with no regularizer) is stable with rate* $O(1/m)$*, whenever* ℓ *is* ρ*-Lipschitz and* α*-strongly convex.*

6.2 Convex Learning Algorithms

In the rich literature of convex optimization, a wide variety of algorithms have been devised for solving convex learning problems in a computationally efficient way. We invite the reader in browsing excellent textbooks about this active research topic (Bertsekas 2015; Boyd and Vandenberghe 2004; Bubeck 2015; Kushner and Yin 2010; Nesterov 2004; Nemirovski 1995; Sra et al. 2012). Here, we shall focus on three, well-studied convex learning algorithms: *Stochastic Gradient Descent* (SGD), *Stochastic Coodinate Descent* (SCD), and *Conditional Gradient* (CG).

6.2.1 Stochastic Gradient Descent

Arguably, the *Gradient Descent* algorithm is one of the oldest strategy for solving convex optimization problems (Cauchy 1847). The overall idea of this iterative optimization algorithm is to improve the solution at each iteration, by taking a step along the negative of the gradient of the function to be minimized at the current point. The stochastic version of this algorithm, which dates back to Robbins and Monro (1951), aims at minimizing a stochastic convex objective function of the form $L_D(\boldsymbol{w})$. To this end, SGD takes at each iteration a step along a random direction, for which the expectation is the negative of the gradient. As most convex learning problems are defined over a restricted subset $\mathscr{W}$ of an Euclidean or Hilbert space, the adaptation of SGD to statistical learning typically involves an additional *projection step*, which maintains the current point in the set of feasible solutions $\mathscr{W}$.

Algorithm 3: Stochastic Gradient Descent

input: scalar η, integer m
initialize: $\boldsymbol{v}_1 = \mathbf{0}$
for $t = 1$ to m **do**
  $\boldsymbol{w}_t = argmin_{\boldsymbol{w} \in \mathscr{W}} \ \|\boldsymbol{w} - \boldsymbol{v}_t\|^2$
  $\boldsymbol{z}_t = \text{EX}(\mathscr{D})$
  $\boldsymbol{v}_{t+1} = \boldsymbol{w}_t - \eta \nabla \ell(\boldsymbol{w}_t, \boldsymbol{z}_t)$
end
output: $\boldsymbol{w} = \frac{1}{m} \sum_{t=1}^{m} \boldsymbol{w}_t$

The resulting projected SGD method is described in Algorithm 3. At each iteration t, the algorithm first projects the current point $\boldsymbol{v}_t$ onto the representation class $\mathscr{W}$, next calls the example oracle for an instance $\boldsymbol{z}_t \in \mathscr{Z}$, and then performs a descent step using the gradient of the loss $\ell(\boldsymbol{w}_t, \boldsymbol{z}_t)$. The convergence of SGD has been analyzed for various families of objective functions (Kushner and Yin 2010; Rakhlin et al. 2012; Shalev-Shwartz et al. 2009; Shalev-Shwartz and Ben-David 2014). The next theorem summarizes convergence results obtained for the three aforementioned families.

Theorem 8 (Convergence of SGD) *Let* $(\mathcal{Z}, \mathcal{W}, \ell)$ *be a convex learning problem. Then, the SGD algorithm is*

- *universally consistent with rate* $O(1/\sqrt{m})$ *if* $\mathcal{W}$ *is* B*-bounded, and* ℓ *is* ρ*-Lipschitz or* β*-smooth.*
- *universally consistent with rate* $\tilde{O}(1/m)$ *if* $\mathcal{W}$ *is* B*-bounded, and* ℓ *is both* ρ*-Lipschitz and* α*-strongly convex.*

In other words, stochastic convex optimization problems of the form (14) can be solved directly, using the SGD algorithm, under reasonable assumptions about the representation class $\mathcal{W}$ and the loss function ℓ. The choice of the learning parameter η is governed by the input parameters defining the family of convex learning problems. For example, if $\mathcal{W}$ is B-bounded and ℓ is ρ-Lipschitz then, using $\eta = B/\rho\sqrt{m}$, the convergence rate is bounded by $B\rho/\sqrt{m}$. Thus, given a desired accuracy ε, it suffices to run SGD $m \geq (B\rho/\varepsilon)^2$ iterations in order to achieve, in expectation, a risk that is ε-close to the smallest risk.

The Gradient Descent method and its stochastic variant belong to the larger family of *Mirror Descent* algorithms (Nemirovski and Yudin 1983; Beck and Teboulle 2003), used to solve regularized risk minimization tasks for various regularization functions. The overall idea is to first map the current point $\boldsymbol{w}_t \in \mathcal{W}$ into the dual space $\mathcal{W}^*$, next perform the gradient descent in the dual space, and then mapping back the resulting point into the primal space. Various instances of Mirror Descent schemes include the *Exponentiated Gradient* algorithm (Kivinen and Warmuth 1997), and the *p-norm* algorithms (Gentile 2003). One of key geometric properties of Mirror Descent schemes is that an objective function f over the primal space $\mathcal{W}$ is α-strongly convex with respect to a norm $\|\cdot\|$ if and only if its conjugate f^* on the dual space $\mathcal{W}^*$ is $1/\alpha$-strongly smooth with respect to the dual norm $\|\cdot\|^*$ (Hiriart-Urrut and Lemaréchal 2004; Kakade et al. 2012). This, together with standard properties of Bregman divergences, typically yield convergence rates in $O(1/\sqrt{m})$ or $\tilde{O}(1/m)$ which depend only logarithmically in the dimension n of the data instances.

Algorithm 4: Stochastic Coordinate Descent

input: convex objective $L_S(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \ell(\boldsymbol{w}, \boldsymbol{z}_i)$
initialize: $\boldsymbol{w}_1 = \boldsymbol{0}$
for $t = 1$ to T **do**
  Choose index j uniformly at random in $[n]$
  Choose stepsize η_t
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \left| \frac{\partial L_S(\boldsymbol{w}_t)}{\partial j} \right| \boldsymbol{e}_j$
end
output: $\boldsymbol{w} = \boldsymbol{w}_T$

From a computational viewpoint, the main bottleneck of the SGD algorithm, and more generally Mirror Descent algorithms, lies in the projection step, which is a constrained convex optimization task performed *at each iteration*. For simple

representation classes $\mathscr{W}$, such as balls, hypercubes, simplices, and permutahedra, fast projection methods have been proposed (Duchi et al. 2008; Krichene et al. 2015; Lim and Wright 2016). However, for more complex representation classes, such as polyhedra described by linear inequalities, the projection step has to rely on general, time-consuming, convex optimization techniques. Circumventing this bottleneck by limiting the number of projection steps in gradient descent algorithms is a subject of ongoing research (Mahdavi et al. 2012; Zhang et al. 2013).

6.2.2 Stochastic Coordinate Descent

When the hypothesis class is a simple convex object, characterized by separable or block-separable constraints, the empirical risk can be minimized using the family of *Coordinate Descent* algorithms (Censor and Zenios 1997; Tseng and Yun 2009; Nesterov 2012; Wright 2015). Such methods, inspired from the Gauss-Seidel method for systems of linear equations, solve convex optimization tasks by iteratively performing approximate minimization along coordinate directions.

Algorithm 4 describes a stochastic version of Coordinate Descent for minimizing the empirical risk in the unconstrained setting (i.e. $\mathscr{W} = \mathbb{R}^n$). During each iteration t, the SCD algorithm first selects a coordinate j uniformly at random, and independently of past rounds, and then performs a descent according to the derivative of the empirical risk $L_S(\boldsymbol{w}_t)$ of the current point $\boldsymbol{w}_t$ at coordinate j. As detailed for instance in Wright (2015), the SCD algorithm may be easily upgraded to constrained versions of this task, using block-separable constraints.

Algorithm 5: Conditional Gradient

input: convex objective $L_S(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \ell(\boldsymbol{w}, z_i)$
initialize: $\boldsymbol{w}_1$ is an arbitrary point in $\mathscr{W}$
for $t = 1$ to T **do**
  $\boldsymbol{v}_t = argmin_{\boldsymbol{v}\in\mathscr{W}} \langle \nabla L_S(\boldsymbol{w}_t), \boldsymbol{v}\rangle$
  Choose stepsize $\eta_t \in (0, 1)$
  $\boldsymbol{w}_{t+1} = (1 - \eta_t)\boldsymbol{w}_t + \eta_t \boldsymbol{v}_t$
end
output: $\boldsymbol{w} = \boldsymbol{w}_T$

Although SCD is a fast, easy-to-implement algorithm, its convergence analysis requires more sophisticated conditions on the feasible set and the objective function (Nesterov 2012; Lu and Xiao 2015; Wright 2015). Notably, if L_S satisfies the property of coordinate-wise Lipschitz continuity with constants $\{\beta_j\}_{j=1}^n$, and the diameter of $\mathscr{W}$ with respect to the norm

$$\|\boldsymbol{w}\|_\beta = \sqrt{\sum_{j=1}^{n} \beta_j w_j^2}$$

is bounded by a constant R, then SCD converges to the empirical risk minimizer with rate in $O(1/T)$. Better convergence bounds may be achieved for strongly convex loss functions, or using accelerated versions of SCD.

6.2.3 Conditional Gradient

For hypothesis classes characterized by complex geometric objects, such as cones or polyhedra, convex projection tasks may be computationally demanding. Yet, *linear optimization* tasks on those objects are typically much easier. *Projection-free* algorithms constitute a family of convex optimization methods which replace the convex projection step with a cheaper linear optimization step (Clarkson 2010; Hazan and Kale 2012; Jaggi 2013; Lacoste-Julien and Jaggi 2015; Freund and Grigas 2016; Garber and Hazan 2016; Garber and Meshi 2016). The prototypical algorithm in this family is the *Conditional Gradient* method, due to Franck and Wolfe (1956).

Algorithm 5 describes a simple version of CG. During each iteration t, the algorithm starts by performing a linear optimization step using the gradient of the empirical risk of the current point $\boldsymbol{w}_t$, and then updates its solution according to a convex combination of $\boldsymbol{w}_t$ and the linear minimizer $\boldsymbol{v}_t$. Different strategies for choosing the stepsize η_t at each iteration are reported in Jaggi (2013), Freund and Grigas (2016). Apart from the choice of η_t, CG algorithms essentially differ in the linear optimization step. For example, a *local* linear optimization step is suggested in Garber and Hazan (2016), while *step-away* strategies are advocated in Lacoste-Julien and Jaggi (2015), Garber and Meshi (2016).

Overall, the performance of CG is relatively similar to the performance of SGD (for minimizing the empirical risk), using only linear optimization steps. Specifically, the convergence rate of CG is in

- $O(1/\sqrt{T})$ if $\mathscr{W}$ is B-bounded, and L_S is ρ-Lipschitz,
- $\tilde{O}(1/T)$ if $\mathscr{W}$ is B-bounded, and L_S is both ρ-Lipschitz and α-strongly convex.

We mention in passing that the SGD, SCD, and SG algorithms enjoy convergence rates in $O(\exp(-T))$ when the objective function is both smooth and strongly convex (Bubeck 2015).

6.3 Support Vector Machines

As mentioned in the introduction of this section, convex learning problems constitute the most important family of statistical learning problems where efficient learnability results can be obtained. It is therefore not surprising that convex learning algo-

rithms have been successfully applied to a wide range of statistical learning tasks. In particular, the key tools for handling high-dimensional learning tasks are *Support Vector Machines* (SVMs). Introduced in Boser et al. (1992), SVMs have been a subject of extensive research, both from a theoretical and practical perspective, summarized in various textbooks (Vapnik 1998; Cristianini and Shawe-Taylor 2000; Schölkopf and Smola 2002; Steinwart and Christmann 2008).

Support Vector Machines are defined through two main notions: *margins* and *kernels*. Intuitively, the notion of margin is related to the sample complexity of learning: SVMs handle high-dimensional hypothesis classes by searching for large margin separators. A linear classifier separates a training set with a large margin if it does not only classify examples in a correct way, but also pushes those examples away from the separating hyperplane. Thus, a large margin classifier may require a small sample complexity, even if the dimensionality of the feature space is high, or even infinite. The notion of kernel is related to the runtime complexity of learning. Basically, a kernel is a similarity measure between instances, which can be characterized as an inner product in some Hilbert space. For classifiers involving a feature expansion mapping, the "kernel trick" enables a computationally efficient implementation of learning, without explicitly handling the high dimensional feature expansion vector. Of course, the notions of margins and kernels are not limited to binary classification: they have been extended to various learning task including, for example, multi-class prediction and structured prediction.

There are two main categories of SVMs, depending on whether the training set supplied to the learner is assumed to be separable, or not. For the sake of simplicity, we focus here on zero-threshold linear functions, but the SVM rules defined below can easily be extended to non-homogeneous linear functions, using data points in the extended domain set $\mathscr{X} \times \{1\}$. A training set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ is linearly separable if there exists a vector $\boldsymbol{w}$ such that $y_i = \text{sign}\, \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$ for all $i \in [m]$. In this separable case, the margin of the hyperplane $\boldsymbol{w}$ with respect to the training set S is the minimal distance between an example in S and the hyperplane. In particular, if $\|\boldsymbol{w}\| = 1$, then the distance between $\boldsymbol{w}$ and any example $(\boldsymbol{x}_i, y_i)$ is simply given by $y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$. Therefore, the *Hard*-SVM rule is to find a separating hyperplane $\boldsymbol{w}$ with $\|\boldsymbol{w}\| = 1$ that maximizes the distance $\min_{i \in [m]} y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$. The Hard-SVM rule may be formulated in an equivalent way by the (constrained) convex optimization task:

$$\begin{aligned} &\text{minimize} \quad \|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq 1 \quad \forall i \in [m] \end{aligned} \tag{15}$$

If the training set S is linearly separable, then this optimization task is feasible. In this case, the solution $\boldsymbol{w}$ is normalized by $\|\boldsymbol{w}\|$ to yield the final predictor.

In the more general case where S is not linearly separable, the formulation (15) can be relaxed by allowing separability constraints to be violated by some examples. As usual, this may be formulated by adding slack variables $\xi_1, \ldots, \xi_m$, where each ξ_i captures by how much the the constraint $y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq 1$ is violated. The resulting *Soft*-SVM rule jointly minimizes the margin and the violations of separability constraints, using the following optimization task:

$$\begin{aligned} &\text{minimize} \quad \lambda \|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \\ &\text{subject to} \quad y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq 1 - \xi_i \quad \forall i \in [m] \end{aligned} \tag{16}$$

To this point, recall that the hinge loss between a linear model $\boldsymbol{w}$ and an example $(\boldsymbol{x}, y)$ is given by $\ell(\boldsymbol{w}, (\boldsymbol{x}, y)) = \max\{0, 1 - y \langle \boldsymbol{w}, \boldsymbol{x} \rangle\}$. With this formulation in hand, the Soft-SVM rule (16) can be expressed as a standard RRM task, given by

$$\min_{\boldsymbol{w}, b} \left(\sum_{i=1}^{m} \ell((\boldsymbol{w}, b), (\boldsymbol{x}_i, y_i)) + \lambda \|\boldsymbol{w}\|^2 \right) \tag{17}$$

This RRM objective is referred to as the *primal* formulation of the Soft-SVM rule. Since we are dealing with a convex optimization task, the Soft-SVM rule admits an equivalent *dual* formulation, where the optimal solution is characterized by a linear combination of examples in S, using Lagrangian variables $\alpha_1, \ldots, \alpha_m$:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \geq \mathbf{0}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle \right) \tag{18}$$

and the correspondence between (17) and (18) is given by $\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i$. If $\boldsymbol{w}$ is an optimal solution of (17) then the data points $\boldsymbol{x}_i$ for which α_i is positive are called the *support vectors* of $\boldsymbol{w}$.

Based on the above formulations, various convex optimization algorithms can be exploited for implementing linear Soft-SVMs. For example, (Shalev-Shwartz et al. 2007) solve the primal problem (17) using Stochastic Gradient Descent, and (Hsieh et al. 2008) solve the dual problem (18) using (dual) Coordinate Descent. The Conditional Gradient algorithm was also advocated for solving structured prediction tasks with SVMs (Lacoste-Julien et al. 2013).

Since the expressive power of linear functions is limited, a natural approach for extending SVMs to non-linear functions is to use a feature expander, that is, an embedding $\boldsymbol{\phi}$ of the domain set $\mathscr{X}$ onto some (possibly infinite dimensional) Hilbert space $\mathscr{F}$. Based on this feature expander, the hypothesis class $\mathscr{H}$ is represented by the set of vectors $\boldsymbol{w}$ such that $h_{\boldsymbol{w}}(\boldsymbol{x}) = \text{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle)$. Given an embedding $\boldsymbol{\phi}$, the corresponding *Kernel operator* is defined as

$$K(\boldsymbol{x}, \boldsymbol{x}') = \left\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \right\rangle$$

and the dual formulation (18) of Soft-SVM can be rewritten using the "kernel trick":

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \geq \mathbf{0}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \right) \tag{19}$$

By the Kernel Representer Theorem (Schölkopf et al. 2001), the optimal solution $\boldsymbol{w}$ can be expressed as a linear combination of expanded points, that is, $\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i \boldsymbol{\phi}(x_i)$. Since the dimension of $\boldsymbol{w}$ can be large or infinite, the kernel trick allows us to efficiently encode $\boldsymbol{w}$ as a set of support vectors x_i, each associated with its coefficient α_i. Furthermore, since the kernel operator K associated with a feature expander $\boldsymbol{\phi}$ defines a positive semidefinite matrix, the kernelized SVM rule (19) is a concave optimization problem. Again, convex optimization algorithms can be advocated for efficiently solving this task, provided that the kernel operator K can be computed in polynomial time. Various kernels satisfying this condition have been proposed in the literature, and we refer the reader to Herbrich (2002), Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004), Bottou (2007), Kung (2014), for detailed surveys about kernel methods.

Finally, as kernels provide a way to express prior knowledge about the learning task at hand, an important subject of ongoing research in SVMs is to *learn kernels*, using a kernel family (Lanckriet et al. 2004; Bach 2008; Cortes et al. 2009, 2010).

7 Conclusion

In this chapter, we have drawn a conceptual map of statistical computational learning by providing answers to several questions: what is a statistical learning problem? How to measure the performance of a learning algorithm? Which are the main optimization principles in statistical learning? And, under which conditions a hypothesis class is learnable? Based on these foundations, we have surveyed two important problems in Machine Learning: concept learning and convex learning. In this concluding section, we highlight several topics of research at the intersection of statistical computational learning and AI. Due to space reasons, the list is by no means exhaustive, and we apologize for omitting other topics of interest.

Learning Sparse Models The concept of sparsity is ubiquitous in many scientific and engineering applications, for identifying parsimonious solutions to high-dimensional problems. Informally, a sparse solution can be viewed as a high-dimensional vector or matrix satisfying some *sparsity constraint*, which limits the degrees of freedom of the model. Various sparsity constraints have been proposed in the literature of machine learning and signal processing, ranging from the standard cardinality constraint that restrains the number of nonzero coordinates (Shalev-Shwartz et al. 2010), to more sophisticated sparsity constraints which impose a low-dimensional structure on the set of nonzero features (Hegde et al. 2015; Jain et al. 2016). For example, in the "group-structured" sparsity constraint, the relevant features are partitioned into a small number of contiguous blocks, and in the "tree-structured" sparsity constraint, such features are arranged into a connected acyclic graph. As convex optimization under sparsity constraints is NP-hard (Natarajan 1995), two main approaches have been advocated for learning sparse models: convex relaxation (Shalev-Shwartz and Tewari 2011; Bach et al. 2012), and

approximation algorithms (Bahmani et al. 2013; Jain et al. 2014). A recent survey on sparse modelling and learning can be found in Rish and Grabarnik (2014).

Learning Probabilistic Models In statistical learning, probabilistic models aim at estimating the hidden distribution that generates data instances. Of particular interest in AI are probabilistic graphical models which are able to represent high-dimensional probability distributions (Koller and Friedman 2009; Murphy 2012). As explained in Sect. 2, a probabilistic graphical model is a pair $(G, \boldsymbol{\theta})$, where G is a graphical structure and $\boldsymbol{\theta}$ is a vectorized set of parameters. *Parameter learning* is the task of estimating from data the parameters of a probabilistic model, when the structure is fixed. Correspondingly, *structure learning* is the problem of extracting the graphical structure of a probabilistic model, given a class of candidate structures, such as directed acyclic graphs for Bayesian networks, or hypertrees for bounded-treewidth Markov networks. Various algorithms have been proposed for estimating parameters under the (possibly regularized) log-likelihood loss function. In particular, *Expectation Minimization* (EM) (Dempster et al. 1977) is the prototypical algorithm for estimating parameters in presence of missing values (Lauritzen 1995). A comprehensive treatment of the subject is given in Little and Rubin (2014).

Structure learning is arguably more challenging, since the corresponding regularized risk minimization task involves combinatorial constraints capturing admissible graphical structures. Although structure learning is tractable for tree-directed models (Chow and Liu 1968) and their mixtures (Meila and Jaakkola 2006), the problem is NP-hard for more expressive models, such as Bayesian networks (Chickering 1996), Bayesian polytrees (Dasgupta 1999), and bounded-treewidth Markov networks (Srebro 2003). For this reason, structure learning is an active research topic relying on both statistical and combinatorial methods. Notably, (Cussens 2011; Kumar and Bach 2013; Nie et al. 2014; Bartlett and Cussens 2017) use Integer Linear Programming techniques for learning the structure of Bayesian networks or Markov networks with bounded-treewidth. SAT and CSP techniques have also been proposed for solving these structure learning problems (Cussens 2008; Berg et al. 2014; van Beek and Hoffmann 2015).

Learning Preference Models The spectrum of applications that resort on the ability to learn preferences is extremely wide, ranging from configuration softwares and recommender systems to information retrieval and group decision-making (see e.g. chapter "Compact Representation of Preferences" of Volume 1). It is therefore not surprising that topic of preference learning has gained a considerable interest in statistical and computational learning. As explained in Sect. 2, preference learning problems can be divided into several categories, depending on the type of reference set, the type of preference relation, the examples provided to the learner and, of course, the class of preference models.

In *label ranking* (Vembu and Gärtner 2010), the problem is to associate instances with a total order of predefined labels. With each training instance, we receive supervision given as a binary relation on the labels. More formally, the instance space is given $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$, where $\mathscr{X}$ is the domain set, and $\mathscr{Y}$ is the space of all directed acyclic graphs over the set of labels $[k]$. The goal is to learn from a training set S a

hypothesis $h : \mathscr{X} \to \mathscr{Y}^\dagger$ in the available hypothesis class $\mathscr{H}$ that minimizes some loss function $\ell : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$. For total ranking tasks, $\mathscr{Y}^\dagger$ is the group of permutations over $[k]$, and for more general ranking tasks, $\mathscr{Y}^\dagger$ is a subset of $\mathscr{Y}$. Several families of label ranking problems can be solved by reduction to binary classification (Hüllermeier et al. 2008), boosting (Dekel et al. 2003), multi-label classification (Crammer and Singer 2003), ordinal regression (Herbrich et al. 2000), or regularized least-square minimization (Gärtner and Vembu 2009).

In *object ranking* (Kamishima et al. 2010), $\mathscr{X}$ is a set of objects, and $\mathscr{Y}$ is a space of total rankings (permutations) or partial rankings (DAGs) over $\mathscr{X}$. Each training instance is formed by a pair, or more generally a set, of objects in $\mathscr{X}$, and the supervision is given by a preference ordering on these objects. The goal is to learn a hypothesis $h : \mathscr{X} \to \mathscr{Y}$, chosen from a class $\mathscr{H}$ that minimizes some loss function ℓ. Again, various statistical learning techniques have been successfully applied for solving tractable object ranking problems. They include, among others, boosting methods (Freund et al. 2003; Xu and Li 2007), and SVMs (Joachims 2002; Kazawa et al. 2005; Cao et al. 2006).

Ranking tasks are intrinsically related to preference aggregation problems. Notably, the problem of finding a total ranking of objects minimizing a pairwise loss function is generally NP-hard (Cohen et al. 1999; Alon 2006). The difficulty is even more accute when the hypothesis class is a Mallows model or an exponential family (Vembu et al. 2009; Lu and Boutilier 2014).

Learning Neural Models As mentioned in Sect. 2, neural models and Machine Learning have a long shared history, dating back to Rosenblatt's invention of the Perceptron algorithm (Rosenblatt 1958). Neural networks were extensively studied in the 1980s, but with mixed empirical results. During this past decade, a combination of algorithmic advances in Machine Learning, together with increasing computational power and data size, has led to a breakthrough in the effectiveness of deep neural networks (Goodfellow et al. 2016). In particular, the families of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown impressive performance on a variety of application domains, including computer vision (LeCun et al. 2010; Krizhevsky et al. 2012; Pinheiro and Collobert 2014), speech recognition (Hinton et al. 2012; Graves et al. 2013), and natural language processing (Collobert and Weston 2008; Cho et al. 2014; Kalchbrenner et al. 2014). In the present book, deep neural networks are discussed in chapter "Reinforcement Learning" of Volume 1, and chapter "Designing Algorithms for Machine Learning and Data Mining" of Volume 2.

Despite the undoubled practical success of deep learning, there are many open theoretical questions related to this fascinating subject of research. As discussed in Sect. 5, intersections of separating hyperplanes over $\{0, 1\}^n$ are not efficiently PAC learnable for the zero-one loss (Klivans and Sherstov 2009). This implies that no efficient algorithm can be found for training neural networks, even if we allow additional layers or effective activation functions. For other loss functions advocated in deep learning, the corresponding optimization task remains highly non-convex, and hence, generally intractable. So, there is a fundamental gap between the theory

of statistical computational learning and the practical efficiency of deep learning, achieved by gradient-based methods with backpropagation (Rumelhart et al. 1986). Recent investigations in the theoretical analysis of deep models have attempted to bridge this gap (Kawaguchi 2016; Bach 2017; Kawaguchi et al. 2017; Shalev-Shwartz et al. 2017; Song et al. 2017; Zhang et al. 2017), but much remains to be done before having a comprehensive analysis of practical results.

References

Aggarwal C, Reddy C (2013) Data clustering: algorithms and applications. Taylor and Francis
Alekhnovich M, Braverman M, Feldman V, Klivans A, Pitassi T (2008) The complexity of properly learning simple concept classes. J Comput Syst Sci 74(1):16–34
Alon N (2006) Ranking tournaments. SIAM J Discret Math 20(1):137–142
Alon N, Ben-David S, Cesa-Bianchi N, Haussler D (1997) Scale-sensitive dimensions, uniform convergence, and learnability. J ACM (JACM) 44(4):615–631
Alpaydin E (2009) Introduction to machine learning. MIT, USA
Angluin D, Laird PD (1987) Learning from noisy examples. Mach Learn 2(4):343–370
Anthony M (2001) Discrete mathematics of neural networks: selected topics. SIAM monographs on. discrete mathematics and applications
Anthony M (2010) Probabilistic learning and boolean functions. In: Crama Y, Hammer P (eds) Boolean models and methods in mathematics, computer science, and engineering, encyclopedia of mathematics and its applications. Cambridge University, Cambridge, pp 197–220
Anthony M, Barlett P (1999) Neural network learning: theoretical foundations. Cambridge University, Cambridge
Anthony M, Biggs N (1997) Computational learning theory. Cambridge University, Cambridge
Anthony M, Shawe-Taylor J (1993) Using the perceptron algorithm to find consistent hypotheses. Comb, Probab Comput 2:385–387
Arora S, Babai L, Stern J, Sweedyk Z (1997) The hardness of approximate optima in lattices, codes, and systems of linear equations. J Comput Syst Sci 54(2):317–331
Bach FR (2008) Exploring large feature spaces with hierarchical multiple kernel learning. In: Advances in neural information processing systems 21 (NIPS 2008), pp 105–112
Bach FR (2017) Breaking the curse of dimensionality with convex neural networks. J Mach Learn Res 18:19:1–19:53
Bach FR, Jenatton R, Mairal J, Obozinski G (2012) Optimization with sparsity-inducing penalties. Found Trends Mach Learn 4(1):1–106
Bahmani S, Raj B, Boufounos P (2013) Greedy sparsity-constrained optimization. J Mach Learn Res 14:807–841
Bartlett M, Cussens J (2017) Integer linear programming for the bayesian network structure learning problem. Artif Intell 244:258–271
Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper Res Lett 31(3):167–175
Ben-David S, Eiron N, Long PM (2003) On the difficulty of approximately maximizing agreements. J Comput Syst Sci 66(3):496–514
Berg J, Järvisalo M, Malone B (2014) Learning optimal bounded treewidth bayesian networks via maximum satisfiability. In: Proceedings of the 17th international conference on artificial intelligence and statistics (AISTATS 2014), pp 86–95
Bertsekas D (2015) Convex optimization algorithms. MIT, USA
Birge J, Louveaux F (2011) Introduction to stochastic programming. Springer, Berlin
Bishop C (2006) Pattern recognition and machine learning. Springer, Berlin

Blum A (1992) Rank-r decision trees are a subclass of r-decision lists. Inf Process Lett 42(4): 183–185
Blum A, Rivest RL (1992) Training a 3-node neural network is NP-complete. Neural Netw 5(1): 117–127
Blumer A, Ehrenfeucht A, Haussler D, Warmuth M (1989) Learnability and the Vapnik-Chervonenkis dimension. J ACM (JACM) 36(4):929–965
Boser BE, Guyon L, Vapnik V (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the 5th annual ACM conference on computational learning theory (COLT 1992), pp 144–152
Bottou L (2007) Large-scale kernel machines, Neural information processing series. MIT, USA
Bousquet O, Elisseeff A (2002) Stability and generalization. J Mach Learn Res 2(Mar):499–526
Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University, Cambridge
Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
Breiman L (2001) Random forests. Mach Learn 45(1):5–32
Bubeck S (2015) Convex optimization: algorithms and complexity. Found Trends Mach Learn 8(3–4):231–358
Cao Y, Xu J, Liu T, Li H, Huang Y, Hon H (2006) Adapting ranking SVM to document retrieval. In: Proceedings of the 29th annual international ACM conference on research and development in information retrieval (SIGIR 2006), pp 186–193
Caruana R, Karampatziakis N, Yessenalina A (2008) An empirical evaluation of supervised learning in high dimensions. In: Proceedings of the 25th international conference on machine learning (ICML 2008), pp 96–103
Cauchy A (1847) Méthode générale pour la résolution des systèmes d'équations simultanées. C. R. Acad. Sci. Paris 25:536–538
Censor Y, Zenios SA (1997) Parallel optimization. Oxford University, Oxford
Chickering DM (1996) Learning Bayesian networks is NP-complete. In: Learning from data: artificial intelligence and statistics V, Springer, Berlin, pp 121–130
Cho K, van Merrienboer B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP 2014), pp 1724–1734
Chow CK, Liu CN (1968) Approximating discrete probability distributions with dependence trees. IEEE Trans Inf Theory 14(3):462–467
Clarkson KL (2010) Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. ACM Trans Algorithms 6(4):63:1–63:30
Clémençon S, Vayatis N (2007) Ranking the best instances. J Mach Learn Res 8:2671–2699
Cohen W, Schapire R, Singer Y (1999) Learning to order things. J Artif Intell Res (JAIR) 10:243–270
Cohen WW, Singer Y (1999). A simple, fast, and effictive rule learner. In: Proceedings of the 16th national conference on artificial intelligence (AAAI 1999), pp 335–342
Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the twenty-fifth international conference in machine learning (ICML 2008), pp 160–167
Cortes C, Mohri M, Rostamizadeh A (2009) Learning non-linear combinations of kernels. In: Advances in neural information processing systems 22 (NIPS 2009), pp 396–404
Cortes C, Mohri M, Rostamizadeh A (2010) Generalization bounds for learning kernels. In: Proceedings of the 27th international conference on machine learning (ICML 2010), pp 247–254
Crammer K, Singer Y (2003) A family of additive online algorithms for category ranking. J Mach Learn Res 3:1025–1058
Criminisi A, Shotton J, Konukoglu E (2012) Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found Trends Comput Graph Vis 7(2–3):81–227
Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University, Cambridge

Cussens J (2008) Bayesian network learning by compiling to weighted MAX-SAT. In: Proceedings of the 24th conference in uncertainty in artificial intelligence (UAI 2008), pp 105–112
Cussens J (2011) Bayesian network learning with cutting planes. In: Proceedings of the 27th conference on uncertainty in artificial intelligence (UAI 2011), pp 153–160
Daniely A, Shalev-Shwartz S (2016) Complexity theoretic limitations on learning dnf's. In: Proceedings of the 29th conference on learning theory (COLT 2016), pp 815–830
Darwiche A (2009) Modeling and reasoning with bayesian networks. Cambridge University, Cambridge
DasGupta A (2011) Probability for statistics and machine learning: fundamentals and. advanced topics. Springer, Berlin
Dasgupta S (1999) Learning polytrees. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence (UAI 1999), pp 134–141
De Raedt L (2008) Logical and relational learning. Springer, Berlin
Dekel O, Manning CD, Singer Y (2003) Log-linear models for label ranking. In: Advances in neural information processing systems 16 (NIPS 2003), pp 497–504
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Society Ser B (Methodological) 39(1):1–38
Devroye L, Györfi L, Lugosi G (2013) A probabilistic theory of pattern recognition. Springer, Berlin
Du K-L, Swamy MNS (2013) Neural networks and statistical learning. Springer, Berlin
Duchi JC, Shalev-Shwartz S, Singer Y, Chandra T (2008) Efficient projections onto the l_1-ball for learning in high dimensions. In: Machine learning, proceedings of the twenty-fifth international conference (ICML 2008), pp 272–279
Engel A, Broeck C (2001) Statistical mechanics of learning. Cambridge University, Cambridge
Feldman V, Gopalan P, Khot S, Ponnuswami AK (2009) On agnostic learning of parities, monomials, and halfspaces. SIAM J Comput 39(2):606–645
Feldman V, Guruswami V, Raghavendra P, Wu Y (2012) Agnostic learning of monomials by halfspaces is hard. SIAM J Comput 41(6):1558–1590
Flach P (2012) Machine learning: the art and science of algorithms that make sense of data. Cambridge University, Cambridge
Fligner MA, Verducci JS (1986) Distance based ranking models. J R Stat Soc 48(3):359–369
Franck M, Wolfe P (1956) An algorithm for quadratic programming. Naval Res Logis Q 3:95–110
Freund RM, Grigas P (2016) New analysis and results for the Frank-Wolfe method. Math Program 155(1–2):199–230
Freund Y, Iyer RD, Schapire RE, Singer Y (2003) An efficient boosting algorithm for combining preferences. J Mach Learn Res 4:933–969
Freund Y, Mason L (1999) The alternating decision tree learning algorithm. In: Proceedings of the 16th international conference on machine learning (ICML 1999), pp 124–133
Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
Fürnkranz J, Hüllermeier E (2010) Preference learning. Springer, Berlin
Garber D, Hazan E (2016) A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. SIAM J Optim 26(3):1493–1528
Garber D, Meshi O (2016) Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes. In: Advances in neural information processing systems 29 (NIPS 2016), pp 1001–1009
Gärtner T, Vembu S (2009) On structured output training: hard cases and an efficient alternative. Mach Learn 76(2–3):227–242
Gentile C (2003) The robustness of the p-norm algorithms. Mach Learn 53(3):265–299
Getoor L, Taskar B (2007) Introduction to statistical relational learning. MIT, Cambridge
Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT, USA

Graves A, Mohamed A, Hinton GE (2013) Speech recognition with deep recurrent neural networks. In: IEEE international conference on acoustics, speech and signal processing (ICASSP 2013), pp 6645–6649

Grünwald P (2007) The minimum description length principle. MIT, USA

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin

Haussler D (1988) Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. Artif Intell 36(2):177–221

Haussler D (1992) Decision theoretic generalizations of the PAC model for neural net and other learning applications. Inf Comput 100(1):78–150

Hazan E, Kale S (2012) Projection-free online learning. In: Proceedings of the 29th international conference on machine learning (ICML 2012)

Hegde C, Indyk P, Schmidt L (2015) A nearly-linear time framework for graph-structured sparsity. In: Proceedings of the 32nd international conference on machine learning (ICML 2015), pp 928–937

Helmbold DP, Sloan RH, Warmuth MK (1992) Learning integer lattices. SIAM J Comput 21(2): 240–266

Herbrich R (2002) Learning kernel classifiers: theory and algorithms. MIT, USA

Herbrich R, Graepel T, Obermayer K (2000) Large margin rank boundaries for ordinal regression. In: Advances in large margin classifiers. MIT Press, USA, pp 115–132

Hinton G, Deng L, Yu D, Dahl GE, r. Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97

Hiriart-Urrut JB, Lemaréchal C (2004) Fundamentals of convex analysis. Springer, Berlin

Höffgen K, Simon HU (1992) Robust trainability of single neurons. In: Proceedings of the fifth annual acm conference on computational learning theory (COLT 1992), pp 428–439

Hsieh C, Chang K, Lin C, Keerthi SS, Sundararajan S (2008) A dual coordinate descent method for large-scale linear SVM. In: Proceedings of the 25th international conference on machine learning, pp 408–415

Hüllermeier E, Fürnkranz J, Cheng W, Brinker K (2008) Label ranking by learning pairwise preferences. Artif Intell 172(16–17):1897–1916

Jaggi M (2013) Revisiting frank-wolfe: projection-free sparse convex optimization. In: Proceedings of the 30th international conference on machine learning (ICML 2013), pp 427–435

Jain P, Rao N, Dhillon I (2016) Structured sparse regression via greedy hard thresholding. In: Advances in neural information processing systems 29 (NIPS 2016), pp 1516–1524

Jain P, Tewari A, Kar P (2014) On iterative hard thresholding methods for high-dimensional M-estimation. In: Advances in neural information processing systems 27 (NIPS 2014), pp 685–693

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning: with applications in R. Springer texts in statistics, Springer, New York

Joachims T (2002) Optimizing search engines using clickthrough data. In: Proceedings of the 8th ACM international conference on knowledge discovery and data mining (SIGKDD 2002), pp 133–142

Johnson DS, Preparata FP (1978) The densest hemisphere problem. Theorertical Comput Sci 6:93–107

Kakade SM, Shalev-Shwartz S, Tewari A (2012) Regularization techniques for learning with matrices. J Mach Learn Res 13:1865–1890

Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL 2014), pp 655–665

Kamishima T, Kazawa H, Akaho S (2010) A survey and empirical comparison of object ranking methods. Preference learning. Springer, Berlin, pp 181–201

Kawaguchi K (2016) Deep learning without poor local minima. In: Advances in neural information processing systems 29 (NIPS 2016), pp 586–594

Kawaguchi K, Kaelbling LP, Bengio Y (2017) Generalization in deep learning. CoRR. arXiv:1710.05468
Kazawa H, Hirao T, Maeda E (2005) Order SVM: a kernel method for order learning based on generalized order statistics. Syst Comput Jpn 36(1):35–43
Kearns M, Li M (1993) Learning in the presence of malicious errors. SIAM J Comput 22(4):807–837
Kearns M, Li M, Pitt L, Valiant L (1987) Recent results on boolean concept learning. In: Proceedings of the fourth international workshop on machine learning (ICML 1987), pp 337–352
Kearns M, Li M, Valiant LG (1994a) Learning boolean formulas. J. ACM 41(6):1298–1328
Kearns M, Schapire R, Sellie L (1994b) Toward efficient agnostic learning. Mach Learn 17(2): 115–141
Kearns M, Vazirani U (1994) An introduction to computational learning theory. MIT, USA
Kivinen J, Warmuth MK (1997) Exponentiated gradient versus gradient descent for linear predictors. Inf Comput 132(1):1–63
Klivans AR, O'Donnell R, Servedio RA (2004) Learning intersections and thresholds of halfspaces. J Comput Syst Sci 68(4):808–840
Klivans AR, Servedio RA (2004) Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. J Comput Syst Sci 68(2):303–318
Klivans AR, Sherstov AA (2009) Cryptographic hardness for learning intersections of halfspaces. J Comput Syst Sci 75(1):2–12
Koller D, Friedman N (2009) Probabilistic graphical models. MIT, USA
Krichene W, Krichene S, Bayen AM (2015) Efficient bregman projections onto the simplex. In: Proceedings of the 54th IEEE conference on decision and control, (CDC 2015), pp 3291–3298
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25 (NIPS 2012), pp 1106–1114
Kulkarni S, Harman G (2011) An elementary introduction to statistical learning theory. Wiley series in probability and statistics, Wiley, New York
Kumar KSS, Bach FR (2013) Convex relaxations for learning bounded-treewidth decomposable graphs. In: Proceedings of the 30th international conference on machine learning (ICML 2013), pp 525–533
Kung S (2014) Kernel methods and machine learning. Cambridge University, Cambridge
Kushner HJ, Yin GG (2010) Stochastic approximation and recursive algorithms and applications. Springer, Berlin
Lacoste-Julien S, Jaggi M (2015) On the global linear convergence of frank-wolfe optimization variants. In: Advances in neural information processing systems 28 (NIPS 2015), pp 496–504
Lacoste-Julien S, Jaggi M, Schmidt MW, Pletscher P (2013) Block-coordinate frank-wolfe optimization for structural SVMs. In: Proceedings of the 30th international conference on machine learning (ICML 2013), pp 53–61
Lanckriet GRG, Cristianini N, Bartlett PL, Ghaoui LE, Jordan MI (2004) Learning the kernel matrix with semidefinite programming. J Mach Learn Res 5:27–72
Lauritzen SL (1995) The em algorithm for graphical association models with missing data. Comput Stat Data Anal 19(2):191–201
Lebanon G, Lafferty J (2002) Conditional models on the ranking poset. In: Advances in neural information processing systems 15 (NIPS 2002), pp 415–422
LeCun Y, Kavukcuoglu K, Farabet C (2010) Convolutional networks and applications in vision. In: Proceedings of the international symposium on circuits and systems (ISCAS 2010), pp 253–256
Lim CH, Wright SJ (2016) Efficient bregman projections onto the permutahedron and related polytopes. In: Proceedings of the 19th international conference on artificial intelligence and statistics (AISTATS 2016), pp 1205–1213
Little R, Rubin D (2014) Statistical analysis with missing data. Wiley, New York
Liu T, Lugosi G, Neu G, Tao D (2017) Algorithmic stability and hypothesis complexity. In: Proceedings of the 34th international conference on machine learning (ICML 2017), pp 2159–2167
Lu T, Boutilier C (2014) Effective sampling and learning for Mallows models with pairwise-preference data. J Mach Learn Res 15(1):3783–3829

Lu Z, Xiao L (2015) On the complexity analysis of randomized block-coordinate descent methods. Math Program 152(1–2):615–642
Ma Y, Fu Y (2011) Manifold learning theory and applications. CRC
Mahdavi M, Yang T, Jin R, Zhu S, Yi J (2012) Stochastic gradient descent with only one projection. In: Advances in neural information processing systems 25 (NIPS 2012), pp 503–511
Mallows CL (1957) Non-null ranking models. Biometrika 44(1–2):114–130
Megiddo N (1988) On the complexity of polyhedral separability. Discret Comput Geom 3(4): 325–337
Meila M, Chen H (2010) Dirichlet process mixtures of generalized mallows models. In: Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence (UAI 2010), pp 358–367
Meila M, Jaakkola TS (2006) Tractable bayesian learning of tree belief networks. Stat Comput 16(1):77–92
Mitchell T (1982) Generalization as search. Artif Intell 18(2):203–226
Mitchell T (1997) Machine learning. McGraw-Hill Education
Mohammadi L, van de Geer S (2005) Asymptotics in empirical risk minimization. J Mach Learn Res 6:2027–2047
Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. MIT, USA
Mukherjee S, Niyogi P, Poggio T, Rifkin R (2006) Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. Adv Comput Math 25(1):161–193
Murphy K (2012) Machine learning: a probabilistic perspective. MIT, USA
Natarajan B (1991) Machine learning: a theoretical approach. M. Kaufmann Publishers
Natarajan B (1995) Sparse approximate solutions to linear systems. SIAM J Comput 24(2):227–234
Nemirovski A (1995) Efficient methods in convex programming. http://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf
Nemirovski AS, Yudin DB (1983) Problem complexity and method efficiency in optimization. J. Wiley, New York
Nesterov Y (2004) Introductory lectures on convex optimization: a basic course. Kluwer Academic Publishers
Nesterov Y (2012) Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J Optim 22(2):341–362
Nie S, Mauá DD, de Campos CP, Ji Q (2014) Advances in learning bayesian networks of bounded treewidth. In: Advances in neural information processing systems 27 (NIPS 2014), pp 2285–2293
Parberry I (1994) Circuit complexity and neural networks. MIT, USA
Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo
Pinheiro PHO, Collobert R (2014) Recurrent convolutional neural networks for scene labeling. In: Proceedings of the 31th international conference on machine learning (ICML 2014), pp 82–90
Pitt L, Valiant L (1988) Computational limitations on learning from examples. J ACM 35(4):965–984
Plackett RL (1975) The analysis of permutations. J R Stat Soc 24(10):193–202
Poffio T, Rifkin R, Kukherjee S, Niyogi P (2004) General conditions for predictivity in learning theory. Nature 428(6981):419
Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
Quinlan JR (1993) C4. 5: Programs for machine learning. Morgan Kaufmann
Quinlan JR (1996) Bagging, boosting, and C4.5. In: Proceedings of the 30th national conference on artificial intelligence (AAAI 1996), pp 725–730
Rakhlin A, Shamir O, Sridharan K (2012) Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th international conference on machine learning (ICML 2012)
Rish I, Grabarnik G (2014) Sparse modeling: theory, algorithms, and applications. CRC
Rissanen J (1983) A universal prior for integers and estimation by minimum description length. Ann stat 416–431

Rissanen J (1985) Minimum description length principle. Wiley, New York
Rivest RL (1987) Learning decision lists. Mach Learn 2(3):229–246
Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22(3):400–407
Rockafellar T (1970) Convex analysis. Princeton University, Princeton
Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–408
Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Parallel distributed processing: explorations in the microstructure of cognition, vol 1. MIT, USA, pp 318–362
Sauer N (1972) On the density of families of sets. J Comb Theory 13:145–147
Schapire RE (1990) The strength of weak learnability. Mach Learn 5:197–227
Schapire RE, Freund Y (2012) Boosting. MIT, USA
Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. Mach Learn 37(3):297–336
Schölkopf B, Herbrich R, Smola AJ (2001) A generalized representer theorem. In: Proceedings of the 14th annual conference on computational on computational learning theory (COLT 2001), pp 416–426
Schölkopf B, Smola A (2002) Learning with Kernels: support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning, MIT, USA
Shalev-Shwartz S, Ben-David S (2014) Understanding machine learning: from theory to algorithms. Cambridge University, Cambridge
Shalev-Shwartz S, Shamir O, Shammah S (2017) Failures of gradient-based deep learning. In: Proceedings of the 34th international conference on machine learning (ICML 2017), pp 3067–3075
Shalev-Shwartz S, Shamir O, Srebro N, Sridharan K (2009) Stochastic convex optimization. In: Proceedings of the 22nd conference on learning theory (COLT 2009), pp 177–186
Shalev-Shwartz S, Shamir O, Srebro N, Sridharan K (2010) Learnability, stability and uniform convergence. J Mach Learn Res 11:2635–2670
Shalev-Shwartz S, Singer Y, Srebro N (2007) Pegasos: primal estimated sub-gradient solver for SVM. In: Proceedings of the 24th international conference on machine learning (ICML 2007), pp 807–814
Shalev-Shwartz S, Srebro N, Zhang T (2010) Trading accuracy for sparsity in optimization problems with sparsity constraints. SIAM J Optim 20(6):2807–2832
Shalev-Shwartz S, Tewari A (2011) Stochastic methods for l_1-regularized loss minimization. J Mach Learn Res 12:1865–1892
Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Kernel methods for pattern analysis, Cambridge University, Cambridge
Song L, Vempala S, Wilmes J, Xie B (2017) On the complexity of learning neural networks. CoRR. arXiv:1707.04615
Sra S, Nowozin S, Wright S (2012) Optimization for machine learning. Neural information processing series, MIT, USA
Srebro N (2003) Maximum likelihood bounded tree-width Markov networks. Artif Intell 143(1):123–138
Sridharan K (2012) Learning from an optimization viewpoint. Ph.D. thesis, Technicological Institute of Chicago, Toyota
Steinwart I, Christmann A (2008) Support vector machines. Information science and statistics, Springer, Berlin
Sugiyama M (2015) Introduction to statistical machine learning. Elsevier Science
Theodoridis S (2015) Machine learning: a bayesian and optimization perspective. Elsevier Science
Tikhonov A (1943) On the stability of inverse problems. Doklady Akademii Nauk SSSR 39(5):195–198
Tseng P, Yun S (2009) A coordinate gradient descent method for nonsmooth separable minimization. Math Program 117(1–2):387–423

Turing A (1950) Computing machinery and intelligence. Mind 59:433–460
Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142
van Beek P, Hoffmann H (2015) Machine learning of bayesian networks using constraint programming. In: Proceedings of the 21st confernce on principles and practice of constraint programming (CP 2015), pp 429–445
Vapnik V (1998) Statistical learning theory. Wiley, New York
Vapnik V (2013) The nature of statistical learning theory, 3rd edn. Springer, Berlin
Vapnik V, Chervonenkis A (1974) Theory of pattern recognition. Nauka, Moskow (in Russian)
Vembu S, Gärtner T (2010) Label ranking algorithms: a survey. In: Preference learning. Springer, Berlin, pp 45–64
Vembu S, Gärtner T, Boley M (2009) Probabilistic structured predictors. In: Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence (UAI 2009), pp 557–564
Viola PA, Jones MJ (2001) Robust real-time face detection. In: Proceedings of the 8th international conference on computer vision ICCV 2001, p 747
Wainwright M, Jordan M (2008) Graphical models, exponential families, and variational inference. Found Trends Mach Learn 1(1–2):1–305
Watanabe S (2009) Algebraic Geometry and Statistical Learning Theory. Cambridge University, Cambridge
Webb A, Copsey K (2011) Statistical pattern recognition. Wiley, New York
Wibisono A, Rosasco L, Poggio T (2009) Sufficient conditions for uniform stability of regularization algorithms. Technical Report MIT-CSAIL-TR-2009-060. MIT, Computer Science and artificial intelligence laboratory
Wright SJ (2015) Coordinate descent algorithms. Math Program 151(1):3–34
Xu J, Li H (2007) AdaRank: a boosting algorithm for information retrieval. In: Proceedings of the 30th annual international ACM conference on research and development in information retrieval (SIGIR 2007), pp 391–398
Zhang L, Yang T, Jin R, He X (2013) $O(\log T)$ projections for stochastic optimization of smooth and strongly convex functions. In: Proceedings of the 30th international conference on machine learning (ICML 2013), pp 1121–1129
Zhang X (2010) Empirical risk minimization. In: Sammut C, Webb G, (eds) Encyclopedia of machine learning, Springer, Berlin, p 312
Zhang Y, Liang P, Wainwright M (2017) Convexified convolutional neural networks. In: Proceedings of the 34th international conference on machine learning (ICML 2017), pp 4044–4053
Zhao Z, Piech P, Xia L (2016) Learning mixtures of plackett-luce models. In: Proceedings of the 33nd international conference on machine learning (ICML 2016), pp 2906–2914

Reinforcement Learning

Olivier Buffet, Olivier Pietquin and Paul Weng

Abstract Reinforcement learning (RL) is a general framework for adaptive control, which has proven to be efficient in many domains, e.g., board games, video games or autonomous vehicles. In such problems, an agent faces a sequential decision-making problem where, at every time step, it observes its state, performs an action, receives a reward and moves to a new state. An RL agent learns by trial and error a good policy (or controller) based on observations and numeric reward feedback on the previously performed action. In this chapter, we present the basic framework of RL and recall the two main families of approaches that have been developed to learn a good policy. The first one, which is value-based, consists in estimating the value of an optimal policy, value from which a policy can be recovered, while the other, called policy search, directly works in a policy space. Actor-critic methods can be seen as a policy search technique where the policy value that is learned guides the policy improvement. Besides, we give an overview of some extensions of the standard RL framework, notably when risk-averse behavior needs to be taken into account or when rewards are not available or not known.

Olivier Buffet, Olivier Pietquin and Paul Weng—Equally contributed in this chapter.

O. Buffet (✉)
INRIA, Université de Lorraine, CNRS, UMR 7503 - LORIA, Nancy, France
e-mail: olivier.buffet@loria.fr

O. Pietquin
Université de Lille, CNRS, Centrale Lille, Inria, UMR 9189 - CRIStAL, Lille, France
e-mail: olivier.pietquin@univ-lille1.fr

Google Brain, Paris, France

P. Weng
Shanghai Jiao Tong University,
University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai, China
e-mail: paul.weng@sjtu.edu.cn

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_12

1 Introduction

Reinforcement learning (RL) is a general framework for building autonomous agents (physical or virtual), which are systems that make decisions without human supervision in order to perform a given task. Examples of such systems abound: expert backgammon player (Tesauro 1995), dialogue systems (Singh et al. 1999), acrobatic helicopter flight (Abbeel et al. 2010), human-level video game player (Mnih et al. 2015), go player (Silver et al. 2016) or autonomous driver (Bojarski et al. 2016). See also chapter "Artificial Intelligence for Games" of Volume 2 and chapters "Artificial Intelligence and Pattern Recognition, Vision, Learning" and "Robotics and Artificial Intelligence" of Volume 3.

In all those examples, an agent faces a sequential decision-making problem, which can be represented as an interaction loop between an agent and an environment. After observing its current situation, the agent selects an action to perform. As a result, the environment changes its state and provides a numeric reward feedback about the chosen action. In RL, the agent needs to learn how to choose good actions based on its observations and the reward feedback, without necessarily knowing the dynamics of the environment.

In this chapter, we focus on the basic setting of RL that assumes a single learning agent with full observability. Some work has investigated the partial observability case (see Spaan (2012) for an overview of both the model-based and model-free approaches). The basic setting has also been extended to situations where several agents interact and learn simultaneously (see Busoniu et al. (2010) for a survey). RL has also been tackled with Bayesian inference techniques, which we do not mention here for space reasons (see Ghavamzadeh et al. (2015) for a survey).

In Sect. 2, we recall the Markov decision process model on which RL is formulated and the RL framework, along with some of their classic solution algorithms. We present two families of approaches that can tackle large-sized problems for which function approximation is usually required. The first, which is value-based, is presented in Sect. 3. It consists in estimating the value function of an optimal policy. The second, called policy search, is presented in Sect. 4. It searches for an optimal policy directly in a policy space. In Sect. 5, we present some extensions of the standard RL setting, namely extensions to the case of unknown rewards and risk-sensitive RL approaches. Finally, we conclude in Sect. 6.

2 Background for RL

Before presenting the RL framework, we recall the Markov decision process (MDP) model, on which RL is based. See also chapter "Decision under Uncertainty" of this volume and chapter "Planning in Artificial Intelligence" of Volume 2.

Markov decision process. MDPs and their multiple variants (e.g., Partially Observable MDP or POMDP) (Puterman 1994) have been proposed to represent and solve sequential decision-making problems under uncertainty. An MDP is defined as a tuple $\mathscr{M} = \langle \mathbf{S}, \mathbf{A}, T, R, \gamma, H \rangle$ where $\mathbf{S}$ is a set of states, $\mathbf{A}$ is a set of actions, transi-

tion function $T(s, a, s')$ specifies the probability of reaching state s' after performing action a in state s, reward function $R(s, a) \in \mathbb{R}$ yields the immediate reward after performing action a in state s, $\gamma \in [0, 1]$ is a discount factor and $H \in \mathbb{N} \cup \{\infty\}$ is the horizon of the problem, which is the number of decisions to be made. An immediate reward, which is a scalar number, measures the value of performing an action in a state. In some problems, it can be randomly generated. In that case, $R(s, a)$ is simply the expectation of the random rewards. In this MDP formulation, the environment is assumed to be stationary. Using such an MDP model, a system designer needs to define the tuple $\mathscr{M}$ such that an optimal policy performs the task s/he wants.

Solving an MDP (i.e., *planning*) amounts to finding a controller, called a *policy*, which specifies which action to take in every state of the environment in order to maximize the expected discounted sum of rewards (standard decision criterion). A policy π can be deterministic (i.e., $\pi(s) \in A$) or randomized (i.e., $\pi(\cdot \mid s)$ is a probability distribution over $\mathbf{A}$). It can also be stationary or time-dependent, which is useful in finite-horizon or non-stationary problems.

A t-step history (also called trajectory, rollout or path) $h = (s_1, a_1, s_2, \ldots, s_{t+1}) \in (\mathbf{S} \times \mathbf{A})^t \times \mathbf{S}$ is a sequence of past states and actions. In the standard case, it is valued by its return defined as $\sum_t \gamma^{t-1} R(s_t, a_t)$. As a policy induces a probability distribution over histories, the *value function* $v^\pi : \mathbf{S} \to \mathbb{R}$ of a policy π is defined by:

$$v_H^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=1}^{H} \gamma^{t-1} R(S_t, A_t) \mid S_1 = s\right],$$

where $\mathbb{E}_\pi$ is the expectation with respect to the distribution induced by π in the MDP, and S_t and A_t are random variables respectively representing a state and an action at a time step t. We will drop subscript H if there is no risk of confusion. The value function can be computed recursively. For deterministic policy π, we have:

$$\begin{aligned} v_0^\pi(s) &= 0, \\ v_t^\pi(s) &= R(s, \pi(s)) + \gamma \sum_{s' \in \mathbf{S}} T(s, \pi(s), s') v_{t-1}^\pi(s'). \end{aligned}$$

In a given state, policies can be compared via their value functions. Interestingly, in standard MDPs, there always exists an optimal deterministic policy whose value function is maximum in every state. Its value function is said to be optimal.

In the infinite horizon case, when $\gamma < 1$, v_t^π is guaranteed to converge to v^π, which is the solution of the *Bellman evaluation equations*:

$$v^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in \mathbf{S}} T(s, \pi(s), s') v^\pi(s'). \tag{1}$$

Given v^π, a better policy can be obtained with the following improvement step:

$$\pi'(s) = \operatorname*{argmax}_{a \in \mathbf{A}} R(s, a) + \gamma \sum_{s' \in \mathbf{S}} T(s, a, s') v^\pi(s'). \tag{2}$$

The policy iteration algorithm consists in alternating between a policy evaluation step (1) and a policy improvement step (2), which converges to the optimal value function $v^* : \mathbf{S} \to \mathbb{R}$.

Alternatively, the optimal value function $v^*_H : S \to \mathbb{R}$ can also be iteratively computed for any horizon H by:

$$\begin{aligned} v^*_0(s) &= 0 \\ v^*_t(s) &= \max_{a \in \mathbf{A}} R(s,a) + \gamma \sum_{s' \in \mathbf{S}} T(s,a,s') v^*_{t-1}(s'). \end{aligned} \tag{3}$$

In the infinite horizon case, when $\gamma < 1$, v^*_t is guaranteed to converge to v^*, which is the solution of the *Bellman optimality equations*:

$$v^*(s) = \max_{a \in \mathbf{A}} R(s,a) + \gamma \sum_{s' \in \mathbf{S}} T(s,a,s') v^*(s'). \tag{4}$$

In that case, (3) leads to the value iteration algorithm.

Two other related functions are useful when solving an RL problem: the action-value function $Q^\pi_t(s,a)$ (resp. the optimal action-value function $Q^*_t(s,a)$) specifies the value of choosing an action a in a state s at time step t and assuming policy π (resp. an optimal policy) is applied thereafter, i.e.,

$$Q^x_t(s,a) = R(s,a) + \gamma \sum_{s' \in \mathbf{S}} T(s,a,s') v^x_{t-1}(s') \quad \text{where } x \in \{\pi, *\}.$$

Reinforcement learning. In the MDP framework, a complete model of the environment is assumed to be known (via the transition function) and the task to be performed is completely described (via the reward function). The RL setting has been proposed to tackle situations when those assumptions do not hold. An RL agent searches for (i.e., during the *learning phase*) a best policy while interacting with the unknown environment by trial and error. In RL, the standard decision criterion used to compare policies is the same as in the MDP setting. Although the reward function is supposed to be unknown, the system designer has to specify it.

In RL, value and action-value functions have to be estimated. For v^π of a given policy π, this can be done with the standard TD(0) evaluation algorithm, where the following update is performed after applying π in state s yielding reward r and moving to new state s':

$$v^\pi_t(s) = v^\pi_{t-1}(s) - \alpha_t(s) \left(v^\pi_{t-1}(s) - \left(r + v^\pi_{t-1}(s') \right) \right), \tag{5}$$

where $\alpha_t(s) \in [0,1]$ is a learning rate. For Q^π, the update is as follows, after the agent executed action a in state s, received r, moved to new state s' and executed action a' (chosen by π):

$$Q_t^\pi(s,a) = Q_{t-1}^\pi(s,a) - \alpha_t(s,a)\big(Q_{t-1}^\pi(s,a) - \big(r + \gamma\, Q_{t-1}^\pi(s',a')\big)\big), \quad (6)$$

where $\alpha_t(s,a) \in [0,1]$ is a learning rate. This update leads to the SARSA algorithm (named after the variables s, a, r, s', a'). In the same way that the policy iteration algorithm alternates between an evaluation step and a policy improvement step, one can use the SARSA evaluation method and combine it with a policy improvement step. In practice, we do not wait for the SARSA evaluation update rule to converge to the actual value of the current policy to make a policy improvement step. We rather continuously behave according to the current estimate of the Q-function to generate a new transition. One common choice is to use the current estimate in a softmax (Boltzmann) function of temperature τ and behave according to a randomized policy:

$$\pi_t(a \mid s) = \frac{e^{Q_{\theta_t}(s,a)/\tau}}{\sum_b e^{Q_{\theta_t}(s,b)/\tau}}.$$

Notice that we chose to use the Bellman evaluation equations to estimate the targets. However we could also use the Bellman optimality equations in the case of the Q-function and replace $r + \gamma\, Q(s',a')$ by $r + \max_b Q(s',b)$. Yet this only holds if we compute the value Q^* of the optimal policy π^*. This gives rise to the Q-learning update rule, which directly computes the value of the optimal policy. It is called an *off-policy* algorithm (whereas SARSA is *on-policy*) because it computes the value function of another policy than the one that selects the actions and generates the transitions used for the update. The following update is performed after the agent executed action a (e.g., chosen according to the softmax rule) in state s, received r and moved to new state s':

$$Q_t^*(s,a) = Q_{t-1}^*(s,a) - \alpha_t(s,a)\big(Q_{t-1}^*(s,a) - (r + \gamma \max_{a'} Q_{t-1}^*(s',a'))\big). \quad (7)$$

Updates (5), (6) and (7) can be proved to converge if the learning rates satisfy standard stochastic approximation conditions (i.e., $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$). Besides, for (6), temperature τ would also need to converge to 0 while ensuring sufficient exploration in order for SARSA to converge to the optimal Q-function. In practice, $\alpha_t(s,a)$ is often chosen constant, which would also account for the case where the environment is non-stationary.

Those two general framework (MDP and RL) have been successfully applied in many different domains. For instance, MDPs or their variants have been used in finance (Bäuerle and Rieder 2011) or logistics (Zhao et al. 2010). RL has been applied to soccer (Bai et al. 2013) or power systems (Yu and Zhang 2013), to cite a few. To tackle real-life large-sized problems, MDP and RL have to be completed with other techniques, such as compact representations (Boutilier et al. 2000; Guestrin et al. 2004; van Otterlo 2009) or function approximation (de Farias and Van Roy 2003; Geist and Pietquin 2011; Mnih et al. 2015).

3 Value-Based Methods with Function Approximation

In many cases, the state-action space is too large so as to be able to represent exactly the value functions v^π or the action-value function Q^π of a policy π. For this reason, function approximation for RL has been studied for a long time, starting with the seminal work of Bellman and Dreyfus (1959). In this framework, the functions are parameterized by a vector of d parameters $\boldsymbol{\theta} = [\theta_j]_{j=1}^d$, with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ (we will always consider column vectors) and the algorithms will aim at learning the parameters from data provided in the shape of transitions $\{s_t, a_t, s'_t, r_t\}_{t=1}^N$ where s'_t is the successor state of s_t drawn from $T(s_t, a_t, \cdot)$. We will denote the parameterized versions of the functions as $v_{\boldsymbol{\theta}}$ and $Q_{\boldsymbol{\theta}}$. Popular approximation schemes are linear function approximation and neural networks. The former gave birth to a large literature in the theoretical domain as it allows studying convergence rates and bounds (although it remains non-trivial). The latter, although already used in the 90s (Tesauro 1995), has known a recent growth in interest following the Deep Learning successes in supervised learning.

The case of neural networks will be addressed in Sect. 3.4 but we will start with linear function approximation. In this particular case, a set of basis functions $\boldsymbol{\phi}(\cdot) = [\phi_j(\cdot)]_{j=1}^d$ has to be defined by the practitioner (or maybe learned through unsupervised learning) so that the value functions can be approximated by:

$$v_{\boldsymbol{\theta}}(s) = \sum_j \theta_j \phi_j(s) = \boldsymbol{\theta}^\intercal \boldsymbol{\phi}(s) \quad \text{or} \quad Q_{\boldsymbol{\theta}}(s,a) = \sum_j \theta_j \phi_j(s,a) = \boldsymbol{\theta}^\intercal \boldsymbol{\phi}(s,a).$$

The vector space defined by the span of $\boldsymbol{\phi}$ is denoted Φ.

Notice that the exact case in which the different values of the value functions can be stored in a table (tabular case) is a particular case of linear function approximation. Indeed, if we consider that the state space is finite and small $\left(s = \{s_k\}_{k=1}^{|\mathbf{S}|} \in \mathbf{S}\right)$, then the value function can be represented in a table of $|\mathbf{S}|$ values $\{v_k \mid v_k = v(s_k)\}_{k=1}^{|\mathbf{S}|}$ where $|\mathbf{S}|$ is the number of states. This is equivalent to defining a vector of $|\mathbf{S}|$ parameters $\boldsymbol{v} = [v_k]_{k=1}^{|\mathbf{S}|}$ and a vector of $|\mathbf{S}|$ basis functions $\boldsymbol{\delta}(s) = [\delta_k(s)]_{k=1}^{|\mathbf{S}|}$ where $\delta_k(s) = 1$ if $s = s_k$ and 0 otherwise. The value function can thus be written $v(s) = \sum_k v_k \delta_k(s) = \boldsymbol{v}^\intercal \boldsymbol{\delta}(s)$.

3.1 Stochastic Gradient Descent Methods

3.1.1 Bootstrapped Methods

If one wanted to cast the Reinforcement Learning problem into a supervised learning problem (see chapter "Statistical Computational Learning" of this Volume and chapter "Designing Algorithms for Machine Learning and Data Mining" of Volume 2), one could want to fit the parameters to the value function directly. For instance, to

evaluate the value of a particular policy π, one would solve the following regression problem (for some ℓ_p-norm and distribution μ over states):

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \|v^\pi_{\boldsymbol{\theta}} - v^\pi\|_{p,\mu} = \operatorname*{argmin}_{\boldsymbol{\theta}} \|v^\pi_{\boldsymbol{\theta}} - v^\pi\|^p_{p,\mu}$$

where $\|\cdot\|_{p,\mu}$ denotes the weighted ℓ_p-norm defined by $\left(\mathbb{E}_\mu \|\cdot\|^p\right)^{1/p}$, $\mathbb{E}_\mu$ is the expectation with respect to μ. Yet, as said before, we usually cannot compute these values everywhere and we usually only have access to some transition samples $\{s_t, a_t, s'_t, r_t\}_{t=1}^N$ generated according to distribution μ. So we could imagine casting the RL problem into the following minimization problem:

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{N} \sum_{t=1}^{N} |v^\pi_{\boldsymbol{\theta}}(s_t) - v^\pi(s_t)|^p.$$

This cost function can be minimized by stochastic gradient descent (we will consider an ℓ_2-norm):

$$\begin{aligned}\boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} - \frac{\alpha}{2} \nabla_{\boldsymbol{\theta}_{t-1}} \left(v^\pi_{\boldsymbol{\theta}_{t-1}}(s_t) - v^\pi(s_t)\right)^2 \\ &= \boldsymbol{\theta}_{t-1} - \alpha \nabla_{\boldsymbol{\theta}_{t-1}} v^\pi_{\boldsymbol{\theta}_{t-1}}(s_t) \left(v^\pi_{\boldsymbol{\theta}_{t-1}}(s_t) - v^\pi(s_t)\right).\end{aligned}$$

Of course, it is not possible to apply this update rule as it is since we do not know the actual value $v^\pi(s_t)$ of the states we observe in the transitions. But, from the Bellman evaluation equations (1), we can obtain an estimate by replacing $v^\pi(s_t)$ by $r_t + \gamma v^\pi_{\boldsymbol{\theta}_{t-1}}(s_{t+1})$. Notice that this replacement uses bootstrapping as we use the current estimate of the target to compute the gradient. We finally obtain the following update rule for evaluating the current policy π:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \nabla_{\boldsymbol{\theta}_{t-1}} v^\pi_{\boldsymbol{\theta}_{t-1}}(s_t) \left(v^\pi_{\boldsymbol{\theta}_{t-1}}(s_t) - \left(r_t + \gamma v^\pi_{\boldsymbol{\theta}_{t-1}}(s'_t)\right)\right).$$

In the case of linear function approximation, i.e., $v^\pi_{\boldsymbol{\theta}}(s) = \boldsymbol{\theta}^\intercal \boldsymbol{\phi}(s)$, we obtain:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \boldsymbol{\phi}(s_t) \left(\boldsymbol{\theta}^\intercal_{t-1} \boldsymbol{\phi}(s_t) - \left(r_t + \gamma \boldsymbol{\theta}^\intercal_{t-1} \boldsymbol{\phi}(s'_t)\right)\right).$$

Everything can be written again in the case of the action-value function, which leads to the SARSA update rule with linear function approximation $Q^\pi_{\boldsymbol{\theta}}(s,a) = \boldsymbol{\theta}^\intercal \boldsymbol{\phi}(s,a)$:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \boldsymbol{\phi}(s_t, a_t) \left(\boldsymbol{\theta}^\intercal_{t-1} \boldsymbol{\phi}(s_t, a_t) - \left(r_t + \gamma \boldsymbol{\theta}^\intercal_{t-1} \boldsymbol{\phi}(s'_t, a'_t)\right)\right).$$

Changing the target as in the Q-learning update, we obtain for $Q^*_{\boldsymbol{\theta}}(s,a) = \boldsymbol{\theta}^\intercal \boldsymbol{\phi}(s,a)$:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \boldsymbol{\phi}(s_t, a_t) \left(\boldsymbol{\theta}^\intercal_{t-1} \boldsymbol{\phi}(s_t, a_t) - \left(r_t + \gamma \max_b \boldsymbol{\theta}^\intercal_{t-1} \boldsymbol{\phi}(s'_t, b)\right)\right).$$

3.1.2 Residual Methods

Instead of using the Bellman equations to provide an estimate of the target after deriving the update rule, one could use it directly to define the loss function to be optimized. We would then obtain the following minimization problem:

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{N} \sum_{t=1}^{N} \left(v_{\boldsymbol{\theta}}^{\pi}(s_t) - \left(r_t + \gamma v_{\boldsymbol{\theta}}^{\pi}(s_t') \right) \right)^2 .$$

This can also be seen as the minimization of the Bellman residual. Indeed the Bellman evaluation equations ($v^{\pi}(s) = \mathbb{E}_{\pi}[R(s, A) + \gamma v^{\pi}(S')]$) can be rewritten as $v^{\pi}(s) - \mathbb{E}_{\pi}[R(s, A) + \gamma v^{\pi}(S')] = 0$. So by minimizing the quantity $v^{\pi}(s) - \mathbb{E}_{\pi}[R(s, A) + \gamma v^{\pi}(S')]$, called the Bellman residual, we reach the objective of evaluating $v^{\pi}(s)$. Here, we take the observed quantity $r + \gamma v^{\pi}(s')$ as an unbiased estimate of its expectation. The Bellman residual can also be minimized by stochastic gradient descent as proposed by Baird et al. (1995) and the update rule becomes:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \nabla_{\boldsymbol{\theta}_{t-1}} \left(v_{\boldsymbol{\theta}_{t-1}}^{\pi}(s_t) - \left(r_t + \gamma v_{\boldsymbol{\theta}_{t-1}}^{\pi}(s_t') \right) \right) \left(v_{\boldsymbol{\theta}_{t-1}}^{\pi}(s_t) - \left(r_t + \gamma v_{\boldsymbol{\theta}_{t-1}}^{\pi}(s_t') \right) \right) .$$

In the case of a linear approximation, we obtain:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \left(\boldsymbol{\phi}(s_t) - \gamma \boldsymbol{\phi}(s_t') \right) \left(\boldsymbol{\theta}_{t-1}^{\intercal} \boldsymbol{\phi}(s_t) - \left(r_t + \gamma \boldsymbol{\theta}_{t-1}^{\intercal} \boldsymbol{\phi}(s_t') \right) \right) .$$

This approach, called R-SGD (for residual stochastic gradient descent), has a major flaw as it computes a biased estimate of the value-function. Indeed, $v_{\boldsymbol{\theta}}^{\pi}(s_t)$ and $v_{\boldsymbol{\theta}}^{\pi}(s_t')$ are correlated as s_t' is the result of having taken action a_t chosen by $\pi(s_t)$ (Werbos 1990). To address this problem, Baird et al. (1995) suggest to draw two different next states s_t' and s_t'' starting from the same state s_t and to update as follows:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \nabla_{\boldsymbol{\theta}_{t-1}} \left(v_{\boldsymbol{\theta}_{t-1}}^{\pi}(s_t) - \left(r_t + \gamma v_{\boldsymbol{\theta}_{t-1}}^{\pi}(s_t') \right) \right) \left(v_{\boldsymbol{\theta}_{t-1}}^{\pi}(s_t) - \left(r_t + \gamma v_{\boldsymbol{\theta}_{t-1}}^{\pi}(s_t'') \right) \right) .$$

Of course, this requires that a generative model or a simulator is available and that transitions can be generated on demand.

The same discussions as in previous section can apply to learning an action-value function. For instance, one could want to solve the following optimization problem to learn the optimal action-value function:

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{N} \sum_{t=1}^{N} \left(Q_{\boldsymbol{\theta}}^*(s_t, a_t) - \left(r_t + \gamma \max_b Q_{\boldsymbol{\theta}}^*(s_t', b) \right) \right)^2 . \tag{8}$$

Yet this optimal residual cannot directly be minimized in the case of the Q-function as the max operator is not differentiable. Notice that a sub-gradient method can still be used.

3.2 Least-Squares Methods

Gradient descent was used to minimize the empirical norm of either the bootstrapping error or the Bellman residual in the previous section. As the empirical norm is generally using the ℓ_2-norm and that linear function approximation is often assumed, another approach could be to find the least squares solution to these problems. Indeed, least squares is a powerful approach as it is a second-order type of method and offers a closed-form solution to the optimization problem. Although there is no method that explicitly applies least squares to the two aforementioned empirical errors, one can see the fixed-point Kalman Filter (FPKF) algorithm (Choi and Van Roy 2006) as a recursive least squares method applied to the bootstrapping error minimization. Also, the Gaussian Process Temporal Difference (GPTD) (Engel et al. 2005) or the Kalman Temporal Difference (KTD) (Geist and Pietquin 2010a) algorithms can be seen as recursive least squares methods applied to Bellman residual minimization. We invite the reader to refer to Geist and Pietquin (2013) for further discussion on this.

Yet, the most popular method inspired by least squares optimization does apply to a different cost function. The Least-Squares Temporal Difference (LSTD) algorithm (Bradtke and Barto 1996) aims at minimizing:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left(v_{\boldsymbol{\theta}}^{\pi}(s_i) - v_{\boldsymbol{\omega}^*}^{\pi}(s_i) \right)^2,$$

where $\boldsymbol{\omega}^* = \operatorname{argmin}_{\boldsymbol{\omega}} \frac{1}{N} \sum_{i=1}^{N} \left(v_{\boldsymbol{\omega}}^{\pi}(s_i) - \left(r_i + \gamma v_{\boldsymbol{\theta}}^{\pi}(s_i') \right) \right)^2$ can be understood as a projection on the space Φ spanned by the family of functions $\boldsymbol{\phi}_j$'s used to approximate v^{π}. It can be seen as the composition of the Bellman operator and of a projection operator. This cost function is the so-called *projected Bellman residual*. When using linear function approximation, this optimization problem admits a closed-form solution:

$$\boldsymbol{\theta}^* = \left[\sum_{i=1}^{N} \boldsymbol{\phi}(s_i) \left[\boldsymbol{\phi}(s_i) - \gamma \boldsymbol{\phi}(s_i') \right]^{\intercal} \right]^{-1} \sum_{i=1}^{N} \boldsymbol{\phi}(s_i) r_i.$$

Note that the projected Bellman residual can also be optimized with a stochastic gradient approach (Sutton et al. 2009).

Extensions to non-linear function approximation exist and rely on the kernel trick (Xu et al. 2007) or on statistical linearization (Geist and Pietquin 2010b). LSTD can be used to learn an approximate Q-function as well and can be combined with policy improvement steps into an iterative algorithm, similar to policy iteration, to learn an optimal policy from a dataset of sampled transitions. This gives rise to the so-called Least Squares Policy Iteration (LSPI) algorithm (Lagoudakis and Parr 2003), which is one of the most popular batch-RL algorithm.

3.3 Iterative Projected Fixed-Point Methods

As we have seen earlier, dynamic programming offers a set of algorithms to compute value functions of a policy in the case the dynamics of the MDP is known. One of these algorithms, Value Iteration, relies on the fact that the Bellman equations define contraction operators when $\gamma < 1$. For instance, if we define the Bellman evaluation operator B^π such that $B^\pi Q(s,a) = R(s,a) + \gamma \mathbb{E}_\pi\big[Q(S',A') \mid S = s, A = a\big]$, one can show that iteratively applying B^π to a random initialization of Q converges to Q^π, because B^π defines a contraction for which the only fixed point is Q^π (Puterman 1994). The Bellman optimality operator B^*, defined as $B^*Q(s,a) = R(s,a) + \gamma \mathbb{E}\big[\max_b Q(S',b) \mid S = s, A = a\big]$, is also a contraction. The same holds for the sampled versions of the Bellman operators. For instance, let us define the sampled evaluation operator $\hat{B}^*$ such that $\hat{B}^*Q(s,a) = r + \gamma \max_b Q(s',b)$, where the expectation has been removed (the sampled operator applies to a single transition). Unfortunately, there is no guarantee that this remains a contraction when the value functions are approximated. Indeed when applying a Bellman operator to an approximate $Q_{\boldsymbol{\theta}}$, the result might not lie in the space spanned by $\boldsymbol{\theta}$. One has thus to project back on the space Φ spanned by $\boldsymbol{\phi}$ using a projection operator Π_Φ, i.e., $\Pi_\Phi f = \text{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{\theta}^\intercal \boldsymbol{\phi} - f\|_2$. If the composition of Π_Φ and $\hat{B}^\pi$ (or $\hat{B}^*$) is still a contraction, then recursively applying this composition to any initialization of $\boldsymbol{\theta}$ still converges to a good approximate $Q^\pi_{\boldsymbol{\theta}}$ (or $Q^*_{\boldsymbol{\theta}}$). Unfortunately, the exact projection is often impossible to get as it is a regression problem. For instance, one would need to use least squares methods or stochastic gradient descent to learn the best projection from samples. Therefore the projection operator itself is approximated and will result in some $\hat{\Pi}_\Phi$ operator. So the iterative projected fixed-point process is defined as:

$$Q_{\boldsymbol{\theta}_t} = \hat{\Pi}_\Phi \hat{B}^\pi Q_{\boldsymbol{\theta}_{t-1}} \quad \text{or} \quad Q_{\boldsymbol{\theta}_t} = \hat{\Pi}_\Phi \hat{B}^* Q_{\boldsymbol{\theta}_{t-1}}.$$

In practice, the algorithm consists in collecting transitions (e.g., $\{s_i, a_i, r_i, s'_i\}_{i=1}^N$), initialize $\boldsymbol{\theta}_0$ to some random value, compute a regression database by applying the chosen sampled Bellman operator (e.g., $\{\hat{B}^* Q_{\boldsymbol{\theta}_0}(s_i,a_i) = r_i + \gamma \max_b Q_{\boldsymbol{\theta}_0}(s_i,b)\}_{i=1}^N$), apply a regression algorithm to find the next value of parameters (e.g., $Q_{\boldsymbol{\theta}_1} = \hat{\Pi}_\Phi \hat{B}^* Q_{\boldsymbol{\theta}_0} = \text{argmin}_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^N \Big(Q_{\boldsymbol{\theta}}(s_i,a_i) - \hat{B}^* Q_{\boldsymbol{\theta}_0}(s_i,a_i)\Big)^2$) and iterate.

This method finds its roots in early papers on dynamic programming (Samuel 1959; Bellman et al. 1963) and convergence properties have been analyzed by Gordon (1995). The most popular implementations use regression trees (Ernst et al. 2005) or neural networks (Riedmiller 2005) as regression algorithms and have been applied to many concrete problems such as robotics (Antos et al. 2008).

3.4 Value-Based Deep Reinforcement Learning

Although the use of Artificial Neural Networks (ANN, see chapter "Designing Algorithms for Machine Learning and Data Mining" of Volume 2) in RL is not

new (Tesauro 1995), there has been only a few successful attempts to combine RL and ANN in the past. Most notably, before the recent advances in Deep Learning (DL) (LeCun et al. 2015), one can identify the work by Riedmiller (2005) as the biggest success of ANN as a function approximation framework for RL. There are many reasons for that, which are inherently due to the way ANN learns and assumptions that have to be made for both gradient descent and most value-based RL algorithms to converge. Especially, Deep ANNs (DNN) require a tremendous amount of data as they contain a lot of parameters to learn (typically hundreds of thousands to millions). To alleviate this issue, Tesauro (1995) trained his network to play backgammon through a self-play procedure. The model learned at iteration t plays again itself to generate data for training the model at iteration $t + 1$. It could thus reach super-human performance at the game of backgammon using RL. This very simple and powerful idea was reused in Silver et al. (2016) to build the first artificial Go player that consistently defeated a human Go master. Yet, this method relies on the assumption that games can easily be generated on demand (backgammon and Go rules are simple enough even though the game is very complex). In more complex settings, the agent faces an environment for which it does not have access to the dynamics, maybe it cannot start in random states and has to follow trajectories, and it can only get transitions through actual interactions. This causes two major issues for learning with DNNs (in addition to intensive usage of data). First, gradient descent for training DNNs assume the data to be independent and identically distributed (i.i.d. assumption). Second, the distribution of the data should remain constant over time. Both these assumptions are normally violated by RL since transitions used to train the algorithms are part of trajectories (so next states are functions of previous states and actions, violating the i.i.d. assumption) and because trajectories are generated by a policy extracted from the current estimate of the value function (learning the value function influences the distribution of the data generated in the future). In addition, we also have seen in Sect. 3.1.2 that Bellman residual minimization suffers from the correlation between estimates of value functions of successive states. All these problems make RL unstable (Gordon 1995).

To alleviate these issues, Mnih et al. (2015) used two tricks that allowed to reach super-human performances at playing Atari 2600 games from pixels. First, they made use of a biologically inspired mechanism, called experience replay (Lin 1992), that consists in storing transitions in a Replay Buffer D before using them for learning. Instead of sequentially using these transitions, they are shuffled in the buffer and randomly sampled for training the network (which helps breaking correlation between successive samples). The buffer is filled on a first-in-first-out basis so that the distribution of the transitions is nearly stationary (transitions generated by old policies are discarded first). Second, the algorithm is based on asynchronous updates of the network used for generating the trajectories and a slow learning network. The slow learning network, called the target network, will be updated less often than the network that actually learns from the transitions stored in the replay buffer (the Q-network). This way, the distribution of transitions in the replay buffer remains constant for a longer time from the fast learning network point of view. In addition, the update rule of the Q-network is built such that correlation between estimates of

$Q(s, a)$ and $Q(s', a')$ is reduced. Indeed, the resulting algorithm (Deep Q-Network or DQN) is inspired by the gradient-descent update on the optimal Bellman residual (8). But instead of using the double-sampling trick mentioned in Sect. 3.1.2, two different estimates of the Q-function are used. One according to the target network parameters ($\boldsymbol{\theta}^-$) and the other according to Q-network parameters ($\boldsymbol{\theta}$). The parameters of the Q-network are thus computed as:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{(s_t, a_t, s'_t, r_t) \in D} \left[\left(r_t + \gamma \max_b Q_{\boldsymbol{\theta}^-}(s'_t, b) \right) - Q_{\boldsymbol{\theta}}(s_t, a_t) \right]^2 ,$$

With this approach, the problem of non-differentiability of the max operator is also solved as the gradient is computed w.r.t. $\boldsymbol{\theta}$ and not $\boldsymbol{\theta}^-$. Once in a while, the target network parameters are updated with the Q-network parameters ($\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}^*$) and new trajectories are generated according to the policy extracted from $Q_{\boldsymbol{\theta}^-}$ to fill again the replay buffer and train again the Q-network. The target network policy is actually a softmax policy based on $Q_{\boldsymbol{\theta}^-}$ (see Sect. 3.1.1). Many improvements have been brought to that method since its publication, such as a prioritized replay mechanism (Schaul et al. 2016) that allows to sample more often from the replay buffer transitions for which the Bellman residual is larger, or the Double-DQN trick (Van Hasselt et al. 2016) used to provide more stable estimates of the max operator.

4 Policy-Search Approaches

Value-based approaches to RL rely on approximating the optimal value function V^* (typically using Bellman's optimality principle), and then acting greedily with respect to this function. Policy Search algorithms directly optimize control policies, which typically depend on a parameter vector $\boldsymbol{\theta} \in \Theta$ (and are thus noted $\pi_{\boldsymbol{\theta}}$), and whose general shape is user-defined.[1] Possible representations include linear policies, (deep) neural networks, radial basis function networks, and dynamic movement primitives (in robotics). Using such approaches avoids issues with discontinuous value functions, and makes it possible, in some cases, to deal with high-dimensional (possibly continuous) state and action spaces. They also allow providing expert knowledge through the shaping of the controller, or through example trajectories—to initialize the parameters.

In the following, we mainly distinguish between *model-free* and *model-based* algorithms—i.e., depending on whether a model is being learned or not.

[1] This section is mainly inspired by Deisenroth et al. (2011), although that survey focuses on a robotic framework.

4.1 Model-Free Policy Search

In model-free policy search, sampled trajectories are used directly to update the policy parameters. The discussion will follow the three main steps followed by the algorithms: (i) how they *explore* the space of policies, (ii) how they *evaluate* policies, and (iii) how policies are *updated*.

4.1.1 Policy Exploration

Exploring the space of policies implies either sampling the parameter vector the policy depends on, or perturbing the action choice of the policy. Often, the sampling of parameters takes place at the beginning of each episode (in episodic scenarios), and action perturbations are different at each time step, but other options are possible. Stochastic policies can be seen as naturally performing a step-based exploration in action space. Otherwise, the exploration strategy can be modeled as an *upper-level policy* $\pi_\omega(\theta)$—sampling θ according to a probability distribution governed by parameter vector ω—, while the actual policy $\pi_\theta(a|s)$ is refered to as a *lower-level policy*. In this setting, the policy search aims at finding the parameter vector ω that maximizes the expected return given this vector. If $\pi_\omega(\theta)$ is a Gaussian distribution (common in robotics), then its covariance matrix can be diagonal—typically in step-based exploration—or not—which leads to more stability, but requires more samples—, meaning that the various parameters in θ can be treated in a correlated manner or not.

4.1.2 Policy Evaluation

Policy evaluation can also be step-based or episode-based. Step-based approaches evaluate each state-action pair. They have low variance and allow crediting several parameter vectors. They can rely on Q-value estimates, which can be biased and prone to approximation errors, or Monte-Carlo estimates, which can suffer from high variance. Episode-based approaches evaluate parameters using complete trajectories. They allow more performance criteria than step-based approaches—e.g., minimizing the final distance to the target. They also allow for more sophisticated exploration strategies, but suffer all the more from noisy estimates and high variance that the dynamics are more stochastic.

4.1.3 Policy Update

Finally, the policy can be updated in rather different manners. We will discuss approaches relying on gradient ascents, inference-based optimization, information-theoretic ideas, stochastic optimization and path-integral optimal control.

Policy Gradient (PG) algorithms first require estimating the gradient. Some (episode-based) PG algorithms perform this estimate using a finite difference (FD) method by perturbing the parameter vector. Other algorithms instead exploit the *Likelihood ratio* trick, which allows estimating the gradient from a single trajectory, but requires a stochastic policy. These can be step-based as REINFORCE (Williams 1992) or G(PO)MDP (Baxter and Bartlett 2001; Baxter et al. 2001), or episode-based as PEPG (Sehnke et al. 2010).

Policy gradients also include natural gradient algorithms (NPG), i.e., algorithms that try to limit the distance between distributions $P_\theta(h)$ and $P_{\theta+\delta\theta}(h)$ using the KL divergence (estimated through the Fisher information matrix (FIM)). In step-based NPGs (Bagnell and Schneider 2003; Peters and Schaal 2008b), using appropriate ("*compatible*") function approximation removes the need to estimate the FIM, but requires estimating the value function, which can be difficult. On the contrary, episodic Natural Actor-Critic (eNAC) (Peters and Schaal 2008a) uses complete episodes, and thus only estimates $v(s_1)$. NAC (Peters and Schaal 2008b) addresses infinite horizon problems, the lack of episodes leading to the use of Temporal Difference methods to estimate values.

Policy gradient usually applies to randomized policies. Recent work (Silver et al. 2014; Lillicrap et al. 2016) has adapted it to deterministic policies with a continuous action space, which can potentially facilitate the gradient estimation. Building on DQN, actor-critic methods have been extended to asynchronous updates with parallel actors and neural networks as approximators (Mnih et al. 2016).

Inference-based algorithms avoid the need to set learning rates. They consider that (i) the return R is an observed binary variable (1 meaning success),[2] (ii) the trajectory h is a latent variable, and (iii) one looks for the parameter vector that maximizes the probability of getting a return of 1. Then, an Expectation-Maximization algorithm can address this Bayesian inference problem. Variational inference can be used in the E-step of EM (Neumann 2011), but most approaches rely on Monte-Carlo estimates instead, despite the fact that they perform maximum likelihood estimates over several modes of the reward function (and thus do not distinguish them). These can be episode-based algorithms as RWR (Peters and Schaal 2007) (uses a linear upper-level policy) or CrKR (Kober et al. 2010) (a kernelized version of RWR, i.e., which does not need to specify feature vectors, but cannot model correlations). These can also be step-based algorithms as PoWER (Kober and Peters 2010), which allows a more structured exploration strategy, and gives more influence to data points with less variance.

Information-theoretic approaches (see chapter "Theoretical Computer Science: Computational Complexity" of Volume 3) try to limit changes in trajectory distributions between two consecutive time steps, which could correspond to degradations rather than improvements in the policy. Natural PGs have the same objective, but need a user-defined learning rate. Instead, REPS (Peters et al. 2010) combines advantages from NPG (smooth learning) and EM-based algorithms (no

[2] Transformations can bring us in this setting.

learning-rate). Episode-based REPS (Daniel et al. 2012) learns a higher-level policy while bounding parameter changes by solving a constrained optimization problem. Variants are able to adapt to multiple contexts or learn multiple solutions. Step-based REPS (Peters et al. 2010) solves an infinite horizon problem (rather than an episodic one), optimizing the average reward per time step. It requires enforcing the stationarity of state features, and thus solving another constrained optimization problem. A related recent method, TRPO (Schulman et al. 2015), which notably constrains the changes of $\pi(\cdot \mid s)$ instead of those of state-action distributions, proves to work well in practice.

Stochastic Optimization relies on black-box optimizers, and thus can easily be used for episode-based formulations, i.e., working with an upper-level policy $\pi_\omega(\theta)$. Typical examples are CEM (de Boer et al. 2005; Szita and Lörincz 2006), CMA-ES (Hansen et al. 2003; Heidrich-Meisner and Igel 2009), and NES (Wierstra et al. 2014), three evolutionary algorithms that maintain a parametric probability distribution (often Gaussian) $\pi_\omega(\theta)$ over the parameter vector. They sample a population of candidates, evaluate them, and use the best ones (weighted) to update the distribution. Many rollouts may be required for evaluation, as examplified with the game of Tetris (Szita and Lörincz 2006).

Path Integral (PI) approaches were introduced for optimal control, i.e., to handle non-linear continuous-time systems. They handle squared control costs and arbitrary state-dependent rewards. *Policy Improvement with PIs* (PI^2) applies PI theory to optimize Dynamic Movement Primitives (DMPs), i.e., representations of movements with parameterized differential equations, using Monte-Carlo rollouts instead of dynamic programming.

4.2 Model-Based Policy Search

Typical model-based policy-search approaches repeatedly (i) sample real-world trajectories using a fixed policy; (ii) learn a forward model of the dynamics based on these samples (and previous ones); (iii) optimize this policy using the learned model (generally as a simulator). As can be noted, this process does not explicitly handle the exploration-exploitation trade-off as policies are not chosen so as to improve the model where this could be appropriate. We now discuss three important dimensions of these approaches: how to learn the model, how to make reliable long-term predictions, and how to perform the policy updates.

Model learning often uses probabilistic models. They first allow accounting for uncertainty due to sparse data (at least in some areas) or an inappropriate model class. In robotics, where action and state spaces are continuous, non-parametric probabilistic methods can be used such as Linearly Weighted Bayesian Regression (LWBR) of Gaussian Processes (GPs), which may suffer from increasing time and memory requirements. But probabilistic models can also be employed to represent stochastic dynamics. An example is that of propositional problems, which are often modeled as Factored MDPs (Boutilier et al. 1995), where the dynamics and rewards

are DBNs whose structure is *a priori* unknown. A variety of approaches have been proposed, which rely on different representations (such as rule sets, decision trees, Stochastic STRIPS, or PPDDL) (Degris et al. 2006; Pasula et al. 2007; Walsh et al. 2009; Lesner and Zanuttini 2011). See chapter "Planning in Artificial Intelligence" of Volume 2.

Long-term predictions are usually required to optimize the policy given the current forward model. While the real world is its own best (unbiased) model, using a learned model has the benefit of allowing to control these predictions. A first approach, similar to paired statistical tests, is to always use the same random initial states and the same sequences of random numbers when evaluating different policies. It has been introduced for policy-search in the PEGASUS framework (Ng and Jordan 2000) and drastically reduces the sampling variance. Another approach is, when feasible, to compute a probability distribution over trajectories using deterministic approximations such as linearization (Anderson and Moore 2005), sigma-point methods (e.g., Julier and Uhlmann 2004) or moment-matching.

Policy updates can rely on gradient-free optimization (e.g., Nelder-Mead method or hill-climbing) (Bagnell and Schneider 2001), on sampling-based gradients (e.g., finite difference methods), as in model-free approaches, although they require many samples, or on analytical gradients (Deisenroth and Rasmussen 2011), which require the model as well as the policy to be differentiable, scale favorably with the number of parameters, but are computationally involved.

5 Extensions: Unknown Rewards and Risk-sensitive Criteria

In the previous sections, we recalled different techniques for solving RL problems, with the assumption that policies are compared with the expected cumulated rewards as a decision criterion. However, rewards may not be scalar, known or numeric, and the standard criterion based on expectation may not always be suitable. For instance, multiobjective RL has been proposed to tackle situations where an action is evaluated over several dimensions (e.g., duration, length, power consumption for a navigation problem). The interested reader may refer to Roijers et al. (2013) for a survey and refer to chapter "Multicriteria Decision Making" of this volume for an introduction to multicriteria decision-making. For space reasons, we focus below only on three extensions: reward learning (Sect. 5.1), preference-based RL (Sect. 5.2) and risk sensitive RL (Sect. 5.3).

5.1 Reward Learning

From the system designer's point of view, defining the reward function can be viewed as programming the desired behavior in an autonomous agent. A good choice of

reward values may accelerate learning (Matignon et al. 2006) while an incorrect choice may lead to unexpected and unwanted behaviors (Randløv and Alstrøm 1998). Thus, designing this function is a hard task (e.g., robotics (Argall et al. 2009), natural language parsers (Neu and Szepesvari 2009) or dialogue systems (El Asri et al. 2012)).

When the reward signal is not known, a natural approach is to learn from demonstration. Indeed, in some domains (e.g., autonomous driving), it is much simpler for an expert to demonstrate how to perform a task rather than specify a reward function. Such an approach has been called apprenticeship learning (Abbeel and Ng 2004), learning from demonstration (Argall et al. 2009), behavior cloning or imitation learning (Hussein et al. 2017). Two families of techniques have been developed to solve such problems. The first group tries to directly learn a good policy from (near) optimal demonstrations (Argall et al. 2009; Pomerleau 1989) while the second, called inverse RL (IRL) (Ng and Russell 2000; Russell 1998), tries to first recover a reward function that explains the demonstrations and then computes an optimal policy from it. The direct methods based on supervised learning usually suffer when the reward function is sparse and even more when dynamics is also perturbed (Piot et al. 2013).

As the reward function is generally considered to be a more compact, robust and transferable representation of a task than a policy (Abbeel and Ng 2004; Russell 1998), we only discuss reward learning approaches here.

As for many inverse problems, IRL is ill-posed: any constant function is a trivial solution that makes all policies equivalent and therefore optimal. Various solutions were proposed to tackle this degeneracy issue, differing on whether a probabilistic model is assumed or not on the generation of the observation. When the state and/or action spaces are large, the reward function is generally assumed to take a parametric form: $R(s,a) = f_{\boldsymbol{\theta}}(s,a)$ for $f_{\boldsymbol{\theta}}$ a parametric function of $\boldsymbol{\theta}$. One important case, called *linear features*, is when f is linear in $\boldsymbol{\theta}$, i.e., $R(s,a) = \sum_i \theta_i \phi_i(s,a)$ where ϕ_i are basis functions.

No generative model assumption. As underlined by Neu and Szepesvari (2009), many IRL methods can be viewed as finding the reward function R that minimizes a dissimilarity measure between the policy π_R^* optimal for R and the expert demonstrations. Most work assume a linear-feature reward function, with some exceptions that we mention below. Abbeel and Ng (2004) introduced the important idea of expected feature matching, which says that the expected features of π_R^* and those estimated from the demonstrations should be close. Thus, they notably proposed the projection method, which amounts to minimizing the Euclidean distance between those two expected features. Neu and Szepesvari (2007) proposed a natural gradient method for minimizing this objective function. Syed and Schapire (2008) reformulated the projection method problem as a zero-sum two-player game, with the nice property that the learned policy may perform better than the demonstrated one. Abbeel and Ng (2004)'s work was extended to the partially observable case (Choi and Kim 2011).

Besides, (Ratliff et al., 2006) proposed a max-margin approach enforcing that the found solution is better than any other one by at least a margin. Interestingly, the

method can learn from multiple MDPs. It was later extended to the non-linear feature case (Ratliff et al. 2007).

Another technique (Klein et al. 2012; Piot et al. 2014) consists in learning a classifier based on a linearly parametrized score function to predict the best action for a state given the set of demonstrations. The learned score function can then be interpreted as a value function and can be used to recover a reward function.

Traditional IRL methods learn from (near) optimal demonstration. More recent approaches extend IRL to learn from other types of observations, e.g., a set of (non-necessarily optimal) demonstrations rated by an expert (El Asri et al. 2016; Burchfield et al. 2016), bad demonstrations (Sebag et al. 2016) or pairwise comparisons (da Silva et al. 2006; Wirth and Neumann 2015). In the latter case, the interactive setting is investigated with a reliable expert (Chernova and Veloso 2009) or unreliable one (Weng et al. 2013).

Generative model assumption. Another way to tackle the degeneracy issue is to assume a probabilistic model on how observations are generated. Here, most work assumes that the expert policy is described by Boltzmann distributions, where higher-valued actions are more probable. Two notable exceptions are the work of Grollman and Billard (2011), which shows how to learn from failed demonstrations using Gaussian mixture models, and the Bayesian approach of Ramachandran and Amir (2007), with the assumption that state-action pairs in demonstrations follow such a Boltzmann distribution. This latter approach has been extended to Boltzmann distribution-based expert policy and for multi-task learning (Dimitrakakis and Rothkopf 2011), and to account for multiple reward functions (Choi and Kim 2012). This Bayesian approach has been investigated to interactive settings where the agent can query for an optimal demonstration in a chosen state (Lopes et al. 2009) or for a pairwise comparison (Wilson et al. 2012; Akrour et al. 2013, 2014).

Without assuming a prior, Babes-Vroman et al. (2011) proposed to recover the expert reward function by maximum likelihood. The approach is able to handle the possibility of multiple intentions in the demonstrations. Furthermore, Nguyen et al. (2015) suggested an Expectation-Maximization approach to learn from demonstration induced by locally consistent reward functions.

To tackle the degeneracy issue, Ziebart et al. (2010) argued for the use of the maximum entropy principle, which states that among all solutions that fit the observations, the least informative one (i.e., maximum entropy) should be chosen, with the assumption that a reward function induces a Boltzmann probability distribution over trajectories. When the transition function is not known, Boularias et al. (2011) extended this approach by proposing to minimize the relative entropy between the probability distribution (over trajectories) induced by a policy and a baseline distribution under an expected feature matching constraint. Wulfmeier et al. (2015) extended this approach to the case where a deep neural network is used for the representation of the reward function, while Bogert et al. (2016) took into account non-observable variables.

5.2 Preference-Based Approaches

Another line of work redefines policy optimality directly based on pairwise comparisons of histories without assuming the existence of a scalar numeric reward function. This notably accounts for situations where reward values and probabilities are not commensurable. In this context, different decision criteria e.g., quantile (Gilbert and Weng 2016) may be used. One popular decision model (Yue et al. 2012; Fürnkranz et al. 2012) is defined as follows: a policy π is preferred to another policy π' if

$$\mathbb{P}[h^{\pi} \succsim h^{\pi'}] \geq \mathbb{P}[h^{\pi'} \succsim h^{\pi}], \tag{9}$$

where $\succsim$ is a preorder over histories, h^{π} is a random variable representing the history generated by policy π and therefore $\mathbb{P}[h^{\pi} \succsim h^{\pi'}]$ is the probability that a history generated by π is not less preferred than a history generated by π'. Based on (9), Fürnkranz et al. (2012) proposed a policy iteration algorithm. However, one crucial issue with (9) is that the concept of optimal solution is not well-defined as (9) can lead to preference cycles (Gilbert et al. 2015). Busa-Fekete et al. (2014) circumvented this problem by refining this decision model with criteria from social choice theory. In Gilbert et al. (2015), the issue was solved by considering mixed solutions: an optimal mixed solution is guaranteed to exist by interpreting it as a Nash equilibrium of a two-player zero-sum game. Gilbert et al. (2016) proposed a model-free RL algorithm based on a two-timescale technique to find such a mixed optimal solution.

5.3 Risk-Sensitive Criteria

Taking into account risk is important in decision-making under uncertainty (see chapter "Decision under Uncertainty" of this volume). The standard criterion based on expectation is risk-neutral. When it is known that a policy will only be used a few limited number of times, variability in the obtained rewards should be penalized. Besides, in some hazardous domains, good policies need to absolutely avoid bad or error states. In those two cases, preferences over policies need to be defined to be risk-sensitive.

In its simplest form, risk can directly be represented as a probability. For instance, Geibel and Wysotzky (2005) adopted such an approach and consider MDP problems with two objectives where the first objective is the standard decision criterion and the second objective is to minimize the probability of reaching a set of bad states.

A more advanced approach is based on risk-sensitive decision criteria (Barbera et al. 1999). Variants of Expected Utility (Machina 1988), which is the standard risk-sensitive criterion, were investigated in two cases when the utility function is exponential (Borkar 2010; Moldovan and Abbeel 2012) and when it is quadratic (Tamar et al. 2012, 2013; Gosavi 2014). In the latter case, the criterion amounts to penalizing the standard criterion by the variance of the cumulated reward. While the

usual approach is to transform the cumulated reward, Mihatsch and Neuneier (2002) proposed to directly transform the temporal differences during learning.

Other approaches consider risk measures (Denuit et al. 2006) and in particular coherent risk measures (Artzner et al. 1999). Value-at-risk, popular in finance, was considered in Gilbert and Weng (2016). Policy gradient methods (Chow and Ghavamzadeh 2014; Tamar et al. 2015b) were proposed to optimize Conditional Value-at-Risk (CVaR) and were extended to any coherent risk measure (Tamar et al. 2015a). Jiang and Powell (2017) proposed dynamic quantile-based risk measures, which encompasses VaR and CVaR, and investigated an approximate dynamic programming scheme to optimize them.

In risk-constrained problems, the goal is to maximize the expectation of return while bounding a risk measure. For variance-constrained problems, Prashanth and Ghavamzadeh (2016) proposed an actor-critic algorithm. For CVaR-constrained problems, Borkar and Jain (2014) proposed a two-timescale stochastic approximation technique, while Chow et al. (2016) investigated policy gradient and actor-critic methods.

One important issue to consider when dealing with risk-sensitive criteria is that the Bellman optimality principle generally does not hold anymore: a sub-policy of an optimal risk-sensitive policy may not be optimal. However, in most cases, the Bellman optimality principle may be recovered by considering a state-augmented MDP, where the state includes the rewards cumulated so far (Liu and Koenig 2006).

6 Conclusion

Recently, thanks to a number of success stories, reinforcement learning (RL) has become a very active research area. In this chapter, we recalled the basic setting of RL. Our focus was to present an overview of the main techniques, which can be divided into value-based and policy search methods, for solving large-sized RL problems with function approximation. We also presented some approaches for tackling the issue of unknown rewards that a system designer would encounter in practice and recalled some recent work in RL when risk-sensitivity needs to be taken into account in decision-making.

Currently RL still has too large sample and computational requirements for many practical domains (e.g., robotics). Research work is very active on improving RL algorithms along those two dimensions, notably by exploiting the structure of the problem (Kulkarni et al. 2016) or other a priori knowledge, expressed in temporal logic (Wen et al. 2017) for instance, or by reusing previous learning experience with transfer learning (Taylor and Stone 2009), lifelong learning (Bou Ammar et al. 2015), multi-task learning (Wilson et al. 2007) or curriculum learning (Wu and Tian 2017), to cite a few. Having more efficient RL algorithms is important as it will pave the way to more applications in more realistic domains.

References

Abbeel P, Coates A, Ng AY (2010) Autonomous helicopter aerobatics through apprenticeship learning. Int J Robot Res 29(13):1608–1639

Abbeel P, Ng A (2004) Apprenticeship learning via inverse reinforcement learning. In: International conference machine learning

Akrour R, Schoenauer M, Sebag M (2013) ECML PKDD. Interactive robot education. Lecture notes in computer science

Akrour R, Schoenauer M, Souplet J-C, Sebag M (2014) Programming by feedback. In: ICML

Anderson BDO, Moore JB (2005) Optimal filtering. Dover Publications

Antos A, Szepesvári C, Munos R (2008) Fitted Q-iteration in continuous action-space MDPs. In: Advances in neural information processing systems, pp 9–16

Argall B, Chernova S, Veloso M, Browning B (2009) A survey of robot learning from demonstration. Robot Auton Syst 57(5):469–483

Artzner P, Delbaen F, Eber J, Heath D (1999) Coherent measures of risk. Math Financ 9(3):203–228

Babes-Vroman M, Marivate V, Subramanian K, Littman M (2011) Apprenticeship learning about multiple intentions. In: ICML

Bagnell JA, Schneider JG (2001) Autonomous helicopter control using reinforcement learning policysearch methods. In: Proceedings of the international conference on robotics and automation, pp 1615–1620

Bagnell JA, Schneider JG (2003) Covariant policy search. In: Proceedings of the international joint conference on artifical intelligence

Bai A, Wu F, Chen X (2013) Towards a principled solution to simulated robot soccer. In: RoboCup-2012: robot soccer world cup XVI. Lecture notes in artificial intelligence, vol 7500

Baird L et al (1995) Residual algorithms: Reinforcement learning with function approximation. In: Proceedings of the twelfth international conference onmachine learning, pp 30–37

Barbera S, Hammond P, Seidl C (1999) Handbook of utility theory. Springer, Berlin

Bäuerle N, Rieder U (2011) Markov decision processes with applications to finance. Springer Science and Business Media

Baxter J, Bartlett P (2001) Infinite-horizon policy-gradient estimation. J Artif Intell Res 15:319–350

Baxter J, Bartlett P, Weaver L (2001) Experiments with infinite-horizon, policy-gradient estimation. J Artif Intell Res 15:351–381

Bellman R, Dreyfus S (1959) Functional approximations and dynamic programming. Math Tables Aids Comput 13(68):247–251

Bellman R, Kalaba R, Kotkin B (1963) Polynomial approximation-a new computational technique in dynamic programming: allocation processes. Math Comput 17(82):155–161

Bogert K, Lin JF-S, Doshi P, Kulic D (2016) Expectation-maximization for inverse reinforcement learning with hidden data. In: AAMAS

Bojarski M, Testa DD, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang J, Zhang X, Zhao J (2016) End to end learning for self-driving cars. Technical report, NVIDIA

Borkar V, Jain R (2014) Risk-constrained Markov decision processes. IEEE Trans Autom Control 59(9):2574–2579

Borkar VS (2010) Learning algorithms for risk-sensitive control. In: International symposium on mathematical theory of networks and systems

Bou Ammar H, Tutunov R, Eaton E (2015) Safe policy search for lifelong reinforcement learning with sublinear regret. In: ICML

Boularias A, Kober J, Peters J (2011) Relative entropy inverse reinforcement learning. In: AISTATS

Boutilier C, Dearden R, Goldszmidt M (1995) Exploiting structure in policy construction. In: Proceedings of the fourteenth international joint conference on artificial intelligence, pp 1104–1111

Boutilier C, Dearden R, Goldszmidt M (2000) Stochastic dynamic programming with factored representations. Artif Intell 121(1–2):49–107

Bradtke SJ, Barto AG (1996) Linear least-squares algorithms for temporal difference learning. Machine Learning 22:33–57
Burchfield B, Tomasi C, Parr R (2016) Distance minimization for reward learning from scored trajectories. In: AAAI
Busa-Fekete R, Szörenyi B, Weng P, Cheng W, Hüllermeier E (2014) Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. Mach Learn 97(3):327–351
Busoniu L, Babuska R, De Schutter B (2010) Innovations in multi-agent systems and applications – 1, vol 310, chapter Multi-agent reinforcement learning: an overview, Springer, Berlin, pp 183–221
Chernova S, Veloso M (2009) Interactive policy learning through confidence-based autonomy. J Artif Intell Res 34:1–25
Choi D, Van Roy B (2006) A generalized Kalman filter for fixed point approximation and efficient temporal-difference learning. Discret Event Dyn Syst 16(2):207–239
Choi J, Kim K-E (2011) Inverse reinforcement learning in partially observable environments. JMLR 12:691–730
Choi J, Kim K-E (2012) Nonparametric Bayesian inverse reinforcement learning for multiple reward functions. In: NIPS
Chow Y, Ghavamzadeh M (2014) Algorithms for CVaR optimization in MDPs. In: NIPS
Chow Y, Ghavamzadeh M, Janson L, Pavone M (2016) Risk-constrained reinforcement learning with percentile risk criteria. JMLR 18(1)
da Silva VF, Costa AHR, Lima P (2006) Inverse reinforcement learning with evaluation. In: IEEE ICRA
Daniel C, Neumann G, Peters J (2012) Hierarchical relative entropy policy search. In: Proceedings of the international conference of artificial intelligence and statistics, pp 273–281
de Boer P, Kroese D, Mannor S, Rubinstein R (2005) A tutorial on the cross-entropy method. Ann Oper Res 134(1):19–67
de Farias D, Van Roy B (2003) The linear programming approach to approximate dynamic programming. Oper Res 51(6):850–865
Degris T, Sigaud O, Wuillemin P-H (2006) Learning the structure of factored Markov decision processes in reinforcement learning problems. In: Proceedings of the 23rd international conference on machine learning
Deisenroth MP, Neumann G, Peters J (2011) A survey on policy search for robotics. Found Trends Robot 2(1–2):1–142
Deisenroth MP, Rasmussen CE (2011) PILCO: a model-based and data-efficient approach to policy search. In: Proceedings of the international conference on machine learning, pp 465–472
Denuit M, Dhaene J, Goovaerts M, Kaas R, Laeven R (2006) Risk measurement with equivalent utility principles. Stat Decis 24:1–25
Dimitrakakis C, Rothkopf CA (2011) Bayesian multitask inverse reinforcement learning. In: EWRL
El Asri L, Laroche R, Pietquin O (2012) Reward function learning for dialogue management. In: STAIRS
El Asri L, Piot B, Geist M, Laroche R, Pietquin O (2016) Score-based inverse reinforcement learning. In: AAMAS
Engel Y, Mannor S, Meir R (2005) Reinforcement learning with Gaussian processes. In: Proceedings of the 22nd international conference on Machine learning, ACM, pp 201–208
Ernst D, Geurts P, Wehenkel L (2005) Tree-based batch mode reinforcement learning. J Mach Learn Res 6(Apr):503–556
Fürnkranz J, Hüllermeier E, Cheng W, Park S (2012) Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. Mach Learn 89(1):123–156
Geibel P, Wysotzky F (2005) Risk-sensitive reinforcement learning applied to control under constraints. JAIR 24:81–108
Geist M, Pietquin O (2010a) Kalman temporal differences. J Artif Intell Res 39:483–532

Geist M, Pietquin O (2010b) Statistically linearized least-squares temporal differences. In: 2010 international congress on ultra modern telecommunications and control systems and workshops (ICUMT), IEEE, pp 450–457
Geist M, Pietquin O (2011) Parametric value function approximation: a unified view. In: ADPRL
Geist M, Pietquin O (2013) Algorithmic survey of parametric value function approximation. IEEE Trans Neural Netw Learn Syst 24(6):845–867
Ghavamzadeh M, Mannor S, Pineau J, Tamar A (2015) Bayesian reinforcement learning: a survey. Found Trends Mach Learn 8(5–6):359–492
Gilbert H, Spanjaard O, Viappiani P, Weng P (2015) Solving MDPs with skew symmetric bilinear utility functions. In: International joint conference in artificial intelligence (IJCAI), pp 1989–1995
Gilbert H, Weng P (2016) Quantile reinforcement learning. In: Asian workshop on reinforcement learning
Gilbert H, Zanuttini B, Viappiani P, Weng P, Nicart E (2016) Model-free reinforcement learning with skew-symmetric bilinear utilities. In: International conference on uncertainty in artificial intelligence (UAI)
Gordon GJ (1995) Stable function approximation in dynamic programming. In: Proceedings of the twelfth international conference onmachine learning, pp 261–268
Gosavi AA (2014) Variance-penalized Markov decision processes: dynamic programming and reinforcement learning techniques. Int J General Syst 43(6):649–669
Grollman DH, Billard A (2011) Donut as I do: learning from failed demonstrations. In: IEEE ICRA
Guestrin C, Hauskrecht M, Kveton B (2004) Solving factored MDPs with continuous and discrete variables. In: AAAI, pp 235–242
Hansen N, Muller S, Koumoutsakos P (2003) Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evol Comput 11(1):1–18
Heidrich-Meisner V, Igel C (2009) Neuroevolution strategies for episodic reinforcement learning. J Algorithms 64(4):152–168
Hussein A, Gaber MM, Elyan E, Jayne C (2017) Imitation learning: a survey of learning methods. ACM Comput Surv
Jiang DR, Powell WB (2017) Risk-averse approximate dynamic programming with quantile-based risk measures. Math Oper Res 43(2):347–692
Julier SJ, Uhlmann JK (2004) Unscented filtering and nonlinear estimation. Proc IEEE 92(3):401–422
Klein E, Geist M, Piot B, Pietquin O (2012) Inverse reinforcement learning through structured classification. In: NIPS
Kober J, Oztop E, Peters J (2010) Reinforcement learning to adjust robot movements to new situations. In: Proceedings of the 2010 robotics: science and systems conference
Kober J, Peters J (2010) Policy search for motor primitives in robotics. Mach Learn 1–33
Kulkarni T, Narasimhan KR, Saeedi A, Tenenbaum J (2016) Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation. In: NIPS
Lagoudakis MG, Parr R (2003) Least-squares policy iteration. J Mach Learn Res 4(Dec):1107–1149
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
Lesner B, Zanuttini B (2011) Handling ambiguous effects in action learning. In: Proceedings of the 9th European workshop on reinforcement learning, p 12
Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2016) Continuous control with deep reinforcement learning. In: ICLR
Lin L-H (1992) Self-improving reactive agents based on reinforcement learning, planning and teaching. Mach Learn 8(3/4):69–97
Liu Y, Koenig S (2006) Functional value iteration for decision-theoretic planning with general utility functions. In: AAAI, AAAI, pp 1186–1193
Lopes M, Melo F, Montesano L (2009) Active learning for reward estimation in inverse reinforcement learning. In: ECML/PKDD. vol 5782, Lecture notes in computer science, pp 31–46
Machina M (1988) Expected utility hypothesis. In: Eatwell J, Milgate M, Newman P (eds) The new palgrave: a dictionary of economics. Macmillan, pp 232–239

Matignon L, Laurent GJ, Le Fort-Piat N (2006) Reward function and initial values: better choices for accelerated goal-directed reinforcement learning. Lect Notes CS 1(4131):840–849
Mihatsch O, Neuneier R (2002) Risk-sensitive reinforcement learning. Mach Learn 49:267–290
Mnih V, Badia AP, Mirza M, Graves A, Lillicrap TP, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement. learning. In: ICML
Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. Nature 518:529–533
Moldovan T, Abbeel P (2012) Risk aversion Markov decision processes via near-optimal Chernoff bounds. In: NIPS
Neu G, Szepesvari C (2007) Apprenticeship learning using inverse reinforcement learning and gradient methods. In: UAI
Neu G, Szepesvari C (2009) Training parsers by inverse reinforcement learning. Mach Learn 77:303–337
Neumann G (2011) Variational inference for policy search in changing situations. In: Proceedings of the international conference on machine learning, pp 817–824
Ng A, Russell S (2000) Algorithms for inverse reinforcement learning. In: ICML, Morgan Kaufmann
Ng AY, Jordan M (2000) PEGASUS : A policy search method for large MDPs and POMDPs. In: Proceedings of the conference on uncertainty in artificial intelligence
Nguyen QP, Low KH, Jaillet P (2015) Inverse reinforcement learning with locally consistent reward functions. In: NIPS
Pasula HM, Zettlemoyer LS, Kaelbling LP (2007) Learning symbolic models of stochastic domains. J Artif Intell Res 29:309–352
Peters J, Mülling K, Altun Y (2010) Relative entropy policy search. In: Proceedings of the national conference on artificial intelligence
Peters J, Schaal S (2007) Applying the episodic natural actor-critic architecture to motorprimitive learning. In: Proceedings of the European symposium on artificial neural networks
Peters J, Schaal S (2008a) Natural actor-critic. Neurocomputation 71(7–9):1180–1190
Peters J, Schaal S (2008b) Reinforcement learning of motor skills with policy gradients. Neural Netw 4:682–697
Piot B, Geist M, Pietquin O (2013) Learning from demonstrations: is it worth estimating a reward function? In: ECML PKDD, Lecture notes in computer science
Piot B, Geist M, Pietquin O (2014) Boosted and Reward-regularized classification for apprenticeship learning. In: AAMAS, France, Paris, pp 1249–1256
Pomerleau D (1989) Alvinn: an autonomous land vehicle in a neural network. In: NIPS
Prashanth L, Ghavamzadeh M (2016) Variance-constrained actor-critic algorithms for discounted and average reward MDPs. Mach Learn
Puterman M (1994) Markov decision processes: discrete stochastic dynamic programming. Wiley, New York
Ramachandran D, Amir E (2007) Bayesian inverse reinforcement learning. In: IJCAI
Randløv J, Alstrøm P (1998) 1998. Learning to drive a bicycle using reinforcement learning and shaping. In: ICML
Ratliff N, Bagnell J, Zinkevich M (2006) Maximum margin planning. In: ICML
Ratliff N, Bradley D, Bagnell JA, Chestnutt J (2007) Boosting structured prediction for imitation learning. In: NIPS
Riedmiller M (2005) Neural fitted Q iteration-first experiences with a data efficient neural reinforcement learning method. In: ECML, vol 3720. Springer, Berlin, pp 317–328
Roijers D, Vamplew P, Whiteson S, Dazeley R (2013) A survey of multi-objective sequential decision-making. J Artif Intell Res 48:67–113
Russell S (1998) Learning agents for uncertain environments. In: Proceedings of the eleventh annual conference on Computational learning theory, ACM, pp 101–103

Samuel A (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev 3(3):210–229
Schaul T, Quan J, Antonoglou I, Silver D (2016) Prioritized experience replay. In: ICLR
Schulman J, Levine S, Abbeel P, Jordan M, Moritz P (2015) Trust region policy optimization. In: ICML
Sebag M, Akrour R, Mayeur B, Schoenauer M (2016) Anti imitation-based policy learning. In: ECML PKDD, Lecture notes in computer science
Sehnke F, Osendorfer C, Rückstieß T, Graves A, Peters J, Schmidhuber J (2010) Parameter-exploring policy gradients. Neural Netw 23(4):551–559
Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneerschelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489
Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M (2014) Deterministic policy gradient algorithms. In: ICML
Singh S, Kearns M, Litman D, Walker M (1999) Reinforcement learning for spoken dialogue systems. In: NIPS
Spaan MT (2012) Reinforcement Learning, chapter Partially observable Markov decision processes. Springer, Berlin
Sutton R, Maei H, Precup D, Bhatnagar S, Silver D, Szepesvári C, Wiewiora E (2009) Fast gradient-descent methods for temporal-difference learning with linear function approximation. In: ICML
Syed U, Schapire RE (2008) A game-theoretic approach to apprenticeship learning. In: NIPS
Szita I, Lörincz A (2006) Learning tetris using the noisy cross-entropy method. Neural Comput 18:2936–2941
Tamar A, Chow Y, Ghavamzadeh M, Mannor S (2015a) Policy gradient for coherent risk measures. In: NIPS
Tamar A, Di Castro D, Mannor S (2012) Policy gradient with variance related risk criteria. In: ICML
Tamar A, Di Castro D, Mannor S (2013) Temporal difference methods for the variance of the reward to go. In: ICML
Tamar A, Glassner Y, Mannor S (2015b) Optimizing the CVaR via sampling. In: AAAI
Taylor ME, Stone P (2009) Transfer learning for reinforcement learning domains: a survey. J Mach Learn Res 10:1633–1685
Tesauro G (1995) Temporal difference learning and td-gammon. Commun ACM 38(3):58–68
Van Hasselt H, Guez A, Silver D (2016) Deep reinforcement learning with double q-learning. In: AAAI, pp 2094–2100
van Otterlo M (2009) The logic of adaptive behavior. IOS
Walsh T, Szita I, Diuk C, Littman M (2009) Exploring compact reinforcement-learning representations with linear regression. In: Proceedings of the 25th conference on uncertainty in artificial intelligence
Wen M, Papusha I, Topcu U (2017) Learning from demonstrations with high-level side information. In: IJCAI
Weng P, Busa-Fekete R, Hüllermeier E (2013) Interactive Q-learning with ordinal rewards and unreliable tutor. In: Workshop reinforcement learning with. generalized feedback, ECML/PKDD
Werbos PJ (1990) Consistency of HDP applied to a simple reinforcement learning problem. Neural Netw 3:179–189
Wierstra D, Schaul T, Glasmachers T, Sun Y, Peters J, Schmidhuber J (2014) Natural evolution strategies. JMLR 15:949–980
Williams R (1992) Simple statistical gradient-following algorithms for connectionnist reinforcement learning. Mach Learn 8(3):229–256
Wilson A, Fern A, Ray S, Tadepalli P (2007) Multi-task reinforcement learning: A hierarchical Bayesian approach. In: ICML

Wilson A, Fern A, Tadepalli P (2012) A Bayesian approach for policy learning from trajectory preference queries. In: Advances in neural information processing systems
Wirth C, Neumann G (2015) Model-free preference-based reinforcement learning. In: EWRL
Wu Y, Tian Y (2017) Training agent for first-person shooter game with actor-critic curriculum learning. In: ICLR
Wulfmeier M, Ondruska P, Posner I (2015) Maximum entropy deep inverse reinforcement learning. In: NIPS, Deep reinforcement learning workshop
Xu X, Hu D, Lu X (2007) Kernel-based least squares policy iteration for reinforcement learning. IEEE Trans Neural Netw 18(4):973–992
Yu T, Zhang Z (2013) Optimal CPS control for interconnected power systems based on SARSA on-policy learning algorithm. In: Power system protection and control, pp 211–216
Yue Y, Broder J, Kleinberg R, Joachims T (2012) The k-armed dueling bandits problem. J Comput Syst Sci 78(5):1538–1556
Zhao Q, Chen S, Leung S, Lai K (2010) Integration of inventory and transportation decisions in a logistics system. Transp Res Part E: Logist Transp Rev 46(6):913–925
Ziebart B, Maas A, Bagnell J, Dey A (2010) Maximum entropy inverse reinforcement learning. In: AAAI

Argumentation and Inconsistency-Tolerant Reasoning

Leila Amgoud, Philippe Besnard, Claudette Cayrol, Philippe Chatalic and Marie-Christine Lagasquie-Schiex

Abstract This chapter is devoted to logical models for reasoning from contradictory information. It deals with methods, such as argumentation, that refrain from giving up any piece of information (by contrast with revision, as discussed in chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume). The baseline is to get the best, resorting to various possibilities, from the available information in order to reason in the most sensible way despite contradictions.

1 Introduction

Intelligent agents reason. Should they also be autonomous, they have to be able to reason no matter what the available information, and whatever defects may plague available information. Chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" of this volume deals with the case that not enough information is available. This chapter deals with the case that there is too much information—but not in the sense of an excessive amount (this chapter is not concerned with a situation in which there is plethora of information or there are pieces of information such that one of them entails the others). How could there be "too much information" then? Well, when two pieces of information contradict each other, there seems to be undue information. This chapter is devoted to approaches to reasoning in

L. Amgoud (✉) · P. Besnard · C. Cayrol · M.-C. Lagasquie-Schiex
IRIT-CNRS, Université Paul Sabatier, Toulouse, France
e-mail: amgoud@irit.fr

P. Besnard
e-mail: besnard@irit.fr

C. Cayrol
e-mail: ccayrol@irit.fr

M.-C. Lagasquie-Schiex
e-mail: lagasq@irit.fr

P. Chatalic
LRI, Université Paris-Sud-CNRS, Université Paris-Saclay, Orsay, France
e-mail: chatalic@lri.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_13

the presence of contradictory information. Reasoning from contradictory statements is not to be confused with reasoning that merely mentions contradictory statements. As an illustration for the latter, from the fact that "it rains" and the principle that "if it snows then it does not rain", the *assumption* that "it snows" implies that "it does not rain", in contradiction with the fact that "it rains"; this refutes the assumption giving rise to this contradiction, and it can be concluded that "it does not snow". This chapter focuses on reasoning involving contradictory statements such that these statements are to be taken for granted instead of being of hypothetical nature.

An example of such reasoning is getting, on the one hand, the information that Mr Thestre is tall, and, and the other hand, that Mr Thestre is short. Question is, what to do, then, in terms of handling information and reasoning?

The answer "eradicate bad information" falls short of settling the matter. First, although the terms of a contradiction may happen to originate from well-identified sources (with the benefit that such extra data can sometimes be used to solve dilemmas), this is not always the case. In particular, a knowledge base needs not have been created with recording of the source for each item in the base. For instance, the format of items may preclude such recording, or knowledge of the source may have been unknown or lost. Moreover, even in the case that each piece of information comes with a well-identified source, this needs not be enough to discriminate "bad information" (also, despite some rather widespread prejudice, introducing a temporal representation is no panacea).

All these reasons underlie the study and development of computational models of reasoning from contradictory information.

2 Reasoning from Inconsistent Information

2.1 Introduction

Having two contradictory pieces of information is not always a disaster, though there are better situations.

For instance, "he is the elder of the family" and "only one of his brothers is older than him". It is not likely to affect other pieces of information, even less if this information has very little to do with the above-mentioned statements; for instance, "my supervisor is leaving tomorrow for a family holiday".

Unfortunately, this observation cannot be extended to a classical logic formalization. If two contradictory statements E and E' are formalized by two classical logic formulae A and A', such that there exists a consequence B of A with the negation of B being a consequence of A', then the consequences of the conjunction $A \wedge A'$ are all the classical logic formulae!

Under these conditions, reasoning has no value: each formula being a consequence, any statement would be a valid conclusion. It is obviously inaccept-

able for a reasoning model. So, inconsistent information is harmful for classical deduction.

However, when an autonomous agent reasons and interacts with its environment, she may be faced with different sources of inconsistency: false beliefs, unreliable observations, exchange of information with other agents having diverging opinions. As shown above, an intelligent agent must give up classical deduction in order to exploit her knowledge base, when the base contains contradictory formulae. So an autonomous intelligent agent must be provided with mechanisms for reasoning from inconsistent information.

Possible options range from getting rid of classical logic as a formal reasoning model, to different methods for pre-processing contradictory information before applying classical logic. This range of options is the topic of the following section.

2.2 Models for Reasoning from Inconsistency

Reasoning from inconsistent pieces of information, represented as logical formulae, is a fundamental issue in Artificial Intelligence. Its importance is reflected by the number of approaches developed so far: belief revision, belief merging, reasoning from preferred consistent subbases, as well as paraconsistent logics and argumentative formalisms.

Two kinds of approaches have emerged, corresponding to two attitudes in front of inconsistent knowledge. The first one is to avoid inconsistency by modifying the contents of the knowledge base, using extra knowledge. The second one is to accept the available knowledge and to cope with inconsistency. This attitude is particularly relevant when it is not possible to get new pieces of information.

Within the first kind of approaches, two directions have been followed:

- In order to avoid inconsistency: when the pieces of information are introduced successively, belief revision offers the most appropriate setting, and when conflicting beliefs stem from different sources, belief merging is the right setting (these formalisms are presented in chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume).
- Once inconsistency has been discovered, it is still possible to remove it by acquiring additional knowledge. The idea is to identify wrong pieces of belief through knowledge-gathering actions (also called tests) or to check some of the sources, in order to retrieve a unique consistent knowledge base (Lang and Marquis 2000; Konieczny et al. 2003).

These approaches will not be further discussed in this chapter, which is mainly devoted to formalisms able to cope with inconsistency without removing any piece of information. The issue is to deal with an available knowledge base containing contradictory pieces of information. So, the set of consequences that can be derived

from the given base must be weakened. This can be achieved either by weakening the input base while keeping classical entailment, or by weakening the consequence relation of classical logic while keeping the input base intact.

2.2.1 Weakening the Input Base

The first idea is to select one or several consistent subsets of the input base and then to apply classical entailment for inferring from these subsets. Note that no piece of information is removed from the input base. Weakening merely consists in inhibiting some pieces of information when computing consistent subsets. Hence the name of virtual restoration of consistency, or weakening by inhibition.

Restoring consistency consists in computing the preferred consistent subsets of the input base. Several criteria for defining preferred subsets have been proposed, ranging from the maximality for set-inclusion to criteria induced by a priority ordering between the formulae of the input base.

This technique goes back to Rescher and Manor (1970), and has been extensively developed from the work of Poole (1988), Brewka (1989). Reasoning by virtual restoration of consistency is the topic of Sect. 3.

Variable forgetting (Lang and Marquis 2010) is another more recent technique for weakening pieces of information so as to restore consistency. The idea is that inconsistency may be caused by information carried by some of the variables. Then consistency can be restored by focussing on the variables "responsible for inconsistency" and simplifying the knowledge base by ignoring these variables.

A given formula is weakened by forgetting a set of variables in this formula. As for weakening by inhibition, several criteria for choosing these sets of variables can be proposed, depending for instance on a priority ordering on the variables.

Note that in the weakening based approaches, the available knowledge is represented by a set of formulae, which is not interpreted as a conjunction of formulae. That is why these approaches are often referred to as "syntax-based" approaches.

2.2.2 Weakening the Consequence Relation

The idea is to accept inconsistent information and to propose systems for reasoning in presence of inconsistency. These systems are generally based on paraconsistent logics, or argumentative formalisms.

Paraconsistent logics avoid trivialization by weakening classical entailment. They are presented in Sect. 4.

Argumentative logics have been introduced by Elvang-Gøransson and Hunter (1995) for reasoning with classically inconsistent information. The idea is to justify each inference that follows from consistent subsets of the inconsistent input base. The justification is based on the notions of argument and acceptability of an argument: An argument is a consistent subset of the input base together with the inference of a conclusion from that subset.

Due to the inconsistency of the input base, multiple conflicting arguments may be produced. Then, acceptability criteria enable to differentiate arguments and to select the most acceptable of them. Only conclusions justified by acceptable arguments are finally considered.

A comparative study between approaches for virtual consistency restoration and argumentative approaches to reasoning from inconsistency was carried out in (Cayrol 1995; Amgoud 2012a, b). More generally, argumentation systems are presented in Sect. 5.

Following the same line of accepting inconsistency, another approach has been proposed in a decentralized setting (Chatalic et al. 2006).

The available information is distributed between several agents (called peers). Each peer has her own knowledge base, which is assumed to be consistent, and also hosts information about other peers. The global knowledge base may be inconsistent even if the base of each peer is consistent.

Peer-to-peer reasoning consists in computing well-founded consequences of a formula, i.e. consequences of the formula w.r.t. a consistent subset of the global knowledge base. Such a computation is novel in the sense that it is decentralized and distributed between different peers, without any global control.

This approach will be presented in Sect. 6.

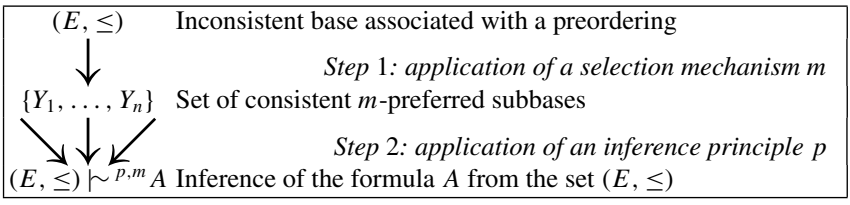
3 Reasoning Based on Virtual Restoration of Consistency

3.1 Introduction

The aim of this approach is to remove "virtually" some pieces of information from an inconsistent base in order to "restore consistency" (this removal is said virtual since the base is not really modified). This approach follows two steps:

- first a *selection mechanism* is needed for *choosing* the pieces of information that must be removed; generally, this is done using a preference relation over these pieces of information (generally a preordering); the use of this mechanism induces the production of some "preferred subbases" that are consistent by construction;
- then, an inference principle manages the preferred subbases using classical entailment.

This process can be synthetized as follows:



Thus, an entailment relation based on the virtual restoration of consistency can be represented by the pair p-m (p for the used inference principle and m for the used selection mechanism).

Historically, this approach was introduced in (Rescher and Manor 1970) using a first version of selection mechanisms and inference principles in the definition of "strong" consequence (a formula is entailed by each subbase that is maximal for set-inclusion).

Then, from the late 1980s and in all the 1990s, many works were realized in order to explore this subject. The aim was either the definition of new selection mechanisms (Poole 1988; Brewka 1989; Cayrol et al. 1993; Dubois et al. 1991; Benferhat et al. 1993a), or the proposal of some inference principles (Pinkas and Loui 1992), or the study and the characterization of the resulting entailment relations following at least two points of view: axiomatisation (Kraus et al. 1990; Pinkas and Loui 1992; Gärdenfors and Makinson 1994; Cayrol and Lagasquie-Schiex 1995; Cayrol et al. 1998) and computational complexity (Nebel 1991; Gottlob 1992; Cayrol et al. 1998).

More recently, in (Martinez et al. 2013), it has been shown that this approach can be considered as a particular case of a more general framework for reasoning from inconsistency.

3.2 Presentation of Some Variants

The reasoning based on consistency restoration needs a selection mechanism and an inference principle, each of them existing in many different variants. We will present the most common variants, knowing that the preference relation $\leq$ associated with the base E is often a total preordering over the formulae of the base expressing a *stratification* of this base.[1]

Selection mechanisms.
The most common method consists in the use of consistent subbases of $(E, \leq)$ maximal for set-inclusion; this is the method introduced in (Rescher and Manor 1970). However, there exist other possibilities for using the relation $\leq$ over E in order to select preferred consistent subbases. Here, we present three of them: the preference relation "Best-Out" issued from possibilistic logic, a preference relation combining the preordering $\leq$ and the maximality for set-inclusion, and a preference relation combining the preordering $\leq$ and the maximality for the cardinality.

Let $(E, \leq)$ be a stratified base. It is usual to represent the stratification of the base by a partition $\{E_1, \ldots, E_n\}$ meaning that any formula of E_i is strictly preferred to any formula of E_j for $i < j$. Let $X = X_1 \cup \ldots \cup X_n$ and $Y = Y_1 \cup \ldots \cup Y_n$ be two consistent subbases of $(E, \leq)$ (with $X_i = X \cap E_i$ and $Y_i = Y \cap E_i$).

[1] Note that there exist other works, not described here, exploiting partial preorderings (Cayrol et al. 1993; Brewka 1994; Benferhat and Garcia 2002; Benferhat and Yahi 2012; Cayrol et al. 2014).

Best-Out preference (Benferhat et al. 1993a): *let X be a consistent subbase of E, consider the notation $a(X) = \min\{i \mid \exists A \in E_i \setminus X\}$. Then Y is* BO*-preferred to X if and only if $a(X) \leq a(Y)$.*[2]

Preference based on inclusion (Cayrol et al. 1993; Geffner 1992):[3] *Y is* INCL*-preferred to X if and only if $\exists i$ such that $X_i \subset Y_i$ and $\forall j,\ 1 \leq j < i, X_j = Y_j$.*

Preference based on cardinality (Lehmann 1995; Benferhat et al. 1993a): *Y is* CARD*-preferred to X if and only if $\exists i$ such that $|X_i| < |Y_i|$ and $\forall j, 1 \leq j < i, |X_j| = |Y_j|$ ($|Z|$ denoting the cardinality of Z).*

Links exist between these different mechanisms (for instance, a CARD-preferred subbase is also INCL-preferred). We will use T (resp. INCL, CARD, BO) to denote the selection mechanism producing the set of the consistent subbases maximal for set-inclusion (resp. INCL-preferred, CARD-preferred, BO-preferred) of $(E, \leq)$.

Inference principles.

There exist at least two well-known principles applicable in presence of conflicting subbases (the skeptical principle and the credulous principle) that are used in (Rescher and Manor 1970). Moreover, a more complete taxomony was established by Pinkas and Loui (1992) with regard to the studied principles. For instance:

Let $(E, \leq)$ be a base equipped with a preordering, and $m(E, \leq)$ denote a set of consistent subbases of $(E, \leq)$ (for instance, $m(E, \leq)$ can be obtained using one of the mechanisms T, INCL, CARD *or* BO*) and let A be a propositional formula, the following inference principles can be defined:*

Uni *principle*: *A is skeptically inferred from $m(E, \leq)$ if and only if A is classically inferred from each element of $m(E, \leq)$.*

Exi *principle*: *A is credulously inferred from $m(E, \leq)$ if and only if A is classically inferred from at least one element of $m(E, \leq)$.*

Arg *principle*:[4] *A is argumentatively inferred from $m(E, \leq)$ if and only if A is classically inferred from at least one element of $m(E, \leq)$ and there is no element of $m(E, \leq)$ that classically infers $\neg A$.*

3.3 An Illustrative Example

Consider the following base $(E, \leq)$ stratified in four levels and representing a variant of the very well known problem of the penguins, the penguin being here replaced by an emu (e means "emu", b means "bird", f means "flying", w means "having wings", a means "having atrophied wings"). The stratification corresponds to a preordering between formulae (reflecting their reliability, their importance, …).

[2] This ordering only depends on the stratum having the highest priority in which at least one formula has been removed for restoring consistency.

[3] The subbases maximal for the preference based on set-inclusion are also called subtheories in (Brewka 1989) and correspond exactly to the strongly maximal consistent subbases of Dubois et al. (1991).

[4] This principle consists in keeping among the credulous consequences only those such that their negation is not credulously inferred (Benferhat et al. 1993b). This inference is said argumentative.

Tweety is an emu	e
an emu is a bird	$e \rightarrow b$
an emu does not fly	$e \rightarrow \neg f$
a bird flies	$b \rightarrow f$
a bird has wings	$b \rightarrow w$
having wings allows one to fly	$w \rightarrow f$
if a bird does not fly then it has atrophied wings	$(b \wedge \neg f) \rightarrow a$

(the formulae e and $e \rightarrow b$ have a higher priority than $e \rightarrow \neg f$ and $b \rightarrow f$ which themselves have a higher priority than $b \rightarrow w$ and $w \rightarrow f$, and the formula $(b \wedge \neg f) \rightarrow a$ has the lowest priority).

We present five subbases among the consistent subbases of this base:[5]

- $Y_1 = \{e, e \rightarrow b, e \rightarrow \neg f, b \rightarrow w, (b \wedge \neg f) \rightarrow a\}$ which entails $e, b, \neg f, w, a$
- $Y_2 = \{e, e \rightarrow b, e \rightarrow \neg f, w \rightarrow f, (b \wedge \neg f) \rightarrow a\}$ which entails $e, b, \neg f, \neg w, a$
- $Y_3 = \{e, e \rightarrow b, b \rightarrow f, b \rightarrow w, w \rightarrow f, (b \wedge \neg f) \rightarrow a\}$ which entails e, b, f, w,
- $Y_4 = \{e, e \rightarrow \neg f, b \rightarrow f, b \rightarrow w, w \rightarrow f, (b \wedge \neg f) \rightarrow a\}$ which entails e, $\neg b$, $\neg f, \neg w$,
- $Y_5 = \{e \rightarrow b, e \rightarrow \neg f, b \rightarrow f, b \rightarrow w, w \rightarrow f, (b \wedge \neg f) \rightarrow a\}$ which entails $\neg e$.

These five subbases are T-preferred, Y_1 to Y_3 are also INCL-preferred and the third one is CARD-preferred.

The literals inferred according to the different principles are:

- with UNI-T, nothing whereas with UNI-INCL, e and b are inferred;
- with EXI-T, $f, \neg f, w, \neg w, e, \neg e, b, \neg b$ and a are inferred, whereas with EXI-INCL, only $f, \neg f, w, \neg w, e, b$ and a are inferred;
- with ARG-T, a is inferred whereas with ARG-INCL, e, b and a are inferred;
- with EXI-CARD, ARG-CARD and UNI-CARD, e, b, f and w are inferred.

Note that EXI-T allows the inference of $f, \neg f, w, \neg w, e, \neg e, b, \neg b$ and a, but not $\neg a$ (as it would be the case with classical deduction). Particularly, f and $\neg f$ cannot be combined in order to generate other conclusions by deduction. Thus EXI-T does not authorize the entailment of all literals and this is the reason that justifies this method as an inconsistency treatment; nevertheless this does not mean that the set of conclusions is "purged" of any source of inconsistency.

[5]Note that the BO-preferred subbases are not presented here. Nevertheless all of them respect a common property: they contain the two formulae e and $e \rightarrow b$ and so they entail e and b.

3.4 Discussion

This combination of selection criteria and inference principles leads to the definition of weak consequence relations that do not respect the monotonicity property (i.e. the addition of new information to the base can call into question a conclusion previously obtained). For instance, if we add in the first stratum of the previous example a new statement "Tweety does not have atrophied wings" (expressed by the formula $\neg a$), we can illustrate the non-monotonicity with the relation ARG-INCL; indeed, the INCL-preferred subbases are modified (the formula $\neg a$ is added at any subbase, whereas the formula $(b \wedge \neg f) \rightarrow a$ is removed from Y_1 and Y_2) and so some conclusions previously obtained disappear: so, with ARG-INCL, $\neg a$ is inferred and not a.

The very important number of such deductive relations has induced many other works in order to analyze these relations in this new framework (Kraus et al. 1990; Gärdenfors and Makinson 1994; Da Silva Neves et al. 2002; Benferhat et al. 2005).

4 Paraconsistent Logics

4.1 Foundations

Except for Jaśkowski (1948), seminal work introducing paraconsistent logics starts with da Costa (1974), Anderson and Belnap (1975), followed by Rescher and Brandom (1979) and others (a recent development over some of these is (Payette 2015). The work by da Costa is motivated by the idea of a formalization of naive set theory, which requires to address the notion of inference from contradictory premises. Anderson and Belnap's motivation is less strongly tied to paraconsistent inference (despite explicit mention of rejecting the principle that a contradiction entails everything) and is instead grounded in the notion of pure implication: for a statement "if E then C" to be true, it must involve a link between the content of E and the content of C.

The intuition underlying paraconsistent logics is that classical inference must be "restricted" in some way (Ripley 2015). In this sense, paraconsistent logics are an approach which is "dual" to the approach described in Sect. 3. It must be stressed that paraconsistent logics *do not assume that there is something wrong* with contradictory premises (Priest 1987).

Formally, premises $\{A, \neg A\}$ are first class citizens over which inference is to apply. What inference can it be? It must take into account the reasons why classical inference collapses in the presence of contradictions: *ex falso quodlibet sequitur*. This is the traditional name for the following inference schema

$$\frac{A \qquad \neg A}{B}$$

which expresses that two contradictory formulae A and $\neg A$ are enough to deduce any formula B.

The *ex falso quodlibet sequitur* is not a primitive principle of classical logic. It is not possible to say "let us take classical logic except for the *ex falso quodlibet sequitur*". Indeed, the *ex falso quodlibet sequitur* is a *derived* principle. Two main ways to derive the *ex falso quodlibet sequitur* are as follows. The first resorts to weakening

$$\frac{A}{B \rightarrow A}$$

as well as non-constructive contraposition

$$\frac{\neg A \rightarrow B}{\neg B \rightarrow A}$$

so that a derivation for the *ex falso quodlibet sequitur* is:

$$\begin{array}{lll} (i) & A & premise \\ (ii) & \neg B \rightarrow A & weakening\ over\ (i) \\ (iii) & \neg A \rightarrow B & non\text{–}constructive\ contraposition\ over\ (ii) \\ (iv) & \neg A & premise \\ (v) & B & modus\ ponens\ over\ (iii)\ and\ (iv) \end{array}$$

A second way to derive the *ex falso quodlibet sequitur* uses disjunctive syllogism

$$\frac{A \vee B \qquad \neg A}{B}$$

as well as the rule of disjunction introduction

$$\frac{A}{A \vee B}$$

so that a derivation of the *ex falso quodlibet sequitur* is:

$$\begin{array}{lll} (i) & A & premise \\ (ii) & A \vee B & disjunction\ introduction\ over\ (i) \\ (iii) & \neg A & premise \\ (iv) & B & disjunctive\ syllogism\ over\ (ii)\ and\ (iii) \end{array}$$

Details differ depending on language, proof theory, … Further proofs are examined in (Øgaard 2016).

4.2 Paraconsistent Inference

When founding a paraconsistent logic, there is no way out but giving up some principle(s) at work in the above derivations so that they break down. For instance, relevant logic admit neither weakening nor disjunctive syllogism. These two points are in full agreement with Anderson and Belnap's idea: for example, as regards weakening, "if E' then E" is untrue when E and E' have nothing to do with each other, even in the event that E would be true. Alternatively, it is also possible to impose an ordering over the application of inference rules (Besnard and Hunter 1995). There are even more striking options, for instance a logic (Tennant 1987) that fails transitivity of inference (i.e., the sequence (i), (ii), (iii), ...).

Let us now digress a little bit so as to mention that paraconsistency is compatible with other options in reasoning (an illustration is (Rahman 2001)).

Back in track, here is, as an example, an axiomatization of the relevant logic R of Anderson and Belnap:

Implication:
$A \rightarrow A$
$(A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$
$(A \rightarrow (B \rightarrow C)) \rightarrow (B \rightarrow (A \rightarrow C))$
$(A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B)$
Negation:
$(A \rightarrow \neg A) \rightarrow \neg A$
$\neg\neg A \rightarrow A$
$(A \rightarrow \neg B) \rightarrow (B \rightarrow \neg A)$
Disjunction:
$A \rightarrow A \vee B$
$A \rightarrow B \vee A$
$(A \rightarrow C) \wedge (B \rightarrow C) \rightarrow ((A \vee B) \rightarrow C)$
Conjunction:
$A \wedge B \rightarrow A$
$B \wedge A \rightarrow A$
$(A \rightarrow B) \wedge (A \rightarrow C) \rightarrow (A \rightarrow (B \wedge C))$
Conjunction and disjunction:
$A \wedge (B \vee C) \rightarrow (A \wedge B) \vee C$
with the inference rules:

$$\frac{A \quad A \rightarrow B}{B} \qquad \frac{A \quad B}{A \wedge B}$$

R is paraconsistent as evidenced by the fact that

$$\frac{A \quad \neg A}{B}$$

is not derivable from the above axiomatics (formally, $\{A, \neg A\} \not\vdash_R B$).

Another insightful hint to found paraconsistent inference is the idea of minimizing inconsistency. Pioneer for such an approach is Priest (1991), who was followed by Arieli and Avron (1996), Besnard and Schaub (1998). The original idea (Priest 1991) is that, as well as the truth-values *true* and *false*, there exists a truth-value *contradictory* and a model of a set of formulae X is an interpretation in which all formulae of X are true or contradictory (intuitively, "at least" true) and such that the set of propositional symbols assigned *contradictory* is minimal.

For such logics, it is crystal-clear that a contradiction does not mean that something is wrong: the inference

$$\{A, \neg A\} \vdash_L A \wedge \neg A$$

makes perfect sense (what is concluded is the fact that there is a contradiction about A) and this is not equivalent to

$$\{A, \neg A\} \vdash_L B.$$

A recent attempt at a unifying approach to paraconsistent logics is (Carnielli and Coniglio 2016).

4.3 An Example: In the Beginning Was the Egg…

Here is a small exercise in formalization about the chicken or the egg dilemma:

e	first was the egg
c	first was the chicken
r	first was the rooster
efc	the egg comes from the chicken
cfe	the chicken comes from the egg

Consider the propositional language based on these five propositional symbols e, c, r, efc, and cfe. As premises, consider (i) the egg comes from the chicken, (ii) the chicken comes from the egg, (iii) if the egg comes from the chicken then it is not the case that first was egg, (iv) if the chicken comes from the egg then it is not the case that first was the chicken, and (v) "first was the egg" and "first was the chicken" are not equivalent. Formally,

$$X = \left\{ \begin{array}{c} efc \\ cfe \\ efc \rightarrow \neg e \\ cfe \rightarrow \neg c \\ \neg(e \leftrightarrow c) \end{array} \right\}$$

According to classical logic, $X \models A$ for each formula A. In contrast, by virtue of a logic such as R,

$$X \vdash_R \neg e$$
$$X \vdash_R \neg c$$
$$\vdots$$

and for R as well as any paraconsistent logic L

$$X \nvdash_L r$$

unlike classical logic that, from these premises, entails that first was the rooster (as a result of the *ex falso quodlibet sequitur*).

5 Argumentation

5.1 Introduction

Argumentation is a cognitive process based upon constructing and evaluating arguments designed with the aim of increasing or decreasing adherence to a viewpoint.

According to Plantin (1996), foundations of a theory of argumentation date back 467 B.C. At the time, Corax and Tisias had already developed a method of argumentation in order to defend landlords in court.

Up to the 1950s, argumentation was studied through rhetorics and logic. In the 1960 and 1970s, philosophers Perelman and Toulmin were the most influential authors on the topic. Perelman described techniques used by people (Perelman and Olbrechts-Tyteca 1958) and Toulmin developed a theory of how argumentation takes place. Since the dawn of the 1990s, argumentation has made its way as a topic in Artificial Intelligence, becoming a major keyword in the domain. Argumentation is indeed regarded as an intuitive paradigm for nonmonotonic reasoning (Dung 1995), reasoning from inconsistent information (Amgoud and Ben-Naim 2015; Besnard and Hunter 2008; Simari and Loui 1992; Aubry and Risch 2005),[6] merging information from multiple sources (Amgoud and Kaci 2007), decision making under uncertainty (Amgoud and Prade 2009; Bonet and Geffner 1996), learning concepts (Mozina et al. 2007), and other applications involving conflicting pieces of information. Argumentation also plays a major role in the analysis and formalization of dialogues. For example, (Fouqueré and Quatrini 2012) presented a linear logic formalization of argumentative dialogues.

[6]There are plenty of other references.

5.2 Architecture of an Argumentation System

5.2.1 Outline

An argumentation process begins with constructing arguments from a knowledge base, continuing with a definition of interactions among these arguments, as well as an evaluation of the intrinsic strength of arguments, and, lastly, an evaluation of arguments so as to determine what arguments are to be used when it comes to drawing conclusions from the knowledge base.

5.2.2 Logical Formalism

An argumentation system is defined from a *logic* $(\mathscr{L}, \mathtt{CN})$ where $\mathscr{L}$ is a set of well-formed *formulae* and CN is a *consequence* operator over $\mathscr{L}$. The language $\mathscr{L}$ is used to represent information whereas CN is essential to the definition of arguments and their interactions. In the literature, there are two main families of logics involved in the definition of argumentation systems: Tarskian logics (Amgoud and Besnard 2009) and rule-based logics (Prakken and Sartor 1997).

After (Tarski 1956), CN is a function from $2^{\mathscr{L}}$ to $2^{\mathscr{L}}$ that satisfies various properties (being a closure operator, …). There is no constraint over the language $\mathscr{L}$. Most of the well-known logics (classical logic, intuitionistic logic, modal logics, …) are Tarskian. The second family of logics takes $\mathscr{L}$ to be a set of literals (i.e., atoms possibly governed by negation $\neg$), a set of rules (possibly split into strict rules and defeasible rules) of the form $l_1, \ldots, l_{n-1} \rightsquigarrow l_n$ (where $\rightsquigarrow$ can either be a strict version $\rightarrow$ or a non-strict version $\Rightarrow$) such that each l_i is a literal. The meaning of such a rule is that if $l_1, \ldots, l_{n-1}$ are true then so is l_n (conditions apply when $\rightsquigarrow$ is of the non-strict kind $\Rightarrow$).

Importantly, logics in either family must admit a notion of consistency. In the second family for example, a set of formulae X of $\mathscr{L}$ is said to be *consistent* for $(\mathscr{L}, \mathtt{CN})$ iff $\mathtt{CN}(X)$ contains no literals l and l' such that l is equivalent to $\neg l'$.

5.2.3 The Notion of an Argument

An argument is a reason to hold a conclusion. It is defined from formulae of a knowledge base $\mathscr{K} \subseteq \mathscr{L}$ using the consequence operator CN. Such an argument is accordingly relative to $\mathscr{K}$.

As to the notion of a formal argument, the following definition is widely used.

Let $\mathscr{K}$ be a knowledge base. An *argument* of $\mathscr{K}$ is a pair (X, x) such that

1. $X \subseteq \mathscr{K}$
2. X is consistent
3. $x \in \mathtt{CN}(X)$

4. $\nexists X' \subset X$ such that X' satisfies all three conditions above.

X is said to be the *support* and x the *conclusion* of the argument.

The following example illustrates this notion of an argument for the case of propositional logic.

Let $\mathscr{K} = \{x, y, x \rightarrow \neg y\}$ be a base in propositional logic. Please observe that the set of all arguments of $\mathscr{K}$ is infinite, some of them are:

$A_1 = (\{x\}, x)$	$A_4 = (\{x, x \rightarrow \neg y\}, \neg y)$
$A_2 = (\{y\}, y)$	$A_5 = (\{y, x \rightarrow \neg y\}, \neg x)$
$A_3 = (\{x \rightarrow \neg y\}, x \rightarrow \neg y)$	$A_6 = (\{x, y\}, x \wedge y)$

5.2.4 Interactions Among Arguments

Arguments constructed from a knowledge base can interact in two ways: by *attacking* or by *supporting*. An attack expresses disagreement or conflict between two arguments. It is a binary relation, meant to capture inconsistency in a knowledge base. It can be defined in a number of ways. In any case, choices made for such a relation are crucial to an argumentation system. Indeed, a poor choice could cause such a system to produce undesirable outcome. Here are some examples of attack relations between two arguments $A_1 = (X_1, x_1)$ and $A_2 = (X_2, x_2)$:

- A_1 *attacks* A_2 iff the set $\{x_1, x_2\}$ is inconsistent, or
- A_1 *attacks* A_2 iff $\exists x \in X_2$ such that the set $\{x_1, x\}$ is inconsistent, or
- A_1 *attacks* A_2 iff $\exists X' \subseteq X_2$ such that the set $\{x_1\} \cup X'$ is inconsistent.

On the other hand, an argument may support another argument. This is represented by means of a binary relation expressing some confluence between two arguments (Cayrol and Lagasquie-Schiex 2013). The fact that an argument supports another argument needs not entail that the latter be accepted by an argumentation system. Here are some examples of support relations between two arguments $A_1 = (X_1, x_1)$ and $A_2 = (X_2, x_2)$:

- A_1 *supports* A_2 iff $x_1 = x_2$, or
- A_1 *supports* A_2 iff $\exists x \in X_2$ such that $x_1 = x$, or
- A_1 *supports* A_2 iff the set $X_1 \cup X_2$ is consistent and $\exists x \in X_2$ such that $x_1 = x$.

5.2.5 Preferences Among Arguments

Both kinds of interactions (attack and support) are about the logical structure of arguments. They do not take into account *quality*, even regarding the formulae occurring in argument supports. Yet, quality could be taken advantage of when it comes to compare arguments. This idea gives rise to another binary relation, dubbed *preference*

relation. In (Bench-Capon 2003), any argument underlies a value (be it economical, moral, …), whose importance determines whether this argument is preferred to what other arguments. More generally, there are different ways to take into account a (pre-)order over $\mathscr{K}$ when comparing arguments. For example, for inconsistency handling in a knowledge base, an argument based on ironclad information is to be preferred to other arguments (Benferhat et al. 1993b).

Let $\mathscr{K} = \mathscr{K}_1 \cup \ldots \cup \mathscr{K}_n$ be a stratified base such that the formulae of $\mathscr{K}_i$ are more certain than the formulae of $\mathscr{K}_j$ for $j > i$. A certainty level can be ascribed to each subbase as follows:

$\texttt{Level}(X) = \min \{i \mid X_{i+1} \cup \ldots \cup X_n = \emptyset\}$ where $X_i = X \cap \mathscr{K}_i$. By convention, $\texttt{Level}(\emptyset) = 0$.

This certainty level is used to compare arguments as follows:

Let $A_1 = (X_1, x_1)$, $A_2 = (X_2, x_2)$ be two arguments constructed from a stratified base $\mathscr{K} = \mathscr{K}_1 \cup \ldots \cup \mathscr{K}_n$. Then, A_1 is preferred to A_2, denoted $A_1 \geq A_2$, iff $\texttt{Level}(X_1) \leq \texttt{Level}(X_2)$.

5.2.6 Evaluation of Arguments

Since arguments can attack and support one another, it seems important to elicit "good" arguments to support a formula to be inferred from an inconsistent knowledge base. An idea is to define *acceptability semantics* for arguments, for which the seminal work is (Dung 1995). It develops an approach to argumentation whose core notion is acceptability of an argument. To start with, an argumentation system is viewed as a set of arguments endowed with an attack relation between these arguments. Structure and origin of the components are undetermined.

An *argumentation system* is a pair $(\mathscr{A}, \mathscr{R})$ where $\mathscr{A}$ is a set of arguments and $\mathscr{R}$ is a binary relation over $\mathscr{A}$. Intuitively, $(A, B) \in \mathscr{R}$ means that A attacks B.

Hence, an argumentation system can be represented as a directed graph whose nodes are the arguments of $\mathscr{A}$ and arcs are the attacks of $\mathscr{R}$. Semantics are defined in order to provide an evaluation of subsets of arguments of such a system. These semantics are supposed to meet at least two requirements: *consistency* and *defense*.

Let $(\mathscr{A}, \mathscr{R})$ be an argumentation system and $\mathscr{B} \subseteq \mathscr{A}$.

- $\mathscr{B}$ is *conflict-free* iff $\nexists A, B \in \mathscr{B}$ such that $(A, B) \in \mathscr{R}$.
- $\mathscr{B}$ *defends* an argument A iff $\forall B \in \mathscr{A}$, if $(B, A) \in \mathscr{R}$, then $\exists C \in \mathscr{B}$ s.t. $(C, B) \in \mathscr{R}$.

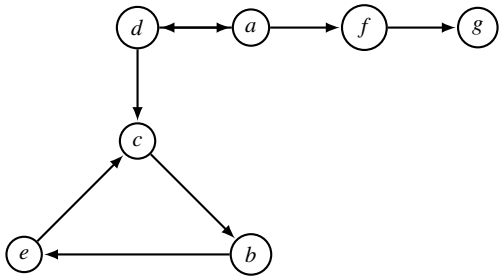
The main semantics proposed by Dung (from which the other semantics can be defined) is based on the principle of admissibility:

Let $\mathscr{B}$ be a conflict-free set of arguments, and let $\mathscr{F} : 2^{\mathscr{A}} \rightarrow 2^{\mathscr{A}}$ be the function defined by $\mathscr{F}(\mathscr{B}) = \{A \in \mathscr{A} \mid \mathscr{B} \text{ defends } A\}$.

- $\mathscr{B}$ is *admissible* iff $\mathscr{B} \subseteq \mathscr{F}(\mathscr{B})$.

- $\mathscr{B}$ is a *complete extension* iff $\mathscr{B} = \mathscr{F}(\mathscr{B})$.
- $\mathscr{B}$ is a *grounded extension* iff $\mathscr{B}$ is a minimal (for set-inclusion) complete extension.
- $\mathscr{B}$ is a *preferred extension* iff $\mathscr{B}$ is a maximal (for set-inclusion) complete extension.
- $\mathscr{B}$ is a *stable extension* iff $\mathscr{B}$ is a preferred extension that attacks (in the sense of $\mathscr{R}$) every argument in $\mathscr{A} \setminus \mathscr{B}$.

Consider the argumentation system represented by the following graph.



This system has only one stable extension $\mathscr{E}_1 = \{b, d, f\}$, it has two preferred extensions $\mathscr{E}_1 = \{b, d, f\}$ and $\mathscr{E}_2 = \{a, g\}$, it has a grounded extension $\mathscr{E}_3 = \emptyset$.

As is proved in (Dung 1995), an argumentation system has only one grounded extension, but it can have several extensions if considering another semantics (as illustrated by the previous example). An argumentation system has always at least one preferred extension but it may happen to have no stable extension.

Once extensions are computed, a qualitative *overall strength* is assigned to each argument as follows: an argument is *skeptically accepted* if it belongs to all extensions, *credulously accepted* if it belongs to some but not all extensions, and *rejected* otherwise.

Many proposals extending the original model (Dung 1995) have been made, for example taking account the support relation (Boella et al. 2010; Cayrol and Lagasquie-Schiex 2013), the relative strength of attacks (Martínez et al. 2008; Dunne et al. 2011), or attacks over attacks (Modgil 2009; Baroni et al. 2011), or audience (Bench-Capon et al. 2007), etc.

Lastly, various instantiations of the original model have been proposed. Some of them capture one or more approaches to non-monotonic reasoning. Dung (1995) presents an instantiation that captures extensions of default logic (Reiter 1980) whereas (Nouioua and Risch 2012) deals with answer set programming (ASP). Cayrol (1995) defines another instantiation, that captures maximal consistent subbases of a knowledge base.

More recently, two other families of semantics are emerging: *gradual* semantics and *ranking* semantics. Gradual semantics assign to each argument a numerical value representing its overall strength. Examples of such semantics are h-Categorizer (Besnard and Hunter 2001; Pu et al. 2014), weighted h-Categorizer, weighted max-based and weighted card-based semantics (Amgoud et al. 2017), game-theoretical semantics (Matt and Toni 2008), social semantics (Leite and Martins 2011), and trust-based semantics (da Costa Pereira et al. 2011). Ranking semantics rank order

arguments from the strongest to the weakest ones. Examples of ranking semantics are tuple-based semantics (Cayrol and Lagasquie-Schiex 2005), Burden-based and Discussion-based semantics (Amgoud and Ben-Naim 2013; Amgoud et al. 2016), and the parametrized one defined in (Bonzon et al. 2017) for persuasion purposes. Obviously, each gradual semantics leads to a ranking one.

It was shown recently in (Amgoud and Ben-Naim 2016; Amgoud et al. 2017) that extension semantics and gradual/ranking semantics are based on different principles. For instance, the number of attackers is taken into account by existing gradual/ranking semantics while it does not play any role in extension semantics. Thus, extension semantics and gradual/ranking semantics may not provide the same evaluations of arguments.

5.2.7 Inference Relations

The last step in an argumentation process consists of defining inference relations permitting to draw plausible conclusions from a knowledge base. This step reuses results from the evaluation of arguments.

Here are a few examples of inference relations in the case of extension-based semantics (Dung 1995):

Let $(\mathscr{A}, \mathscr{R})$ be an argumentation system obtained from a knowledge base $\mathscr{K}$. Let $\mathscr{E}_1, \ldots, \mathscr{E}_n$ be the extensions of this system under a given semantics. Let $x \in \mathscr{L}$.

- $\mathscr{K} \mid\sim x$ iff $\exists A = (X, x) \in \mathscr{A}$ s.t. $A \in \bigcap_i \mathscr{E}_i$, or
- $\mathscr{K} \mid\sim x$ iff $\forall i = 1..n, \exists A = (X, x) \in \mathscr{A}$ s.t. $A \in \mathscr{E}_i$

The plausible conclusions in case of gradual/ranking semantics are simply those supported by at least one argument (Amgoud and Ben-Naim 2015). Note that a formula and its negation may both be plausible. This means that the approach tolerates inconsistency. More importantly, the conclusions are ranked from the most to the least plausible ones. A formula is ranked higher than another formula if it is supported by an argument which is stronger than any argument supporting the second formula.

This is the way argumentation allows us to reach our initial objective of capturing reasoning from inconsistent information.

6 Reasoning in Peer-to-Peer Inference Systems

Peer-to-peer architectures are characterized by the absence of any central control authority or hierarchical organization. Each peer can simultaneously behave as a server and as a client, that can both provide and consume shared resources. This homogeneity contributes to the robustness of the whole, each peer being able to join or leave the network at anytime without compromising the stability of the whole system. Such features appear essential for building flexible and scalable fully distributed

applications over the internet. Well known applications typically share files, computing power or data flows. Peer-to-Peer Inference Systems exploit this paradigm for sharing knowledge and reasoning capabilities.

6.1 Peer-to-Peer Inference Systems

A Peer-to-Peer Inference System (P2PIS) is a finite network of peers $\mathscr{P} = (\mathscr{P}_i)_{i=1..n}$. Each peer $\mathscr{P}_i$ has its own (propositional) language L_i built on a proper alphabet A_i and corresponds to a set of formulae $\mathscr{P}_i = S_i \cup M_i$. The set S_i corresponds to the *proper knowledge* of the peer and is made exclusively from formulae constructed on L_i. The set M_i describes semantic links, called *mappings*, established by $\mathscr{P}_i$ to relate some of its own concepts with those of other peers. M_i is made of formulae of the language L built on the alphabet $A = \bigcup_{i=1..n} A_i$, that contain at least one term of L_i and one term of another language L_j $(j \neq i)$. In the following we assume without loss of generality that such theories are expressed in clausal form.

An important characteristic of P2PIS is that each peer only has a local view of the system in which it belongs. Actually it is only aware of its own knowledge and of the mappings that connect it to its direct neighbours in the network of peers. But each of them knows neither the global theory $\Sigma = \bigcup_{i=1..n} \mathscr{P}_i$ nor the topology of the network of peers, on which no particular assumption can be made. The challenge is thus to propose fully decentralized reasoning algorithms, making it possible for the peers to collaborate in an appropriate way during inference tasks over the global theory, despite the fact that each of them only has a local view of the whole system.

Works by Adjiman et al. (2004, 2005, 2006) have proposed an incremental message-passing algorithm (DECA), that can produce all proper prime implicates of a clause with respect to the global theory. However this algorithm assumes the global theory Σ to be consistent. Yet in a peer-to-peer inference system where each peer is independent and can freely design its local theory and its mappings, this cannot be ensured. In this context, one may wonder whether it is possible to detect inconsistencies in the global theory in a decentralized way and, in such cases, whether it is possible to avoid deriving trivial conclusions.

6.2 Inconsistency in Peer-to-Peer Inference Systems

Among the two classical alternatives : repair or tolerate inconsistencies, the first one must clearly be ruled out. In a peer-to-peer system, each peer being independent, it can only control its own theory and cannot force other peers involved in inconsistencies to repair themselves. Moreover, given the homogeneity of the peers, each peer is as legitimate as others and it is difficult to hold one of them as particularly responsible for the cause of an inconsistency. In many cases, the responsibility is rather collective. Therefore the only realistic approach seems to be considering methods able to tolerate inconsistencies. Particularly, one would like to restrict conclusions that can be derived

to the *well-founded* ones, i.e. those that can be derived from a consistent subset of Σ.

While consistency of the global theory cannot be ensured, assuming the consistency of the local theory $S_i \cup M_i$ of each peer seems reasonable. This can be checked easily with a local satisfiability test. Since the languages L_i used for expressing the proper knowledge of the respective peers $\mathscr{P}_i$ are disjoint, $S = \bigcup_{i=1..n} S_i$ is thus necessarily consistent. In some way, inconsistencies can be considered caused by mappings of $M = \bigcup_{i=1..n} M_i$. This seems intuitively acceptable since, while each peer can be held as qualified for stating knowledge using its own language, when establishing mappings with other peers, a peer does not necessarily have a good perception of the semantics of the concepts introduced by its neighbours.

This approach is followed in (Chatalic et al. 2006), where causes of incoherence, called *nogoods*, are defined as sets ng of mappings from M, such that $S \cup ng \models \bot$. For any minimal nogood ng and any mapping $m \in ng$, $\bot$ is necessarily a proper prime implicate of m relatively to $S \cup ng \setminus \{m\}$. This peculiarity is exploited by the P2P- NG algorithm, that is able to detect all nogoods of Σ. This algorithm can be seen as a specialization of DECA that can produce all minimal sets of mappings (called *mapping supports*) used in a derivation of $\bot$, by resolution from an initial input clause. This algorithm runs in the same way on each peer of the network and proceeds in two steps. From a given clause (initially, a new mapping m that a peer $\mathscr{P}_k$ wants to add), it first produces all implicates that only contain literals of other peers' languages, while keeping track of the (local) mappings used in each proof. Each obtained clause $c = l^1_{j_1} \vee \ldots \vee l^n_{j_n}$ is split and for each literal $l^i_{j_i}$ in the language L_{j_i} of a neighbour peer $\mathscr{P}_{j_i}$, P2P- NG is launched again on $\mathscr{P}_{j_i}$, with the input clause $l^i_{j_i}$. The results of the recursive calls on neighbour peers are then recombined on the queried peer in an incremental way, using a distribution operator. If the final result is non empty, each obtained set of mappings constitutes with the mapping m a nogood, that will be stored on $\mathscr{P}_k$. A history mechanism, included in the transmitted messages, prevents possible problems induced by cycles in the graph of peers, that cannot be excluded, and guarantees the completion of the process.

6.3 Illustrative Example

The behaviour of P2P- NG is illustrated on Fig. 1, assuming that the mappings of the different peers are added successively, according to the order m_3, m_2, m_1, m_4. From m_3, the peer $\mathscr{P}_3$ locally produces b_1 whereas $\mathscr{P}_1$ cannot produce $\bot$ from b_1.

From m_2, the peer $\mathscr{P}_2$ locally produces a_1. It then queries the peer $\mathscr{P}_1$, that produces $\neg b_1$ locally, without using any mapping. $\mathscr{P}_1$ in turn queries the peer $\mathscr{P}_3$, that manages to produce $\bot$ using $\{m_3\}$. Finally, the set $\{m_3\}$ is sent back to $\mathscr{P}_1$, and then to $\mathscr{P}_2$. Hence, the latter has detected a nogood $\{m_2, m_3\}$, which is stored on the peer $\mathscr{P}_2$. When adding m_1 and m_4 to $\mathscr{P}_4$, no further inconsistencies are detected.

Note that the different nogoods are stored in a completely decentralized way. Moreover, among all peers involved in a nogood ng, the only peer that is aware of

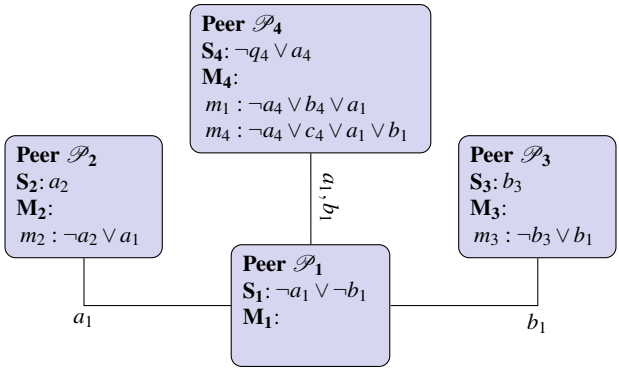


Fig. 1 An inconsistent P2PIS

ng is the one on which it is stored. The completeness of P2P- NG guarantees that all nogoods are detected and they are stored *somewhere* on the network.

The WF- DECA algorithm (Chatalic et al. 2006), can compute *well-founded* proper implicates of a clause with respect to Σ. As P2P- NG, it computes the minimal mapping supports of the produced implicates. Simultaneously, while other peers of the network taking part in the current reasoning are visited, all the nogoods stored on these peers, that could invalidate mapping supports of produced implicates, are collected and sent back together with the implicates. Mapping supports containing such nogoods are discarded, as well as implicates with no remaining mapping support.

In the previous example, if $\mathscr{P}_4$ is asked to compute proper prime implicates of q_4, the local consequents that are obtained are $\{q_4, a_4, b_4 \vee a_1, c_4 \vee a_1 \vee b_1\}$, with respective sets of mapping supports $\{\emptyset\}$, $\{\emptyset\}$, $\{\{m_1\}\}$ and $\{\{m_4\}\}$, and no nogood is collected. On $\mathscr{P}_1$, $\perp$ is obtained as an implicate of a_1, with the set of mapping supports $\{\{m_3\}\}$, and no nogood is collected. On $\mathscr{P}_2$, $\perp$ is obtained as an implicate of b_1, with the set of mapping supports $\{\{m_2\}\}$ and the nogood $\{\{m_2, m_3\}\}$ is collected. Eventually, implicates obtained on $\mathscr{P}_4$ are b_4, supported by $\{\{m_1, m_3\}\}$, and c_4, supported by $\{\{m1, m_2, m_3\}\}$. But since the latter contains the nogood $\{m_2, m_3\}$ (that has been retrieved when visiting $\mathscr{P}_2$) it is not well-founded and is thus discarded.

Note that in this approach, both p and $\neg p$ can be derived as well-founded consequents, from different consistent subsets of Σ. To deal with such cases, (Binas and McIlraith 2008) has proposed an extension of this work, based on argumentation techniques. Their approach comes down to generalizing the notion of support to minimal subsets of *all* formulae from which an implicate can be derived (not only mappings). They assume the existence of a total order on the set of peers and introduce priority degrees, that are used to prefer some implicates over others, according to the priority of arguments supporting them. However, assuming such a *global* total order on the whole set of peers looks unrealistic in the context of peer-to-peer and it goes against the principles of a completely decentralized approach on possibly dynamic networks.

7 Conclusion

Artificial Intelligence, by considering reasoning models, has nurtured the development of various logical approaches to reasoning from contradictory information.

This chapter sketches various approaches distinctive of the principle "make the best of" with such information, without having to temper with it in some way or another. It is all about tolerating inconsistency instead of attempting (in a presumably vain endeavour) to avoid it.

Beside the problem of reasoning from contradictory information, there exist other works on inconsistency measures whose objective is to evaluate to what extent a given body of information is inconsistent (Hunter 2002; Dubois et al. 2003; Hunter and Konieczny 2010; Grant and Martinez 2018).

Since early nineties, most works on argumentation focused either on defining particular semantics for the evaluation of arguments, or on showing how argumentation can be used for solving different problems. Consequently, there is a great number of semantics without formal tools for evaluating and comparing them. For bridging this gap, Amgoud and Ben-Naim (2013) proposed properties for comparing ranking semantics. Amgoud et al. (2017), Bonzon et al. (2017), Amgoud and Ben-Naim (2018), Baroni et al. (2018) discussed other properties for different kinds of argumentation frameworks (attack graphs, support graphs, bipolar graphs, weighted graphs). These properties allow a better understanding of each semantics and a clear comparison of pairs of semantics. They could also provide a classification of semantics in terms of their suitability to particular applications.

References

Adjiman P, Chatalic P, Goasdoué F, Rousset MC, Simon L (2004) Distributed reasoning in a peer-to-peer setting, short paper. In: 16th European conference on artificial intelligence (ECAI'04, (ed) Lopez de Mántaras R, Saitta L. IOS, Valencia, Spain, pp 945–946

Adjiman P, Chatalic P, Goasdoué F, Rousset MC, Simon L (2005) Scalability study of peer-to-peer consequence finding. In: Pack Kaelbling L, Saffiotti A (eds) 19th international joint conference on artificial intelligence (IJCAI'05). Professional Book Center, Edinburgh, Scotland, U.K., pp 351–356

Adjiman P, Chatalic P, Goasdoué F, Rousset MC, Simon L (2006) Distributed reasoning in a peer-to-peer setting: application to the semantic web. J Artif Intell Res 25:269–314

Amgoud L (2012a) The outcomes of logic-based argumentation systems under preferred semantics. In: Hüllermeier E, Link S, Seeger B (eds) 6th international conference on scalable uncertainty management (SUM'12), vol 7520. Lecture notes in artificial intelligence. Springer, Germany, pp 72–84

Amgoud L (2012b) Stable semantics in logic-based argumentation systems. In: Hüllermeier E, Link S, Seeger B (eds) 6th international conference on scalable uncertainty management (SUM'12), vol 7520. Lecture notes in artificial intelligence. Springer, Germany, pp 58–71

Amgoud L, Ben-Naim J (2013) Ranking-based semantics for argumentation frameworks. In: Liu W, Subrahmanian V, Wijsen J (eds) 7th international conference on scalable uncertainty management (SUM'13), vol 8078. Lecture notes in artificial intelligence. Springer, USA, pp 134–147

Amgoud L, Ben-Naim J (2015) Argumentation-based ranking logics. In: Weiss G, Yolum P, Bordini R, Elkind E (eds) 14th international conference on autonomous agents and multiagent systems (AAMAS'15). ACM, Istanbul, Turkey, pp 1511–1519

Amgoud L, Ben-Naim J (2016) Axiomatic foundations of acceptability semantics. In: Baral C, Delgrande J, Wolter F (eds) 15th international conference on principles of knowledge representation and reasoning (KR'16). AAAI, Cape Town, South Africa, pp 2–11

Amgoud L, Ben-Naim J (2018) Weighted bipolar argumentation graphs: Axioms and semantics. In: 27th international joint conference on artificial intelligence (IJCAI'18), Stockholm, Sweden

Amgoud L, Besnard P (2009) Bridging the gap between abstract argumentation systems and logic. In: Godo L, Pugliese A (eds) 3rd international conference on scalable uncertainty management, vol 5785. Lecture notes in artificial intelligence. Springer, USA, pp 12–27

Amgoud L, Kaci S (2007) An argumentation framework for merging conflicting knowledge bases. Int J Approx Reason 45(2):321–340

Amgoud L, Prade H (2009) Using arguments for making and explaining decisions. Artif Intell J 173(3–4):413–436

Amgoud L, Ben-Naim J, Doder D, Vesic S (2016) Ranking arguments with compensation-based semantics. In: Baral C, Delgrande J, Wolter F (eds) 15th international conference on principles of knowledge representation and reasoning (KR'16). AAAI, South Africa, pp 12–21

Amgoud L, Ben-Naim J, Doder D, Vesic S (2017) Acceptability semantics for weighted argumentation frameworks. In: Sierra C (ed) 26th international joint conference on artificial intelligence (IJCAI'17). Melbourne, Australia, pp 56–62

Anderson A, Belnap N (1975) Entailment: the logic of relevance and necessity, vol 1. Princeton University, Princeton

Arieli O, Avron A (1996) Reasoning with logical bilattices. J Log, Lang Inf 5(1):25–63

Aubry G, Risch V (2005) Toward a logical tool for generating new arguments in an argumentation-based framework. In: 17th. IEEE international conference on tools with artificial intelligence (ICTAI'05). IEEE Computer Society, China, pp 599–603

Baroni P, Cerutti F, Giacomin M, Guida G (2011) AFRA: Argumentation framework with recursive attacks. Int J Approx Reason 52(1):19–37

Baroni P, Rago A, Toni F (2018) How many properties do we need for gradual argumentation? In: 32nd AAAI conference on artificial intelligence (AAAI'18), New Orleans, USA

Bench-Capon TJM (2003) Persuasion in practical argument using value-based argumentation frameworks. J Log Comput 13(3):429–448

Bench-Capon TJM, Doutre S, Dunne PE (2007) Audiences in argumentation frameworks. Artif Intell 171(1):42–71

Benferhat S, Garcia L (2002) Handling locally stratified inconsistent knowledge bases. Studia Logica 70(1):77–104

Benferhat S, Yahi S (2012) Étude comparative des relations d'inférence à partir de bases de croyances partiellement préordonnées. Revue d'Intelligence Artificielle 26(1–2):39–61

Benferhat S, Cayrol C, Dubois D, Lang J, Prade H (1993a) Inconsistency management and prioritized syntax-based entailment. In: Bajcsy R (ed) 13th international joint conference on artificial intelligence (IJCAI'93). Morgan Kaufmann, France, pp 640–645

Benferhat S, Dubois D, Prade H (1993b) Argumentative inference in uncertain and inconsistent knowledge bases. In: Heckerman D, Mamdani A (eds) 9th conference on uncertainty in artificial intelligence. Morgan Kaufmann, USA, pp 411–419

Benferhat S, Bonnefon JF, Da Silva Neves R (2005) An overview of possibilistic handling of default reasoning: an experimental study. Synthese 146(1–2):53–70

Besnard P, Hunter A (1995) Quasi-classical logic: Non-trivializable classical reasoning from inconsistent information. In: Froidevaux C, Kohlas J (eds) 3rd European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU'95), Springer, Fribourg, Switzerland, Lecture Notes in Artificial Intelligence, vol 946, pp 44–51

Besnard P, Hunter A (2001) A logic-based theory of deductive arguments. Artif Intell 128(1–2):203–235

Besnard P, Hunter A (2008) Elements of argumentation. MIT, USA
Besnard P, Schaub T (1998) Signed systems for paraconsistent reasoning. J Autom Reason 20(1):191–213
Binas A, McIlraith S (2008) Peer-to-peer query answering with inconsistent knowledge. In: Brewka G, Lang J (eds) 11th international conference on principles of knowledge representation and reasoning, Morgan Kaufmann, Australia, pp 329–339, http://www.cs.toronto.edu/~sheila/publications/bin-mci-kr08.pdf
Boella G, Gabbay D, van der Torre L, Villata S (2010) Support in abstract argumentation. In: Baroni P, Cerutti F, Giacomin M, Simari G (eds) Computational models of argument (COMMA'10), Frontiers in artificial intelligence and applications, vol 216. IOS. Desenzano del Garda, Italy, pp 111–122
Bonet B, Geffner H (1996) Arguing for decisions: A qualitative model of decision making. In: Horvitz E, Jensen FV (eds) 12th conference on uncertainty in artificial intelligence (UAI'96). Morgan Kaufmann, USA, pp 98–105
Bonzon E, Delobelle J, Konieczny S, Maudet N (2017) A parametrized ranking-based semantics for persuasion. In: Moral S, Pivert O, Sánchez D, Marín N (eds) 11th international conference on scalable uncertainty management (SUM'17), vol 10564. Lecture notes in computer science. Granada, Spain, pp 237–251
Brewka G (1989) Preferred subtheories: An extended logical framework for default reasoning. In: Sridharan NS (ed) 11th International Joint Conference on Artificial Intelligence (IJCAI'89). Morgan Kaufmann, Detroit (MI), USA, pp 1043–1048
Brewka G (1994) Reasoning about priorities in default logic. In: Hayes-Roth B, Korf RE (eds) 12th national conference on artificial intelligence (AAAI'94). AAAI/MIT, USA, pp 940–945
Carnielli W, Coniglio ME (2016) Paraconsistent logic: consistency, contradiction and negation, logic, epistemology, and the unity of science, vol 40. Springer, Berlin
Cayrol C (1995) On the relation between argumentation and non-monotonic coherence-based entailment. In: Mellish CS (ed) 14th international joint conference on artificial intelligence (IJCAI'95). Morgan Kaufmann, Canada, pp 1443–1448
Cayrol C, Lagasquie-Schiex MC (1995) Non-monotonic syntax-based entailment: a classification of consequence relations. In: Froidevaux C, Kohlas J (eds) 3rd European conference on symbolic and quantitative approaches to reasoning and uncertainty (ECSQARU'95), vol 946. Lecture notes in artificial intelligence. Springer, Switzerland, pp 107–114
Cayrol C, Lagasquie-Schiex MC (2005) Graduality in argumentation. J Artif Intell Res 23:245–297
Cayrol C, Lagasquie-Schiex MC (2013) Bipolarity in argumentation graphs: towards a better understanding. Int J Approx Reason 5(7):876–899. https://doi.org/10.1016/j.ijar.2013.03.001
Cayrol C, Royer V, Saurel C (1993) Management of preferences in assumption-based reasoning. In: Yager R, Bouchon B (eds) Advanced Methods in Artificial Intelligence, Lecture Notes in Artificial Intelligence, vol 682, Springer, pp 13–22, extended version in Technical Report IRIT-CERT, 92-13R (University Paul Sabatier Toulouse)
Cayrol C, Lagasquie-Schiex MC, Schiex T (1998) Nonmonotonic reasoning: from complexity to algorithms. Ann Math Artif Intell 22(3–4):207–236
Cayrol C, Dubois D, Touazi F (2014) On the semantics of partially ordered bases. In: Beierle C, Meghini C (eds) 8th international symposium on foundations of information and knowledge systems (FoIKS'14), vol 8367. Lecture notes in artificial intelligence. Springer, France, pp 136–153
Chatalic P, Nguyen GH, Rousset MC (2006) Reasoning with inconsistencies in propositional peer-to-peer inference systems. In: Brewka G, Coradeschi S, Perini A, Traverso P (eds) 17th European conference on artificial intelligence (ECAI'06). IOS, Italy, pp 352–357
da Costa NCA (1974) On the theory of inconsistent formal systems. Notre Dame J Form Log 15(4):497–510
da Costa Pereira C, Tettamanzi A, Villata S (2011) Changing one's mind: Erase or rewind? In: Walsh T (ed) 22nd international joint conference on artificial intelligence (IJCAI'11). IJCAI/AAAI, Barcelona, Spain, pp 164–171

Da Silva Neves R, Bonnefon JF, Raufaste E (2002) An empirical test of patterns for nonnomotonic reasoning. Ann Math Artif Intell 34(1–3):107–130
Dubois D, Lang J, Prade H (1991) Inconsistency in possibilistic knowledge bases - to live or not to live with it. In: Zadeh LA, Kacprzyk J (eds) Fuzzy logic for the management of uncertainty. Wiley, New York, pp 335–351
Dubois D, Konieczny S, Prade H (2003) Quasi-possibilistic logic and its measures of information and conflict. Fundamenta Informaticæ 57(2–4):101–125
Dung PM (1995) On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. Artif Intell J 77(2):321–357
Dunne P, Hunter A, McBurney P, Parsons S, Wooldridge M (2011) Weighted argument systems: Basic definitions, algorithms, and complexity results. Artif Intell J 175(2):457–486
Elvang-Gøransson M, Hunter A (1995) Argumentative logics: reasoning from classically inconsistent information. Data Knowl Eng 16(2):125–145
Fouqueré C, Quatrini M (2012) Un cadre formel issu de la théorie de la démonstration pour la théorie de l'argumentation. Math Soc Sci 2(198):49–83
Gärdenfors P, Makinson D (1994) Nonmonotonic inference based on expectations. Artif Intell J 65:197–245
Geffner H (1992) Default reasoning: causal and conditional theories. MIT, USA
Gottlob G (1992) Complexity results for nonmonotonic logics. J Log Comput 2(3):397–425
Grant J, Martinez MV (eds) (2018) Measuring inconsistency in information, studies in logic, vol 73. College Publications
Hunter A (2002) Measuring inconsistency in knowledge via quasi-classical models. In: Dechter R, Sutton R (eds) 18th american national conference on artificial intelligence (AAAI'2002). AAAI/MIT, Canada, pp 68–73
Hunter A, Konieczny S (2010) On the measure of conflicts: shapley inconsistency values. Artif Intell J 174(14):1007–1026
Jaśkowski S (1948) Rachunek zdań dla systemów dedukcyjnych sprzecznych. Studia Societatis Scientiarum Torunensis, Sectio A I:55–77, translated as "Propositional Calculus for Contradictory Deductive System" in Studia Logica 24 (1969), pp. 143–157 and also in Logic and Logical Philosophy 7 (1999), pp. 35–56
Konieczny S, Lang J, Marquis P (2003) Quantifying information and contradiction in propositional logic through test actions. In: Gottlob G, Walsh T (eds) 18th international joint conference on artificial intelligence (IJCAI'03). Morgan Kaufmann, Mexico, pp 106–111
Kraus S, Lehmann D, Magidor M (1990) Nonmonotonic reasoning, preferential models and cumulative logics. Artif Intell J 44(1–2):167–207
Lang J, Marquis P (2000) In search of the right extension. In: Cohn AG, Giunchiglia F, Selman B (eds) 7th international conference on principles of knowledge representation and reasoning (KR'00). Morgan Kaufmann, USA, pp 625–636
Lang J, Marquis P (2010) Reasoning under inconsistency: a forgetting-based approach. Artif Intell 174(12–13):799–823
Lehmann D (1995) Another perspective on default reasoning. Ann Math Artif Intell 15(1):61–82
Leite J, Martins J (2011) Social abstract argumentation. In: Walsh T (ed) 22nd international joint conference on artificial intelligence (IJCAI'11). IJCAI/AAAI, Spain, pp 2287–2292
Martínez DC, García A, Simari G (2008) An abstract argumentation framework with varied-strength attacks. In: Brewka G, Lang J (eds) 11th international conference on principles of knowledge representation and reasoning (KR'08). AAAI, Australia, pp 135–144
Martinez MV, Molinaro C, Subrahmanian VS, Amgoud L (2013) A general framework for reasoning on inconsistency. Springer Briefs in Computer Science, Springer
Matt P, Toni F (2008) A game-theoretic measure of argument strength for abstract argumentation. In: Hölldobler S, Lutz C, Wansing H (eds) 11th European conference on logics in artificial intelligence (JELIA'08). Springer, Germany, pp 285–297
Modgil S (2009) Reasoning about preferences in argumentation frameworks. Artif Intell J 173(9–10):901–1040

Mozina M, Zabkar J, Bratko I (2007) Argument-based machine learning. Artif Intell J 171(10–15):922–937
Nebel B (1991) Belief revision and default reasoning: Syntax-based approaches. In: Allen JA, Fikes R, Sandewall E (eds) 2nd international conference on principles of knowledge representation and reasoning (KR'91). Morgan Kaufmann, USA, pp 417–428
Nouioua F, Risch V (2012) A reconstruction of abstract argumentation admissible semantics into defaults and answer sets programming. In: Filipe J, Fred ALN (eds) 4th international conference on agents and artificial intelligence (ICAART'12), vol 1. SciTe, Portugal, pp 237–242
Øgaard TF (2016) Paths to triviality. Philos Log 45(3):237–276
Payette G (2015) Getting the most out of inconsistency. Philos Log 44(5):573–592
Perelman C, Olbrechts-Tyteca L (1958) Traité de l'argumentation : La nouvelle rhétorique. Éditions de l'Université Libre de Bruxelles, adapted and translated in 1969 as: "The New Rhetoric: A Treatise on Argumentation". Notre Dame University Press
Pinkas G, Loui RP (1992) Reasoning from inconsistency: A taxonomy of principles for resolving conflict. In: Nebel B, Rich C, Swartout WR (eds) 3rd international conference on principles of knowledge representation and reasoning (KR'92). Morgan Kaufmann, USA, pp 709–719
Plantin C (1996) L'argumentation. Mémos Seuil, Seuil
Poole D (1988) A logical framework for default reasoning. Artif Intell 36(1):27–47
Prakken H, Sartor G (1997) Argument-based extended logic programming with defeasible priorities. J Appl Non-Class Log 7(1):25–75
Priest G (1987) In Contradiction. Martinus Nijhoff, The Hague, The Netherlands
Priest G (1991) Minimally inconsistent LP. Studia Logica 50:321–331
Pu F, Luo J, Zhang Y, Luo G (2014) Argument ranking with categoriser function. In: Buchmann R, Kifor C, Yu J (eds) 7th international knowledge science, engineering and management conference (KSEM'14). Springer, Romania, pp 290–301
Rahman S (2001) On Frege's nightmare. a combination of intuitionistic, free and paraconsistent logics. In: Wansing H (ed) Essays on non-classical logic. World Scientific, pp 61–85
Reiter R (1980) A logic for default reasoning. Artif Intell J 13(1–2):81–132
Rescher N, Brandom R (1979) The logic of inconsistency. Blackwell
Rescher N, Manor R (1970) On inference from inconsistent premises. Theory Decis 1(2):179–217
Ripley D (2015) Paraconsistent logic. Philos Log 44(6):771–780
Simari G, Loui RP (1992) A mathematical treatment of defeasible reasoning and its implementation. Artif Intell J 53(2–3):125–157
Tarski A (1956) Logic, semantics, metamathematics. Woodger EH (ed) Chap On some fundamental concepts of metamathematics. Oxford University, Oxford
Tennant N (1987) Natural deduction and sequent calculus for intuitionistic relevant logic. J Symb Log 52(3):665–680

Main Issues in Belief Revision, Belief Merging and Information Fusion

Didier Dubois, Patricia Everaere, Sébastien Konieczny and Odile Papini

Abstract This chapter focuses on the dynamics of information represented in logical or numerical formats, from pioneering works to recent developments. The logical approach to belief change is a topic that has been extensively studied in Artificial Intelligence, starting in the mid-seventies. In this problem, logical formulas represent beliefs held by an intelligent agent that must be revised upon receiving new information that conflicts with prior beliefs and usually has priority over them. In contrast, in the merging problem, the logical theories that must be combined have equal priority. Such logical approaches recalled here make sense for merging beliefs as well as goals, even if each of these problems cannot be reduced to the other. In the last part, we discuss a number of issues pertaining to the fusion and the revision of uncertainty functions representing epistemic states, such as probability measures, possibility measures and belief functions. The need to cope with logical inconsistency plays a major role in these problems. The ambition of this chapter is not to provide an exhaustive bibliography, but rather to propose an overview of basic notions, main results and new research issues in this area.

D. Dubois (✉)
IRIT, CNRS and Université Paul Sabatier, Toulouse, France
e-mail: dubois@irit.fr

P. Everaere
CRIStAL-CNRS, Université de Lille, Lille, France
e-mail: patricia.everaere-caillier@univ-lille.fr

S. Konieczny
CRIL-CNRS, Université d'Artois, Lens, France
e-mail: konieczny@cril.univ-artois.fr

O. Papini
Aix Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
e-mail: odile.papini@univ-amu.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_14

1 Introduction

The dynamics of beliefs, or yet belief change, is an important research topic in Artificial Intelligence. In many practical artificial intelligence problems, we face the situation where an agent entertains assumptions about the world and receives a new reliable piece of information that contradicts prior beliefs of this agent or information previously received. Such information can be imperfect in the sense that it is possibly incomplete, imprecise or uncertain. In this approach, it is assumed that assumptions about the world, called "beliefs", are constructed by an intelligent agent from information received (observations, testimonies) and also from background knowledge or experience. When a new piece of information comes in, and contradicts the current epistemic state of the agent, revision consists in restoring consistency, so as to integrate the new piece of information, while minimally modifying initial beliefs. Besides, if the new input comes from several sources, the corresponding pieces of information can contradict each other, in which case information fusion aims to extract reliable facts, by exploiting complementarity between sources, and solving conflicts so as to reduce imprecision and uncertainty.

First works in belief revision can be found in the literature on subjective probabilities, mainly the works of Richard Jeffrey in the 1960's. In this framework (Jeffrey 1983), the agent's beliefs are represented by a measure of probability and belief revision is couched in terms of what Jeffrey calls "probability kinematics". Some time later, logical approaches to revision have been developed in the area of epistemology, under the name "theory change", the starting point being the study of how scientific theories evolve. First results on logical change operations date back to years 1975–1977 in the field of epistemology (Levi 1980) and the history of sciences (Harper 1975).

Besides, in the area of databases, the issue of updating has been the focus point of several works, in particular (Fagin et al. 1983), where a methodology for updating logical deductive databases based on model theory was proposed. This approach does not rely on a set of formulas, rather on a set of logical interpretations. In the mid-1980's, connections between results obtained in epistemology, artificial intelligence and databases have been laid bare, and this convergence process proved fruitful.

First attempts at formalizing revision in artificial intelligence come from philosophical logics with the works of Carlos Alchourrón, Peter Gärdenfors and David Makinson (Alchourrón et al. 1985; Gärdenfors 1988) in the 1980s. They proposed postulates, now known as AGM postulates, that characterize revision operations. Later on, they proposed first concrete revision operations for theories, i.e., deductively closed sets of logical formulas.[1]

The originality of the AGM approach lies in the abstract standpoint chosen for studying the revision problem. This specificity can be highlighted by the following aspects:

[1] Logical formulas and their logical consequences.

- The study of postulates that any reasonable revision operation should satisfy (instead of focusing on a particular revision operation as in previous works);
- A consistency-based approach, so-called *coherentist*, that insists on the importance of maintaining the consistency of the set of beliefs.[2] This approach does not depend on the nature of beliefs (sources, justification, distinction between explicit and derived beliefs $\cdots$); it relies on a very simple representation framework for pieces of information (beliefs) that play the same role (hence the use of logical theories).

This view is in opposition with the so-called *foundational* approaches, which insist on the different roles played by available pieces of information (e.g. the original ones vs. the derived ones). The latter historically refer to techniques from truth-maintenance systems (Doyle 1979; de Kleer 1986), that handle justifications for each agent belief.

Coherentist approaches have the merit to focus on the belief change process per se, so as to reach a higher level of generality. However, they can be criticized, because the idea of assigning different statuses to information items (justifications, level of reliability, or genericity, etc.) is something natural that is not captured by the AGM framework. The coherentist approach must be seen as a first step enabling general principles for revision to be laid bare. Then the notions thus established can be adapted to the foundational setting where pieces of information do not have the same status.

Let us mention as other examples of approaches that are linked to the foundational point of view, the revision of belief bases that are not deductively closed, after Hansson (1998, 1993, 1999), where a distinction is made between explicit beliefs in the base and implicit beliefs that are deduced from it via inference. Nethertheless these approaches take into account the "coherentist" framework.

Another distinction is made between beliefs and knowledge, the latter being understood as generic information (Dubois 2008). Beliefs pertain to the current state of the world, and evolve as new observations of the same type are acquired. In contrast with beliefs, knowledge, due to its genericity, is more stable and seldom questioned by the arrival of new observations on the current state of the world. However the belief revision process relies on the agent's available knowledge to construct new beliefs in agreement with the new observations. Belief revision is thus clearly related to non-monotonic reasoning and non-monotonic logics as detailed in chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" of this volume.

Regarding information fusion, many publications have appeared since the 1970's in the setting of probability theory (see Genest and Zidek 1986; Cooke 1991 for surveys) and more general uncertainty theories (e.g., the combination rule of Dempster in evidence theory (Shafer 1976) and the conjunction and disjunction connectives in possibility theory (Dubois and Prade 1988)). Besides, the problem of coherent merging of heterogeneous logical databases has produced many works in the area of databases since the 1980's. It is only from the mid 1990's that logical approaches to fusion raised interest in the AI community (Baral et al. 1991; Revesz 1993, 1997;

[2] However, methods for inconsistency management are surveyed at a more general level in chapter "Argumentation and Inconsistency-Tolerant Reasoning" of this volume.

Lin and Mendelzon 1998; Cholvy and Hunter 1997). A set of postulates characterizing the rational behavior of merging operations was proposed by Konieczny and Pino Pérez (2002a), based on the works by Revesz (1993, 1997), quite in the same spirit as the AGM postulates of belief revision. The multisource merging problem has been the focus of numerous studies in the recent past years.

Be it about revision or fusion, there are many concrete methods available, because there does not seem to exist a universal revision or fusion operation that would be satisfactory in all circumstances. The choice of the method depends on the epistemic status of pieces of information to be handled, and also on the application context.

This chapter is composed of six sections. The next one is dedicated to logical approaches to belief revision and Sect. 3 is devoted to iterated belief revision. Then Sect. 4 gives the state of art about logical approaches to theory merging. Section 5 provides a quick survey of numerical approaches to revision and fusion before concluding in Sect. 6. The ambition of this chapter is not to provide an exhaustive bibliography. It rather proposes an overview of the main results and opens on new research issues in revision and merging. However this chapter does not consider problems of updating and reasoning about action, which are the topic of chapter "Reasoning About Action and Change" of this volume.

2 Belief Revision

In this section, an agent's beliefs are represented by logical formulas. A *belief base* is a finite set of logical formulas. The deductive closure of a belief base is called a *belief set* or *theory*. Belief revision consists in making an agent's beliefs evolve in the presence of new information of the same nature as her beliefs. The expected properties of belief revision operators are intuitively summarized by three principles:

- *Success*[3]: change must succeed, i.e. after revision, the new information item must be accepted in the new belief set.
- *Consistency*: after revision, the sets of beliefs must be consistent (In order to avoid trivialization).
- *Minimal change*: the agent's beliefs have to be modified as little as possible in order to ensure that no information is removed without necessity and no unwanted information is added.

2.1 Principles and Belief Revision Approaches

The various revision operators can be classified according to the formal framework used for representing information: logical (syntactic, semantic) or quantitative

[3] Also called priority to new information.

(for instance, numerical). Without being exhaustive, we give an overview of the main logical approaches proposed in the literature.

2.1.1 Coherentist Logical Approaches

Within these approaches, the agent's beliefs are supposed to consist of a logical theory, often represented by a single propositional formula. The revision of a (closed) belief set amounts (in the finite propositional case) to looking for the models of the new information item closest to the models of the formula representing the belief set. The minimal change principle is defined in terms of pre-orders over the language formulas (under the irrelevance of syntax hypothesis), like in the AGM approach detailed below, or in terms of models like in the Katsuno and Mendelzon (1991) approach or (Grove 1988). Within the approaches by Borgida (1985) and Dalal (1988b), this pre-order may be represented in term of distances. For Dalal's operator the chosen distance is the Hamming distance between interpretations, that is the number of propositional variables on which these interpretations differ. The revision operation consists in looking for models of the new information item whose Hamming distance from the models of the belief set is as small as possible. Within the general approach of Katsuno and Mendelzon (1991) one has to look for the models of the new information item whose plausibility is maximal, given a total pre-order over interpretations representing this plausibility relation.

Note that coherentist approaches to revision are syntax-independent. Yet, they may fail to be language-independent, namely changing the set of propositional variables used to describe a problem may affect the result of the revision process (while, for instance, inference in propositional logic is language-independent). This issue, even if already mentioned in the early nineties (Sombé 1994), was seriously studied only recently by Marquis and Schwind (2014).

2.1.2 Syntactic Approaches

With syntactic approaches, greater importance is given to the way beliefs are encoded. Revision deals with finite belief bases, that is finite sets of propositional formulas. Drawing inspiration from theory revision approaches, Nebel (1991) proposed, in particular, an operation related to partial intersections (*partial meet*) within the context of finite belief bases, briefly presented in Sect. 2.3.1. Besides, most approaches stem from the construction of consistent subbases maximal according several criteria (Benferhat et al. 1993; de Kleer 1990; Lehmann 1995). From a dual point of view, other approaches rely on the minimal withdrawal of formulas in order to restore consistency with the new additional information, like *kernel revision* where incision functions remove subsets of formulas minimal according to set inclusion (Hansson 1997) or like the approach based on *removed sets*, i.e., subsets of formulas to remove that are minimal according to cardinality (Papini 1992; Wurbel et al. 2000; Benferhat et al. 2010a). Within these approaches, two equivalent bases may be dif-

ferently revised; for instance $\{a, b\}$ has not the same meaning as the formula $a \wedge b$ any longer, which means that the comma must be interpreted in a non-classical way (Konieczny et al. 2005). These approaches are closely related to consistency restoration methods presented in the chapter "Argumentation and Inconsistency-Tolerant Reasoning" of this volume.

2.2 The AGM Approach and its Variants

Alchourrón, Gärdenfors and Makinson provided a formalization of belief revision principles for logic theories in terms of postulates (Alchourrón et al. 1985; Gärdenfors 1988). This axiomatic approach, called the AGM paradigm, became a standard for the theory of belief revision in artificial intelligence. For more details the reader is referred to the special issue of the *Journal of Philosophical Logic* dedicated to the 25 years of the AGM theory (Fermé and Hansson 2011b), or to the papers dedicated to the genesis of this theory (Gärdenfors 1992; Makinson 2003; Gärdenfors 2011; Fermé and Hansson 2011a). The AGM paradigm considers the problem of the evolution of a theory, i.e., a deductively closed set of classical logic formulas, $K = Cn(K)$ where $Cn(K)$ is the set of logical consequences of K. Actually, the AGM paradigm makes sense for any monotonic (Tarskian) logic equipped with classical connectives, for instance first-order or modal logic.

Let K be an agent's beliefs. A formula α may have 3 different epistemic statuses:

- $\alpha \in K$, the agent believes that α is true. We say that α is accepted by the agent.
- $\neg\alpha \in K$, the agent believes that α is false. We say that α is rejected by the agent.
- $\alpha \notin K$ and $\neg\alpha \notin K$, then the truth value of α is unknown (indeterminate) for the agent.

Belief change operators can be defined as transitions between these different epistemic states, as illustrated in Fig. 1.

When a formula changes from the indeterminate status to the accepted one (or symmetrically rejected one), this transition is called *expansion*, and denoted by $+$, since one only adds information (one moves from K to $K + \alpha = Cn(K \cup \{\alpha\})$). The reverse transition (from accepted/rejected to indeterminate) is called

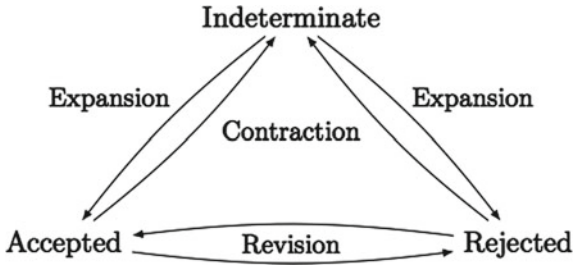


Fig. 1 Transitions between epistemic states

contraction, denoted by $-$, since one needs to remove information from the agent's beliefs. When the formula directly switches from the "accepted" status to "rejected" one (or symmetrically from "rejected" to "accepted") the transition is called *revision* and is denoted by $*$. In this case, the agent changes its mind on the truthfulness of a piece of information. As suggested by Fig. 1, it is possible to split revision, which is for the piece of information α a transition from the rejected status to the accepted one, into a contraction step, which implements the transition from the rejected status to the unknown one, followed by an expansion step, which performs the transition from the unknown status to the accepted one. This is, more formally, expressed by Levi's identity $K * \alpha = (K - \neg\alpha) + \alpha$. Besides, Harper's identity allows one to define a contraction operator from a revision one: $K - \alpha = K \cap (K * \neg\alpha)$. These identities show the very close link between revision and contraction operators, i.e., the ones can be defined from the others. However, logicians and philosophers usually consider the contraction operators as basic, while most researchers in AI select revision operators as primitive because they are the most needed ones in the field of knowledge based systems.

For these three types of belief change operators, we want to ensure that they have the expected behavior, which is reflected by the satisfaction of rationality properties. In the following, we focus on revision but we do not study expansion nor contraction operations. There exists a set of properties any *reasonable* belief change operator should satisfy. These postulates insist on minimal change and the need to maintain the consistency of the belief sets through the revision operations.

2.2.1 AGM Postulates

A revision operator $*$ is a function that maps a theory K and a formula α to a new theory $K * \alpha$ which satisfies the following properties[4]:

(**K*1**) $K * \alpha$ is a theory (*closure*).
(**K*2**) $\alpha \in K * \alpha$ (*success*).
(**K*3**) $K * \alpha \subseteq K + \alpha$ (*inclusion*).
(**K*4**) If $\neg\alpha \notin K$, then $K + \alpha \subseteq K * \alpha$ (*vacuity*).
(**K*5**) $K * \alpha = K_\perp$ if and only if $\models \neg\alpha$ (*consistency preservation*).
(**K*6**) If $\models \alpha \leftrightarrow \beta$, then $K * \alpha = K * \beta$ (*syntax independence*).
(**K*7**) $K * (\alpha \wedge \beta) \subseteq (K * \alpha) + \beta$ (*conjunctive inclusion*).
(**K*8**) If $\neg\beta \notin K * \alpha$, then $(K * \alpha) + \beta \subseteq K * (\alpha \wedge \beta)$ (*conjunctive vacuity*).

The interpretation of these postulates is the following: Postulate **K*1** expresses that the result of revision is a theory. Postulate **K*2** says that the new information item α is true in the new theory. Postulate **K*3** means that revising by new information cannot add any belief that is not a consequence of the new information item and of the theory. Postulates **K*3** and **K*4** together state that when the new information item does not contradict the initial theory, then the result of revision boils down to

[4] $K + \alpha = Cn(K \cup \{\alpha\})$. Besides $K_\perp$ denotes an inconsistent theory.

the expansion of this theory. They reflect an elementary form of minimal change (if the new information item does not contradict a prior information, the latter one is unchanged). **K*5** expresses the fact that the only way to get an inconsistent theory by revision is to revise by an inconsistent information. **K*6** says that the result of revision has to be independent from the syntactic encoding of the new information item. These six postulates are basic for revision operators.

The two postulates **K*7** and **K*8** are called additional postulates. They state that revising by the conjunction of two pieces of information amounts to a revision by the first one and an expansion by the second one whenever possible (whenever the second piece of information does not contradict any belief resulting from the first revision). This property is rather natural within different choice theories (decision, social choice etc.) (Rott 2001).

The expansion operator $K + \alpha$ has also been axiomatized in the framework of the AGM approach in order to justify the fact that it consists of adding α to K and computing the deductive closure of $K \cup \{\alpha\}$. When α is consistent with K, the expansion of K coincides with its revision.

2.2.2 KM Postulates

In order to characterize different semantic approaches within the same framework, Katsuno and Mendelzon (1991) restricted the AGM framework to standard, finite propositional logic. They reformulated the AGM postulates for the revision of a propositional formula representing a set of beliefs or a theory.

Let φ, μ and ψ be propositional formulas, respectively the prior information and the new information item. The operator $\circ$ is a revision operator if it satisfies the following postulates:

(**R1**) $\varphi \circ \mu \models \mu$.
(**R2**) If $\varphi \wedge \mu$ is consistent then $\varphi \circ \mu \equiv \varphi \wedge \mu$.
(**R3**) If μ is consistent then $\varphi \circ \mu$ is consistent.
(**R4**) If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$ then $\varphi_1 \circ \mu_1 \equiv \varphi_2 \circ \mu_2$.
(**R5**) $(\varphi \circ \mu) \wedge \psi \models \varphi \circ (\mu \wedge \psi)$.
(**R6**) If $(\varphi \circ \mu) \wedge \psi$ is consistent then $\varphi \circ (\mu \wedge \psi) \models (\varphi \circ \mu) \wedge \psi$.

Let $*$ be a revision operator over theories and $\circ$ be a revision operator on propositional formulas. We say that the operator $*$ corresponds to the operator $\circ$ if when $K = Cn(\varphi)$, then $K * \alpha = Cn(\varphi \circ \alpha)$. The interpretation is thus clear: **R1** is equivalent to **K*2** (success); **R2** is equivalent to **K*3** and **K*4** (reduction to expansion in case of consistency between prior information and the new information item); **R3** is the consistency postulate **K*5** and **R4** us the irrelevance of syntax **K*6**.[5] Finally, **R5** and **R6** are equivalent to **K*7** and **K*8**, respectively. And de facto one can prove the following theorem (Katsuno and Mendelzon 1991):

[5]This postulate is omitted if formulas are replaced by their sets of models.

Theorem 1 *Let $*$ be a revision operator and $\circ$ its corresponding operator. Then $*$ satisfies the postulates* **K*1–K*8** *if and only if $\circ$ satisfies the postulates* **R1–R6**.

2.3 Representation Theorems

While postulates describe the desirable properties of revision operators, representation theorems characterize revision operators satisfying these postulates. We only present some representation theorems (there are other ones, see for example Alchourrón and Makinson 1985; Fariñas del Cerro and Herzig 1996).

2.3.1 Partial Intersections of Theories

Partial meet revision operators rely on the idea of keeping as many formulas as possible from the initial theory, which expresses the minimal change principle. One could wish to keep the set of all subsets of the theory that do not imply the negation of information to be added. More formally, Let K be a theory and α be a proposition. The set of maximal sub-theories of K not implying $\neg\alpha$ is denoted by $K \perp \neg\alpha$.[6]

Several revision operators can be defined from this set. A first approach, called "full meet" revision, consists in considering the expansion by α of the set $\cap(K \perp \neg\alpha)$ of formulas that can be inferred from all the sub-theories of $K \perp \neg\alpha$. This revision operation, derived from the Levi's identity, is too cautious, since it is possible to show that $K * \alpha = Cn(\{\alpha\})$ whenever $\neg\alpha \in K$, in other words, all initial beliefs are forgotten (Gärdenfors 1988).

A more constructive contraction operation consists in selecting a single maximal sub-theory K' not implying $\neg\alpha$. In this case, if $\neg\alpha \in K$, it is possible to show that either $\neg\alpha \vee \beta \in K - \neg\alpha$ or $\neg\alpha \vee \neg\beta \in K - \neg\alpha$ (Gärdenfors 1988). Consequently, the revision operation obtained thanks to the Levi's identity is extreme, since for any formula β, either $\beta \in K * \alpha$ or $\neg\beta \in K * \alpha$, too many beliefs are added (resulting in complete belief sets, i.e., having a single model).

A compromise solution between the two previous revision operations is to only keep some sub-theories (the "best", the most "typical", etc ...). The contraction operation is defined by selecting a subset $\mathscr{S}(K \perp \neg\alpha)$ of $K \perp \neg\alpha$ and stating $K - \neg\alpha = \cap\mathscr{S}(K \perp \neg\alpha)$ if $\neg\alpha \in K$ (partial meet).

A representation theorem specifies that every partial meet revision operator, produced by the Levi's identity, satisfies the expected logical properties for revision and conversely every operator satisfying these logical properties can be defined by a partial meet revision operator. More formally (Alchourrón et al. 1985):

Theorem 2 *An operator $*$ is a partial meet revision operator if and only if it satisfies the postulates* **K*1–K*6**.

[6] It is reduced to $\{K\}$ if K is consistent with α.

It is possible to constrain the selection function to be *relational*. A selection function $\mathscr{S}$ is relational if and only if for every K there exists a relation $\leq$ over the union of all $K \perp \alpha$ for every non-tautological α of K such that $\mathscr{S}(K \perp \alpha) = \{K' \in K \perp \alpha \mid K' \leq K'', \forall K'' \in K \perp \alpha\}$. If $\leq$ is transitive then $\mathscr{S}$ is called *relational transitive*. Thus the following result holds (Alchourrón et al. 1985):

Theorem 3 *An operator $*$ is a relational transitive partial meet revision operator if and only if $*$ satisfies the postulates* **K*1–K*8**.

2.3.2 Epistemic Entrenchment

The formulas of a theory can be ranked according to their importance, plausibility or reliability: thus only the least entrenched formulas are removed, which reflects their tendency to remain inside a theory throughout contraction. More formally, let two formulas α and β, the notation $\alpha \leq_{EE} \beta$ means that β is at least as entrenched (certain/prioritary) as α and $\alpha <_{EE} \beta$ means that β is more entrenched than α. The following postulates have been proposed (Gärdenfors 1988). The relation $\leq_{EE}$ is called *epistemic entrenchment* if it satisfies the following properties:

(**EE1**) If $\alpha \leq_{EE} \beta$ and $\beta \leq_{EE} \gamma$, then $\alpha \leq_{EE} \gamma$ (*transitivity*).
(**EE2**) If $\alpha \models \beta$, then $\alpha \leq_{EE} \beta$ (*monotonicity*).
(**EE3**) $\alpha \leq_{EE} \alpha \wedge \beta$ or $\beta \leq_{EE} \alpha \wedge \beta$ (*conjunction*).
(**EE4**) If $K \neq K_{\perp}$, $\alpha \notin K$ if and only if $\forall \beta\ \alpha \leq_{EE} \beta$ (*minimality*).
(**EE5**) If $\beta \leq_{EE} \alpha\ \forall \beta$, then $\models \alpha$ (*maximality*).

Thanks to **EE1**, these axioms ensure that the relation $\leq_{EE}$ is a complete pre-order over the formulas of the language. **EE2** means that if a formula is implied by another, the first one is at least as entrenched as the second one, since it cannot be less certain than the latter. In the context of the other postulates **EE3** specifies that if one wishes to give up $\alpha \wedge \beta$ from K, this can be performed only removing either α or β. The loss of information resulting from the withdrawal of $\alpha \wedge \beta$ is the same as the withdrawal of one of the two. **EE4** means that the formulas that do not belong to K are minimal in $\leq_{EE}$ (the formulas at least slightly entrenched thus form a deductively closed set). **EE5** expresses that the formulas most entrenched are the tautologies. Note that two equivalent formulas are equally entrenched and this relation may be defined over sets of models. Moreover K is completely defined by $\leq_{EE}$.

Taking into account the epistemic entrenchment of the formulas of K, a contraction operation can be defined in order to only remove the least important formulas. More precisely, if α is not a tautology $\beta \in K - \alpha$ if and only if $\beta \in K$ and $\alpha \leq_{EE} \alpha \vee \beta$.

Conversely, from a contraction operation an epistemic entrenchment can be defined: $\alpha <_{EE} \beta$ if and only if $\alpha \notin K - (\alpha \wedge \beta)$. The revision operation obtained from the Levi's identity satisfies the postulates **K*1–K*8**. Moreover, the following representation theorem specifies that every revision operation satisfying the AGM postulates can be defined in terms of epistemic entrenchment (Gärdenfors 1988).

Theorem 4 *A revision operation* $*$ *satisfies* **K*1–K*8** *if and only if there exists a binary relation over the formulas of* K*, denoted by* $\leq$*, satisfying the postulates* **EE1–EE5**.

One can directly define $K * \alpha$ from the dual relation $\leq_{\Pi}$ of $\leq_{EE}$: $\alpha \leq_{\Pi} \beta$ if and only if $\neg\beta \leq_{EE} \neg\alpha$. This relation is known from Lewis (1973) as a comparative possibilistic relation and it is characteristic of possibility measures (Dubois 1986), while $\leq_{EE}$ is characteristic of necessity measures (Dubois and Prade 1991). The underlying vision of uncertainty in the AGM approach is thus possibilistic and not probabilistic. In this case, It is clear that $K = \{\alpha : \alpha >_{EE} \neg\alpha\} = \{\alpha : \alpha >_{\Pi} \neg\alpha\}$ and $K * \alpha = \{\beta : \beta \wedge \alpha >_{\Pi} \neg\beta \wedge \alpha\}$.

One can see in $K * \alpha$ the set of beliefs accepted within the environment where α is true (a form of conditioning of the possibility relation). From a plausibility relation $\succeq$ between formulas satisfying a minimal number of properties (like **EE2**), one can retrieve the six basic postulates of revision **K*1–K*6** just by imposing the closure of the set of accepted beliefs in the sense of $\succeq$ (Dubois et al. 2004).

2.3.3 Faithful Assignments

At the semantic level, Katsuno and Mendelzon (1991) interpret formulas φ in terms of *faithful assignments* that rank interpretations of the language in terms of relative plausibility. More formally, a *faithful assignment* is a function that maps any theory represented by a propositional formula φ to a pre-order over interpretations $\leq_{\varphi}$ such that:

(1) If $\omega \models \varphi$ and $\omega' \models \varphi$, then $\omega =_{\varphi} \omega'$.
(2) If $\omega \models \varphi$ and $\omega' \not\models \varphi$, then $\omega <_{\varphi} \omega'$.
(3) If $\varphi_1 \equiv \varphi_2$ then $\leq_{\varphi_1} = \leq_{\varphi_2}$.[7]

The following result holds (Katsuno and Mendelzon 1991):

Theorem 5 *A revision operator* $\circ$ *satisfies the postulates* **R1–R6** *if and only there exists a faithful assignment which maps each formula to a total pre-order* $\leq_{\varphi}$ *such that* $Mod(\varphi \circ \mu) = \min(Mod(\mu), \leq_{\varphi})$.

This result in terms of models is easely linked to Theorem 4, noting that the total pre-order $\leq_{\varphi}$ induces a possibility relation over the formulas ($\beta \leq_{\Pi} \alpha$ if and only if $\exists\omega \models \alpha, \forall\omega' \models \beta, \omega \leq_{\varphi} \omega'$). Within this approach, *more possible than* means *closer to* φ *than*, which allows one to interpret the theorem in terms of minimal change. This form of revision is the ordinal version of possibilistic conditioning, which satisfies all the AGM postulates (Dubois and Prade 1992). One can see that the theory of qualitative possibility initiated by Lewis is central to the AGM approach. De facto, the view of revision according to AGM as expressed by the revision postulates **K*7–K*8** is in agreement with possibility theory, but not with other uncertainty theories like probability theory.

[7] Since it is a semantic approach, this pre-order does not depend on the syntactic form of the formula.

3 Iterated Revision

The AGM approach focuses on the evolution of the agent's beliefs, represented by a theory. The Katsuno and Mendelzon's representation theorem shows that revision is guided by a plausibility pre-order over interpretations, however it does not discuss the evolution of this pre-order. By contrast knowing how to revise a pre-order makes it possible to iterate the revision process. However, it requires a representation richer than a simple belief set. The AGM characterization is not sufficient to model the iteration of the revision process. It requires additional constraints on the evolution of the plausibility pre-order. As we shall see later, this plausibility pre-order models the notion of *epistemic state*.

3.1 Postulates for Iterated Revision

An epistemic state, denoted by Ψ, encodes the agent's current beliefs but also other information on the relative plausibility of formulas. This epistemic state is, within the AGM framework, represented by a total pre-order $\leq_\Psi$ representing the relative plausibility of interpretations. More generally, an epistemic state is an abstract entity which symbolizes an agent's belief state, from which a (closed) belief set, denoted by $Bel(\Psi)$, can be extracted, representing the agent's accepted current beliefs, induced by the epistemic state, thanks to the pre-order $\leq_\Psi$.[8]

Darwiche and Pearl (1997) reformulated the Katsuno and Mendelzon's postulates for the revision of epistemic states and added specific postulates for its iteration. Postulates **R⋆1**, **R⋆2**, **R⋆3**, **R⋆5**, **R⋆6** are directly obtained from the KM postulates replacing φ with $Bel(\Psi)$ and $\varphi \circ \mu$ by $Bel(\Psi \circ \mu)$. In contrast, postulate **R⋆4** is weakened:

(**DP4**) If $\Psi_1 = \Psi_2$ and $\mu_1 \equiv \mu_2$, then $Bel(\Psi_1 \circ \mu_1) \equiv Bel(\Psi_2 \circ \mu_2)$.

It requires the epistemic states to be identical (not only the belief sets). This subtle difference allows for consistently making a link with suitable postulates for iteration. These new postulates will constrain the behavior of the operators during successive iterations:

(**C1**) If $\alpha \models \mu$, then $Bel((\Psi \circ \mu) \circ \alpha) \equiv Bel(\Psi \circ \alpha)$.
(**C2**) If $\alpha \models \neg\mu$, then $Bel((\Psi \circ \mu) \circ \alpha) \equiv Bel(\Psi \circ \alpha)$.
(**C3**) If $Bel(\Psi \circ \alpha) \models \mu$, then $Bel((\Psi \circ \mu) \circ \alpha) \models \mu$.
(**C4**) If $Bel(\Psi \circ \alpha) \not\models \neg\mu$, then $Bel((\Psi \circ \mu) \circ \alpha) \not\models \neg\mu$.

The interpretation of these postulates is the following. **C1** expresses that if two pieces of information are successively incorporated and if the second one implies

[8] Within the AGM approach, the epistemic state is attached to a theory K and its revision. Here the theory $Bel(\Psi)$ is dictated by the epistemic state $\leq_\Psi$. Its models are the minimal interpretations of $\leq_\Psi$.

the first one then only incorporating the second one gives the same result. **C2** does the same if two contradictory pieces of information come in.

C3 states that a piece of information should be kept if revision is performed by a piece of information which, given a belief set, implies the first one.

C4 expresses that no information can contribute to its own dismissal.

In (Darwiche and Pearl 1994) postulates **C1**–**C4** have first been proposed as additional postulates to usual Katsuno and Mendelzon's postulates **R⋆1**–**R⋆6**. Freund and Lehmann (1994) have shown that **C2** is not in agreement with the AGM postulates. Moreover Lehmann (1995) has shown that postulates **C1** and **R⋆1**–**R⋆6** imply **C3** and **C4**. Darwiche and Pearl (1997) rewrote their postulates and the AGM postulates int terms of epistemic states which solves this contradiction and makes **C3** and **C4** not redundant.

Faithful assignments have also been generalized to epistemic states. Conditions (1) and (2) are directly renewed replacing φ with $Bel(\Psi)$. In contrast, condition (3) becomes: if $\Psi_1 = \Psi_2$, then $\leq_{\Psi_1} = \leq_{\Psi_2}$. It requires the epistemic states to be equal. The representation theorem is generalized as follows in (Darwiche and Pearl 1997):

Theorem 6 *A revision operator* $\circ$ *satisfies postulates* **R⋆1**, **R⋆2**, **R⋆3**, **DP 4**, **R⋆5**, **R⋆6** *if and only if there exists a faithful assignment that maps each epistemic state* Ψ *to a total pre-order over interpretations* $\leq_\Psi$ *such that:* $Mod(Bel(\Psi \circ \mu)) = \min(Mod(\mu), \leq_\Psi)$.

A second representation theorem in (Darwiche and Pearl 1997) adds constraints relative to iteration:

Theorem 7 *A revision operator that satisfies* **R⋆1**, **R⋆2**, **R⋆3**, **DP4**, **R⋆5**, **R⋆6** *also satisfies* **C1**–**C4** *if and only if the operator and its corresponding faithful assignment satisfy:*

(**CR1**) *If* $\omega \models \mu$ *and* $\omega' \models \mu$, *then* $\omega \leq_\Psi \omega'$ *iff* $\omega \leq_{\Psi \circ \mu} \omega'$.
(**CR2**) *If* $\omega \models \neg\mu$ *and* $\omega' \models \neg\mu$, *then* $\omega \leq_\Psi \omega'$ *iff* $\omega \leq_{\Psi \circ \mu} \omega'$.
(**CR3**) *If* $\omega \models \mu$ *and* $\omega' \models \neg\mu$, *then* $\omega <_\Psi \omega'$ *only if* $\omega <_{\Psi \circ \mu} \omega'$.
(**CR4**) *If* $\omega \models \mu$ *and* $\omega' \models \neg\mu$, *then* $\omega \leq_\Psi \omega'$ *only if* $\omega \leq_{\Psi \circ \mu} \omega'$.

This representation theorem is important because it means that iterated revision operators can be considered as transition functions between total pre-orders (with the constraints given by **CR1**–**CR4**), and thus total pre-orders can be considered as the canonical representation of epistemic states, since the representation theorem expresses that whatever is the exact representation of epistemic states, it is possible to model their behavior through a faithful assignment.

Other postulates have been proposed. Boutilier (1993, 1996) proposed an *absolute minimization* postulate which can be considered as performing a minimal change on the total pre-order corresponding to epistemic states: **(CB)** If $Bel(\Psi \circ \alpha) \models \neg\mu$ then $Bel((\Psi \circ \alpha) \circ \mu) \equiv Bel(\Psi \circ \mu)$.

However, this change minimization leads to a bad behavior of the revision operator, since it totally erases the effect of α if the second piece of information μ contradicts the belief set after revising by α (Darwiche and Pearl 1997). It is somewhat

problematic that the Darwiche and Pearl's characterization permits such an operator. It thus has been proposed to define admissible revision operators, in order to remove it. These operators are defined by an additional iteration postulate (Booth and Meyer 2006):

(P) If $Bel(\Psi \circ \alpha) \not\models \neg\mu$ then $Bel((\Psi \circ \mu) \circ \alpha) \models \mu$.

A revision operator is called *admissible* (Booth and Meyer 2006) if it satisfies **R⋆1**, **R⋆2**, **R⋆3**, **DP4**, **R⋆5**, **R⋆6**, **C1**, **C2** and **P**[9] and the corresponding representation theorem is obtained in (Booth and Meyer 2006; Jin and Thielscher 2007):

Theorem 8 *Let* $\circ$ *be a revision operator which satisfies* **R⋆1**, **R⋆2**, **R⋆3**, **DP4**, **R⋆5**, **R⋆6**. *The operator* $\circ$ *satisfies* **P** *if and only if the operator and its corresponding faithful assignment satisfy:*

*(***CP***) If* $\omega \models \mu$ *and* $\omega' \models \neg\mu$, *then* $\omega \leq_\Psi \omega'$ *only if* $\omega <_{\Psi\circ\mu} \omega'$.

A generalization of iterated revision using so-called improvement operators has been proposed later in (Konieczny and Pino Pérez 2008). These operators carry out a more cautious form of change, where the plausibility of the new information item increases within the agent's epistemic state, but this new piece of information is not systematically totally accepted after revision, and therefore the success postulate **R⋆2** is not satisfied anymore.

There exists a number of iterated revision operators in the literature. We only present some of them; for a more exhaustive survey see (Rott 2009) and (Konieczny and Pino Pérez 2002b).

For instance, Boutilier proposes an operator called *natural revision* that stems from the principle of absolute minimal change, and that tries to modify as little as possible the total pre-order corresponding to the initial epistemic state. The idea is to let the most plausible models of the new piece of information be minimal in the revised pre-order, the relative ordering between the other interpretations being unchanged. This operator satisfies Darwiche and Pearl's postulates. It is the only revision operator which satisfies properties **R⋆1**, **R⋆2**, **R⋆3**, **DP4**, **R⋆5**, **R⋆6, C1–C4** and the property of absolute minimization **CB** (Darwiche and Pearl 1997). One can note that the more drastic revision rule for revision with memory proposed by Papini (2001) where all the models of the new piece of information are preferred to its counter-models (the relative ordering in both of these subsets of interpretations is preserved) allows for the reversibility of iterated revision operators thanks to a suitable encoding of total pre-orders in terms of polynomials.

3.2 Extension to Partial Pre-orders

In case of partial ignorance, totally pre-ordered information is not available. Then partial pre-orders are more suitable for representing incomplete ordinal information

[9] One can note that **C3** and **C4** are consequences of these postulates.

or incomparability. Benferhat et al. (2005) extended the KM postulates to epistemic states represented by partial pre-orders.

Postulate **R⋆6** splits into two postulates:

(**P6**) If $Bel(\Psi \circ \mu_1) \models \mu_2$ and $Bel(\Psi \circ \mu_2) \models \mu_1$ then $Bel(\Psi \circ \mu_1) \equiv Bel(\Psi \circ \mu_2)$;

(**P7**) $Bel(\Psi \circ \mu_1) \wedge Bel(\Psi \circ \mu_2) \models Bel(\Psi \circ (\mu_1 \vee \mu_2))$.

Postulate **R2** is too strong for partial preorders. It is replaced by:

(**P2**) $Bel(\Psi \circ \top) \equiv Bel(\Psi)$.

P2 expresses that the agent's current beliefs must not change in case where new information is a tautology. Two weakenings of **R⋆2** can be derived from this set of postulates:

(**P2'**) $Bel(\Psi) \wedge \mu \models Bel(\Psi \circ \mu)$.

This postulate stipulates that the common models of the initial beliefs and new information are contained in the models of the new beliefs and:

(**P2w**) If $Bel(\Psi) \models \mu$ then $Bel(\Psi \circ \mu) \equiv Bel(\Psi) \wedge \mu$

This postulate expresses that if the initial beliefs are contained in new information, the new beliefs are equivalent to the conjunction of the initial beliefs and new information.

The notion of faithful assignment has been extended to partial pre-orders using the concept of *P-faithful assignment*. With respect to Darwiche and Pearl, conditions (1) and weakened (3) are unchanged. However condition (2) is not appropriate any longer since it stipulates that each model of the agent's current beliefs is preferred to any counter-model of the current beliefs. This condition is weakened in:

(**2p**) If $\omega' \not\models Bel(\Psi)$, then there exists ω such that $\omega \models Bel(\Psi)$ and $\omega \prec_\Psi \omega'$,

stipulating that each counter-model of the agent's current beliefs is strictly less preferred than at least one model of current beliefs. The Katsuno and Mendelzon's representation theorem is generalized as follows by Benferhat et al. (2005):

Theorem 9 *A revision operator* $\circ$ *satisfies postulates* **R⋆1**, **P⋆2**, **R⋆3**, **DP4**, **R⋆5**, **P6**, **P7** *if and only if there exists a P-faithful assignment which maps each epistemic state* Ψ *to a partial pre-order over interpretations* $\preceq_\Psi$ *such that:* $Mod(Bel(\Psi \circ \mu)) = \min(Mod(\mu), \preceq_\Psi)$.

The iterated revision operators such as natural revision or revision with memory have been easily extended to partial pre-orders from a semantic point of view. In contrast, the extension to partial pre-orders of their syntactic counter-part is more complex since it consists in building a partial pre-order, called comparator, over subsets of formulas from a partial pre-order over formulas. Several comparators have been proposed: the one based on inclusion (Junker and Brewka 1989), the possibilistic order (Benferhat et al. 2004) following research works dating back to Lewis (1973) and Halpern (1997), more recently its lexicographic refinement (Yahi et al. 2008), which enables removed sets revision to be extended to partially pre-ordered belief bases (Sérayet et al. 2011).

3.3 Comments on Iterated Revision

In general, revision assumes that the new information item is of the same nature as the information to be revised. This is clearly the case with probabilistic (Jeffrey 1983), possibilistic (Benferhat et al. 2010b), or Spohn-type approaches (Spohn 2012), (see Sect. 5.1). Within Darwiche and Pearl's logical framework, this is much less clear. The epistemic state itself (the ordering over models) only implicitly appears (except in postulate **DP4**). However, in contrast one could assume that the new information item μ is another epistemic state (partially defined) and formulate postulates where these epistemic states explicitly appear. Indeed, in many iterated revision operators, the new information item μ is clearly interpreted as a constraint of the form $\mu >_\Pi \neg\mu$ which must be satisfied by the final epistemic state.

These elegant mathematical models should not lead us to forget that it is necessary, when defining a belief change operator, to properly highlight the conditions of its application and on which kind of data it operates, which is called its ontology by Friedman and Halpern (1996), in order to avoid technical developments useless in practice. Anyway, it is useful to know the nature of the data under concern in order to have additional intuitions on the way to interpret revision and the AGM framework. For example, Dubois (2008) makes the following distinctions:

- If the total plausibility pre-order encodes generic knowledge (for example, coming from a base of rules of type "birds fly") justifying the agent's beliefs on the current case (the bird Tweety, which is believed to fly), and if the new information item is of the same nature as these factual beliefs (that is, relative to the same current case: "Tweety is a penguin"), there is no reason to change this pre-order (our generic knowledge on birds). One simply restricts this pre-order to interpretations which are in agreement with the new information item and modifies the beliefs (about Tweety: "It does not fly"). This point of view assumes that the new information item is not inconsistent. In this case, belief revision is only another point of view on non-monotonic inference (Gärdenfors 1990).
- If the plausibility pre-order is considered to be of the same nature as the new information item, then this pre-order has to be revised and the AGM theory is not sufficient. This case covers two situations:
  - The plausibility pre-order and the new information item are uncertain factual information: for example, one modifies the plausibility ordering on candidates that are likely to win the elections (e.g., I had thought that Barrack was going to win; however, because Mitt was better at the televised debate, now I believe that Mitt will win). In this case, the change problem can be addressed from a merging point of view, since all information items play the same role for the agent.
  - The plausibility pre-order and the new information item both encode generic information. In this case, one can argue that the new information item does not play the same role as the pre-order (the agent is much less willing to do away with its generic knowledge than factual beliefs) and a pre-order revision

is required (Kern-Isberner 2001). However, one can prefer to directly revise the rule base that induces the plausibility pre-order (Boutilier 1996).

The axiomatic approach to revision theory is similar to the one for decision theory, where postulates deal with observable objects (choices, preferences) and the representation theorem lays bare underlying probabilities and utility functions. In revision theory, postulates are formulated on belief sets and the resulting epistemic states are implicit. However, it seems easy for an agent to declare some propositions more plausible than other ones, or to provide default rules. This remark suggests an iterated revision approach in the style of Arrow axioms, where postulates would directly deal with pre-orders and not only on belief sets that they induce (see Kern-Isberner 2001; Ma et al. 2010 for steps forward in this direction), or with uncertainty distributions (this is the case for the characterization of the Jeffrey rule in Sect. 5.1) or yet with default rules bases (Boutilier 1996).

4 Logical Approaches to Merging

Merging operators aim at combining several pieces of information given by various sources. Each source is individually consistent, however, generally they may be mutually inconsistent. Like in belief revision, some basic requirements can be stated for equally reliable pieces of information:

- *Optimism principle*: all available pieces of information have to be used.
- *Fairness principle*: no source is favored by the result of the merging.
- *Consistency principle*: the result of the merging is consistent.

Merging operators allow for defining a consistent set of pieces of information from sets of pieces of information that may be mutually inconsistent. Moreover, they may produce pieces of information that none of the initial sources alone was able to infer. This behavior illustrates the optimism principle. Suppose, for example, that one source of information knows that a is true and that another source knows $a \rightarrow b$. Then the combination of these two sources may infer that b is true, while no source alone can. The Fairness principle states that if the sources are mutually consistent then the result of merging is consistent with each of them, but if the sources are mutually inconsistent then the result of merging is either consistent with all of them or consistent with none. This principle is valid for information merging as well as for preference aggregation. Nevertheless, these two problems are distinct, even if they share a lot of common tools. In preference aggregation, the result may conflict with the preferences of each source as long as it represents an acceptable compromise. In information merging, some authors propose an alternative to Fairness principle, which states that a piece of information that is rejected by all the sources is not acceptable (Everaere et al. 2010; Dubois 2011). The consistency principle is valid for both revision and merging.

There mainly exist two families of approaches for information merging: the numerical approaches and the logical ones. The oldest approaches are the numerical ones. They mostly concern the domains such as the merging of expert opinions, often represented by probability distributions, and robotics where pieces of information coming from sensors are merged. Symbolic approaches to merging multi-source information have fostered many research works within the AI community since the 1990s (Baral et al. 1991; Revesz 1993; Lin 1996; Revesz 1997; Cholvy 1998; Konieczny and Pino Pérez 2002a). When they are syntax-independent, these symbolic approaches are relatively close to information merging based on possibility theory (Dubois and Prade 1987, 1995; Bloch 1996).

4.1 Semantic Approach to Merging Under Constraint

In the following, we consider a *profile*, denoted by $E = \{K_1, \ldots, K_n\}$ which is a multi-set of n logical bases representing belief sets. The profile represents the available information of a group of agents. We denote by $\bigwedge E$, the base $K_1 \cup \cdots \cup K_n$ whose models are the intersection of the sets of models of the bases K_i. A profile E is consistent if and only if $\bigwedge E$ is consistent. We denote by $E = E_1 \sqcup E_2$ the profile obtained by the concatenation of the two profiles E_1 and E_2. By extension $K_1 \sqcup \ldots \sqcup K_n$ is the profile consisting of the logical bases K_i.

In the same spirit as the AGM postulates for revision, Konieczny and Pino Pérez (2002a) proposed some postulates representing the expected properties of merging operators, when all sources are equally reliable. It is assumed that sources are mutually independent, i.e., there does not exist any link between them. All sources are equally important and provide consistent logical bases. Each source provides pieces of information of the same reliability and priority.

More formally, a constrained merging operator Δ is a function from a profile E and a formula μ representing an integrity constraint,[10] which returns a base $\Delta_\mu(E)$ satisfying the following properties:

(IC0) $\Delta_\mu(E) \models \mu$.
(IC1) If μ is consistent, then $\Delta_\mu(E)$ is consistent.
(IC2) If E is consistent with μ, then $\Delta_\mu(E) \equiv \bigwedge E \wedge \mu$.
(IC3) If $E_1 \equiv E_2$ and $\mu_1 \equiv \mu_2$, then $\Delta_{\mu_1}(E_1) \equiv \Delta_{\mu_2}(E_2)$.
(IC4) If $K \models \mu$ and $K' \models \mu$, then $\Delta_\mu(K \sqcup K') \wedge K \not\models \bot$ implies $\Delta_\mu(K \sqcup K') \wedge K' \not\models \bot$.
(IC5) $\Delta_\mu(E_1) \wedge \Delta_\mu(E_2) \models \Delta_\mu(E_1 \sqcup E_2)$.
(IC6) If $\Delta_\mu(E_1) \wedge \Delta_\mu(E_2)$ is consistent, then $\Delta_\mu(E_1 \sqcup E_2) \models \Delta_\mu(E_1) \wedge \Delta_\mu(E_2)$.
(IC7) $\Delta_{\mu_1}(E) \wedge \mu_2 \models \Delta_{\mu_1 \wedge \mu_2}(E)$.
(IC8) If $\Delta_{\mu_1}(E) \wedge \mu_2$ is consistent, then $\Delta_{\mu_1 \wedge \mu_2}(E) \models \Delta_{\mu_1}(E) \wedge \mu_2$.

[10] When there is no constraint, we state $\mu = \top$, i. e. μ is a tautology.

Most of these properties had already been proposed by Revesz (1997) to characterize (*model-fitting*) operators. The intuitive meaning of these postulates is the following: (**IC0**) assures that the result of merging satisfies the integrity constraint. (**IC1**) is the consistency principle. It states that, if the integrity constraint is consistent, then the result of merging should be consistent, i.e., that some consistent information can always be synthesized from the group of agents. (**IC2**) corresponds to the optimism postulate: when possible, the result of merging is the conjunction of the bases with the integrity constraint. In other words, when there is no conflict between the agents and the integrity constraint, the merging is simply the union of the bases and μ. (**IC3**) expresses that the result of merging is syntax-independent, i.e., only reflects the opinions expressed by sources and is not impacted by the syntactic form of information. (**IC4**) is the Fairness principe. When the opinions of two experts are merged, the merging operator must not give preference to one of them. (**IC5**) and (**IC6**) together express that if one can find two subgroups that agree on at least one alternative, then the result of merging is exactly the set of alternatives the two groups agree on. (**IC7**) and (**IC8**) are a direct generalization of the postulates (**R5**) and (**R6**) for revision (see Sect. 2.2.2). They state some conditions on the conjunctions of integrity constraints. In particular, they ensure that the merging process is based on a complete preordering expressing ideas of plausibility of interpretations, or proximity to belief bases in the profile.

It is possible to require additional constraints on the behavior of these operators. For instance, two important subclasses of constrained merging operators are the *majority* and *arbitration* (egalitarian) operators (Konieczny and Pino Pérez 2002a).

As in the revision case, a representation theorem shows that a constrained merging operator corresponds to a family of pre-orders over interpretations. To this end, a so-called *syncretic assignment* is defined. This is a function mapping each profile E to a pre-order $\leq_E$ over interpretations such that for any profiles E, E_1, E_2 and for any base K, K' the following conditions hold:

(**1**) If $\omega \models \bigwedge E$ and $\omega' \models \bigwedge E$, then $\omega \simeq_E \omega'$.
(**2**) If $\omega \models \bigwedge E$ and $\omega' \not\models \bigwedge E$, then $\omega <_E \omega'$.
(**3**) If $E_1 \equiv E_2$, then $\leq_{E_1} = \leq_{E_2}$.
(**4**) $\forall K, K', \forall \omega \models K \; \exists \omega' \models K' \; \omega' \leq_{K \sqcup K'} \omega$.
(**5**) If $\omega \leq_{E_1} \omega'$ and $\omega \leq_{E_2} \omega'$, then $\omega \leq_{E_1 \sqcup E_2} \omega'$.
(**6**) If $\omega <_{E_1} \omega'$ and $\omega \leq_{E_2} \omega'$, then $\omega <_{E_1 \sqcup E_2} \omega'$.

Conditions (**1**) and (**2**) express the optimism principle. (**3**) states that the merging process is syntax-independent. These three conditions are a generalization of faithful assignment conditions for revision operators (Katsuno and Mendelzon 1991). (**4**) corresponds to the Fairness principle: the pre-order associated to a profile consisting of two logical bases is such that for each model of the first one, there must exist a model of the second one that is as least as good as the first one.

(**5**) is a monotonicity property for aggregation in the broad sense (like a Pareto condition in decision theory) and (**6**) strengthens this property requiring strict monotonicity (strong Pareto condition).

The representation theorem for constrained merging operators (or **IC** merging operators) is the following (Konieczny and Pino Pérez 2002a)[11]:

Theorem 10 *An operator Δ_μ is an* **IC** *merging operator if and only if there exists a syncretic assignment that maps each profile E to a total pre-order $\leq_E$ such that $Mod(\Delta_\mu(E)) = \min(Mod(\mu), \leq_E)$.*

Constrained merging generalizes belief revision. Indeed, conditions (**1**), (**2**) and (**3**) satisfied by a syncretic assignment for merging are similar to the conditions verified by a faithful assignment for revision. Furthermore, a revision operator $\circ$ can be defined from a constrained merging operator Δ_μ by letting $K \circ \mu = \Delta_\mu(\{K\})$. Namely, if Δ_μ satisfies (**IC0**)–(**IC8**) then $\circ$ satisfies (**R1**)–(**R6**) (Konieczny and Pino Pérez 2002a).

4.2 Families of Merging Operators

We present a brief overview of the main families of merging operators.

4.2.1 Model-Based Operators

The following approach is syntax-independent. As a consequence, this approach can be encoded by an aggregation operation on pre-orders over models. It is easier to encode these pre-orders with numerical functions. The link between faithful assignments and logical bases consists in assuming that the plausibility of an interpretation is directly related to its proximity to the models of the base. Then it is natural to encode the relative plausibility (or the preference) using a distance between interpretations $d(\omega, \omega')$.[12]

The model-based operators select the interpretations that are the closest to the profile. These operators are parametrized by a distance and an aggregation function (Konieczny and Pino Pérez 2002a). An aggregation operator (see also Grabisch et al. 2009) is a family of functions $f_n, n > 0, n \in \mathbb{N}$, mapping any finite n-tuple of positive reals $x_1, \ldots, x_n$ to a positive real $y \in \mathbb{R}^+$:

- if $x \leq y$, then $f_n(x_1, \ldots, x, \ldots, x_n) \leq f_n(x_1, \ldots, y, \ldots, x_n)$ (*monotonicity*)
- $f_n(x_1, \ldots, x_n) = 0$ if and only if $x_1 = \cdots = x_n = 0$ (*minimality*)
- $f_1(x) = x$ (*identity*)

Let d be a distance between interpretations and $E = \{K_1, \ldots, K_n\}$; the pre-order $\leq_E$ over interpretations is defined as follows: $\omega \leq_E \omega'$ if and only if $d(\omega, E) \leq d(\omega', E)$, with

[11] For the infinite case, see Chacón and Pino Pérez (2006).

[12] In fact, a pseudo-distance satisfying $d(\omega, \omega') = d(\omega', \omega)$, and $d(\omega, \omega') = 0$ if and only if $\omega = \omega'$ is sufficient, because triangular inequality is not required.

$$d(\omega, E) = f_n(d(\omega, K_1) \ldots, d(\omega, K_n)),$$

where $d(\omega, K_i) = \min_{\omega' \models K_i} d(\omega', \omega)$. A constrained merging operator $\Delta_\mu^{d,f}$ is defined by: $Mod(\Delta_\mu^{d,f}(E)) = \min(Mod(\mu), \leq_E)$. The operators studied by Revesz (1997), and Lin and Mendelzon (1999) are particular cases using Hamming distance and aggregation functions Σ and max. In (Konieczny and Pino Pérez 2002a), it is proved that the properties of these operators are true regardless of the distance used. Furthermore, if the aggregation function f has some desirable properties, as the usual functions (maximum, sum, leximax, nth power sum, leximin), the obtained operators are constrained merging operators regardless of the distance and we have (Konieczny et al. 2004):

Theorem 11 *Let d be a distance between interpretations and f be an aggregation function, the operator $\Delta^{d,f}$ satisfies the properties (***IC0***), (***IC1***), (***IC2***), (***IC3***), (***IC7***) and (***IC8***).*

Theorem 12 *Let d be a distance between interpretations and f be an aggregation function, the operator $\Delta^{d,f}$ satisfies the properties (***IC0***)–(***IC8***) if and only if the aggregation function f satisfies the following properties:*

- *For all permutation of the indexes σ, $f_n(x_1, \ldots, x_n) = f_n(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ (symmetry)*
- *If $f_n(x_1, \ldots, x_n) \leq f_n(y_1, \ldots, y_n)$, then $f_n(x_1, \ldots, x_n, z) \leq f_{n+1}(y_1, \ldots, y_n, z)$. (composition)*
- *If $f_{n+1}(x_1, \ldots, x_n, z) \leq f_n(y_1, \ldots, y_n, z)$, then $f_n(x_1, \ldots, x_n) \leq f_n(y_1, \ldots, y_n)$. (decomposition)*

Instead of using numerical distance between interpretations, the set of conflicting variables can be used. This leads to a more general family of operators studied in (Everaere et al. 2008).

Alternatively, a new family of merging operators, parametrized by a distance and two aggregation functions and called DA^2 merging operators (for a **D**istance and **2 A**ggregation functions) has been introduced in (Konieczny et al. 2004). Consider a distance d between interpretations and two aggregation operators f and g. The pre-order $\leq_E$ over interpretations is defined by $\omega \leq_E \omega'$ if and only if $d(\omega, E) \leq d(\omega', E)$ with $d(\omega, E) = f_n(d(\omega, K_1), \ldots, d(\omega, K_n))$, where $E = \{K_1, \ldots, K_n\}$ and $d(\omega, K_i) = g_{m_i}(d(\omega, \alpha_1), \ldots, d(\omega, \alpha_{m_i}))$, where $K_i = \{\alpha_1, \ldots, \alpha_{m_i}\}$. The merging operator DA^2 $\Delta_\mu^{d,f,g}$ is such that $Mod(\Delta_\mu^{d,f,g}(E)) = \min(Mod(\mu), \leq_E)$. The first aggregation function g is used to extract consistent information from any base K_i even if it is inconsistent.[13] The second function f aggregates the sources. These operators are a generalization of usual semantic merging operators, and recover some formula-based merging operators.

[13] So this approach is no longer typically semantic: every base K_i could be seen as a (sub-)profile.

4.2.2 Formula-Based Operators

For such operators, the result of the merging process depends on the syntactical representation of the sources involved in the process. When the bases are finite sets of formulas, the usual syntactical merging operators select some subbases maximal (for inclusion) from the union of the bases. The combination operators proposed in (Baral et al. 1991, 1992; Benferhat et al. 1997) use inference techniques from inconsistent bases (see chapter "Argumentation and Inconsistency-Tolerant Reasoning" of this volume). These merging methods forget the origin of information, and, as a consequence, they do not take the distribution of information among the sources into account. Some natural requirements, like majority for example, cannot be considered. In (Konieczny 2000), some selection functions inspired from the transitively relational selection functions in revision are used. These operators have better logical properties, hence a better behavior. In (Konieczny 2000), three particular criteria have been studied. The first one (Δ^{D}) selects the maximal subbases (for inclusion) consistent with the maximal number of bases in a profile; it satisfies the postulates (**IC0**)–(**IC2**), (**IC3**), (**IC5**) and (**IC7**). The second one ($\Delta^{S,\Sigma}$) selects the maximal (for inclusion) consistent subbases that have minimal symmetric difference (for cardinality) with the bases in the profile; it satisfies the postulates (**IC0**)–(**IC2**), (**IC3**), (**IC7**) and (**IC8**). The third one ($\Delta^{\cap,\Sigma}$) selects all maximal (for inclusion) subbases consistent with maximal overlap (for cardinality) with the bases in a profile, it satisfies the postulates (**IC0**)–(**IC2**), (**IC5**)–(**IC8**). From a dual point of view, the Removed Sets Revision approach, based on subsets of formulas to remove in order to restore consistency, has been extended to merging. The merging strategies Σ, $Card$, Max, $GMax$ can be seen as total pre-orders on Removed Sets (Hué et al. 2007, 2008) and the results are equivalent to the ones obtained with methods relying on maximal consistent subbases (for cardinality).

The main drawback of usual semantic merging operators is that they do not consider inconsistent bases. However, in certain cases, it may be necessary or simply useful to take this information into account. Besides, the syntactical operators presented above allow for taking inconsistent bases into account but they do not consider the distribution of pieces of information among sources. The merging operators DA^2 avoid these two pitfalls. In particular, from a computational perpective, the computational complexity of these distance-based operators turns out to be not higher than the usual distance-based semantic ones and they still belong to the second level of the polynomial hierarchy. Even if, by construction, semantic merging operators are syntax-independent, they may fail to be language-independent, just like semantic revision or update operations, as studied in (Marquis and Schwind 2014).

4.3 Prioritized Merging, Merging and Iterated Revision

Delgrande et al. (2006) proposed a formal framework which links iterated revision and merging. The principle is to merge a set of formulas[14] more or less prioritized, respecting a strict priority order reflecting their importance. These authors motivate the generality of their approach by showing that the "classical" propositional merging operators (i.e., on flat bases) and the iterated merging operators (à la Darwiche and Pearl) can be seen as two extreme cases of prioritized merging. This discussion highlights the fact that in some papers about iterated revision, it seems that a confusion is made between the hypothesis of more and more reliable information and the one of more more recent information.

Their discussion on iterated revision operators reminds Friedman and Halpern's warnings on the dangers of defining change operators without specifying their ontology (Friedman and Halpern 1996). The main argument is the following. If one makes the assumption that new pieces of information successively arriving during a sequence of revisions, concern a static world (usual assumption), then there is no reason, a priori, to prefer the most recent one. If these pieces of information have different reliabilities, it is possible to take them explicitly into account in the "revision" process, even if their recency does not reflect their reliability. And the correct way to proceed is to perform prioritized merging.

The framework proposed by Delgrande, Dubois and Lang identifies the sequence of successive formulas received by an agent with an epistemic state. This assumption was already proposed in the definition of iterated revision by Lehmann (1995) and in the proposition of memory operators (Konieczny and Pino Pérez 2000; Papini 2001). Delgrande, Dubois and Lang show that the postulates of iterated revision operators can be obtained from the basic postulates of prioritized merging they propose. They also show that some postulates of constrained merging can be retrieved.

4.4 Merging in Other Logical Frameworks

Merging has also been studied within frameworks other than propositional logic. One may need to merge pieces of information that are more structured than the ones expressed in classical logic. It creates additional problems and issues. We present here a brief overview of these works.

Merging in First-Order Logic

Bloch and Lang (2002) have proposed model-based merging operators $\Delta^{d,\max}$, where the aggregation function is the maximum, based on a dilation process. By the way, it is worth noticing that the revision operator in the seminal Dalal paper (Dalal 1988a), is not defined by a distance, but using such a dilation function. Gorogiannis and

[14] Each formula may represent a base, if we want compare with merging within the propositional setting.

Hunter (2008b) have extended this approach in order to define usual model-based merging operators, i.e., not only $\Delta^{d,\max}$, but also $\Delta^{d,\Sigma}$, $\Delta^{d,Gmax}$, and $\Delta^{d,Gmin}$, in terms of dilations.

Contrary to the distance-based approach, this characterization can be extended to first-order logic. Indeed, the usual definition of model-based operators requires the computation of distances between sets of interpretations. However, as soon as a logic more expressive than propositional logic is used, this computation is not possible anymore. The interest of the definition in terms of dilation is that it possible to compute it in these logics. This only requires to select a suitable dilation function. See Gorogiannis and Hunter (2008b) for a discussion and some examples on dilation functions within the first order logic framework.

Default-Based Operators

Delgrande and Schaub (2007) have introduced two default-based merging operators. The idea is to use a specific language for each base, in order to ensure the consistency of the union of these bases and to then add as many default rules as necessary in order to identify the corresponding variables in the different languages (which reminds of Besnard and Schaub 1996's approach to inference in the presence of inconsistency).

This approach may be criticized since, like formula-based operators, these default-based operators do not take the distribution of pieces of information among sources into account. In particular, they are not majoritarian, and one piece of information believed by all sources but one may fail to appear in the result of the merging process. Nevertheless, exactly like for the formula-based operators, it seems possible to define additional policies in order to take these arguments into account, using some selection functions on equivalent maximal subsets.

Besnard et al. (2009) have also proposed an approach for merging of default theories. It relies on the notion of Minimally Unsatisfiable Subformulas (MUSes). The MUSes are computed and each formula in a MUS is replaced by a supernormal default, leading to several extensions. When the set of defaults is empty, this approach corresponds to the merging of propositional bases.

Merging Logic Programs

Some authors have studied merging operators for bases represented by logic programs and the semantic of stable models (*Answer Set Programming*).[15] This question is quite natural, because numerous works on the revision and update of logic programs exist (see for example Zhang and Foo 1998; Alves et al. 1998; Alferes et al. 2000; Eiter et al. 2002), but not for merging until recently. Delgrande et al. (2008, 2009, 2013) proposed a semantic approach in terms of Strong Equivalence (SE)-models (Turner 2003). Model-based revision and merging stemming from a distance between interpretations have been extended to logic programs. Besides, formula-based revision (or base revision) has also been extended to ASP. Krümpelmann and Kern-Isberner (2012) proposed an extension of the "remainder set" approach while Hué et al. (2009)

[15]For a more precise presentation of logic programs and the semantics of stable models, see chapter "Logic Programming" of Volume 2.

extended the "removed sets" one to ASP. These two approaches rely on the removal of rules. More recently, Zhuang et al. (2016a, b) taking advantage of the nonmonotonicity of ASP proposed another approach, called SLP revision, stemming from the removal or the addition of rules.

Similarity-Based Merging Operators

Schockaert and Prade (2009) have proposed merging operators based on a qualitative similarity relation on propositional variables: for each propositional variable, a partial pre-order on variables is provided, in which this variable is the only minimum. Two variables may obviously not be in relation, so the pre-order is necessarily *partial*. This relation can be deduced from a graph on propositional variables, where the similarity between two variables is computed as a function of the number of edges in a path between them. Schockaert and Prade use this similarity relation to try to find the best compromises for merging. The justification comes from the assumption that the conflict does not come from divergent opinions (that is, from a real conflict) but from the choice of an ontology or due to approximation issues, for example, an agent that makes no distinction between two similar concepts, like single and divorced. The use of such similarity relation allows one to use finest techniques, for conflict solving, which share common points with those used for merging systems of constraints.

Merging of Qualitative Constraint Networks

Condotta et al. (2009a, b) have proposed methods for merging qualitative constraint networks. These methods may be very useful when the networks represent spatial regions. For example, in geographical information systems (GIS), it may be necessary to merge spatial information issued from various sources. The conflicts that occur are more subtle than the ones coming from problems represented in propositional logic. In the latter the conflicts could be of type true/false while in the case of constraint networks there might be several types of more or less serious conflicts. This "granularity" between the different conflicts allows one to imagine a wider range of merging policies than in the propositional case.

Merging of Argumentation Systems

A lot of works are devoted to argumentation as a way to reason from contradictory information. Basically, a set of arguments and an attack relation among arguments are used. A general framework for argumentation has been proposed by Dung (1995) see chapter "Argumentation and Inconsistency-Tolerant Reasoning" of this volume. But these works on argumentation are limited to one agent. In (Coste-Marquis et al. 2007; Delobelle et al. 2015) and (Delobelle et al. 2016), it has been studied how to generalize these frameworks in order to take into account the fact that the arguments are distributed among a set of agents. One problem is that different agents may have argumentation systems constructed from different arguments. It is then necessary to represent these argumentation systems to be able to compare and merge them, in order to find the acceptable arguments for the group. A semantic approach has been recently proposed for merging argumentation systems in (Delobelle et al. 2016).

Merging Within Fragments of Propositional Logic

More recently, belief change within the framework of fragments of propositional logic has gained increasing attention. Indeed, when initial beliefs are expressed in a fragment of propositional logic, it may be the case that the result of the change operation does not remain inside the fragment. Several approaches have been proposed to adapt existing model-based change operators such that the result of these operators remains in the fragment under consideration. Belief change operations within the framework of fragments of classical logic constitute a vivid research branch, in particular for contraction (Booth et al. 2011; Delgrande and Wassermann 2013; Zhuang and Pagnucco 2014), revision (Delgrande and Peppas 2015; van de Putte 2013; Zhuang et al. 2013) and merging (Haret et al. 2015; Creignou et al. 2016).

Merging Within Description Logics

In the last decades, there has been an increasing use of ontologies in many application domains like for example in the semantic Web. Description Logics, which are tractable fragments of first order logic, have been recognized as a powerful formalism for representing and reasoning with ontologies (Baader et al. 2010). A DL knowledge base consists of two distinct components: a terminological base (TBox), representing generic knowledge about the application domain, and an assertional base (ABox), containing extensional knowledge (i.e., facts, individuals or constants) that instantiate terminological knowledge (For more details see chapter "Reasoning with Ontologies" of this volume). Originally, Description Logics have been introduced to represent the static aspects of a domain of interest (Baader et al. 2003). However, for some applications, knowledge may not be static and evolves from one situation to another in order to cope with changes that occur over time. Thus belief change within the framework of Description Logics has become a very active direction of research, see for example (Qi et al. 2006a; Ribeiro and Wassermann 2007; Qi and Yang 2008; Calvanese et al. 2010; Wang et al. 2010; Kharlamov et al. 2013; Zhuang et al. 2016c; Benferhat et al. 2017).

5 Non-Boolean Approaches to Information Revision and Fusion

There are also graded approaches to information revision and fusion developed in the framework of uncertainty theories, like probability or possibility theories (Dubois et al. 1998). The oldest revision and fusion techniques actually appeared first in the setting of probability theory. They consider a set of exhaustive and mutually exclusive possible worlds $\omega \in \mathscr{W}$, which can stand for interpretations of a logical language as used in preceding sections. In both probabilistic and possibilistic settings, a value in some totally ordered scale is assigned to each possible world that represents the extent to which it can be considered as the real world. Such distributions model epistemic states in a more refined way than in the pure Boolean setting. Values may

range in the unit interval [0, 1] as in the probabilistic setting, but one may use the set of integers, or even just an ordinal scale that is possibly finite. More precisely:

- In probability theory, the sum of the weights $p(\omega)$ is 1 so that $p(\omega) = 1$ means that ω is the real state of the world. Indeed it implies that $p(\omega') = 0,\ \forall \omega' \neq \omega$ for other worlds ω' considered as impossible.
- In possibility theory (Dubois et al. 1994), $\pi(\omega) = 0$ also means that ω is an impossible world, that cannot be the real one. In contrast with probability theory, $\pi(\omega) = 1$ only means that nothing prevents ω from being the real world, and consistency imposes the condition $\exists \omega, \pi(\omega) = 1$. Here, $\pi(\omega) = 1$ represents complete plausibility or total lack of surprize if ω were the case. In the qualitative setting, function π only reflects a plausibility ordering.
- In the theory of ranking functions[16] by Spohn (1988, 2012), $\kappa(\omega)$ is a natural integer (more generally, an ordinal) that represents a degree of impossibility. The plausibility scale is reversed with respect to probabilistic and possibilistic settings: $\kappa(\omega) = 0$ reflects a total lack of surprize if ω were the real world, while plain impossibility is encoded by $\kappa(\omega) = +\infty$.

There three formal representations are closely related to each other in the numerical setting. Spohn (1990) interprets $\kappa(\omega)$ as the exponent of an infinitesimal probability of the form $p(\omega) = \epsilon^{\kappa(\omega)}$, while letting $\pi(\omega) = k^{-\kappa(\omega)}$ for any real number $k > 1$ leads to the setting of possibility theory (Dubois and Prade 1991). The additive law of probability theory $P(A \cup B) = P(A) + P(B)$ if $A, B \subseteq \mathscr{W}$, $A \cap B = \emptyset$ reduces to the basic property of ranking functions $\kappa(A \cup B) = \min(\kappa(A), \kappa(B)), \forall A, B$, and in particular, $\kappa(A) = \min_{\omega \in A} \kappa(\omega)$. In possibility theory, this basic property reads $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$, where $\Pi(A) = \max_{\omega \in A} \pi(\omega)$. The dual set function $N(A) = 1 - \Pi(\overline{A})$ where $\overline{A}$ is the complement of set A, represents the degree of certainty of A.

Under a very different framework, probability and numerical possibility theories are special cases of the theory of belief functions, and more generally of the one of imprecise probabilities presented in chapter "Representations of Uncertainty in AI: Beyond Probability and Possibility" of this volume. In that framework, a possibility distribution encodes a convex set of probability measures, or yet a consonant belief function.

5.1 Valued Revision

In the above non-Boolean settings, belief change via the arrival of a new piece of information $E_I \subset \mathscr{W}$, stating that the real world lies in set E_I comes down down to a modification of the distribution pl ($= p, \pi$ or κ) into another one pl'. Generally distribution pl' results from a conditioning operation $pl'(\omega) = pl(\omega \mid E_I)$. In the following, index I indicates a new piece of information. A change operation must

[16] Originally called ordinal conditional functions (OCF).

respect the three principles underlying the idea of revision already presented in Sect. 2:

- *Consistency*: pl' is a normalized distribution of the same nature as pl (preservation of the representation),
- *Success*: the input information is considered as certainly true after revision, i.e., $\forall \omega \notin E_I$, $pl'(\omega) = 0$,
- *Minimal change principle*: the distribution after revision pl' should differ as little as possible from the prior distribution pl; for instance some distance between them should be minimal.

These approaches allow one to model what it means to revise an epistemic state by an uncertain piece of information. For instance we can specify the degree of certainty with which the new information must be held in the posterior epistemic state, in other words, how much more plausible should the possible worlds in E_I compared to those outside it. For a detailed comparative survey of valued revision methods in the XXth century, see (Dubois et al. 1998). Let us mention here the main approaches to non-Boolean revision. In contrast to theory revision in the style of AGM, they naturally lend themselves to iteration.

5.1.1 The Bayesian Approach

Among probabilistic approaches to revision, the oldest one is the Bayesian approach (Pearl 1988) in which the modification of the prior probability distribution upon the arrival of a new piece of sure information relies on Bayes rule of conditioning: $P(A|E) = \frac{P(A \cap E)}{P(E)}$ if $P(E) > 0$. In this case, minimal change can be expressed in terms of minimizing relative entropy. Besides, one may also notice that when going from a probability distribution to a conditional probability, probabilities do not change in relative value inside E since all probability values of elements of E are divided by $P(E)$. More generally, one may revise a probability measure P by another probability measure P_I defined on a partition of $\mathscr{W}$ using the same minimal change principles. It can be done by means of Jeffrey (1983)'s rule, probably the oldest revision rule in the literature: if the new piece of information is of the form $P_I(E) = a > 0$ (on the partition $\{E, E^c\}$), the revision operation is defined by: $P'(A) = aP(A|E) + (1 - a)P(A|E^c)$. This method is completely determined by the three above principles, and it also minimizes relative entropy (Fraassen 1981).

5.1.2 Qualitative Possibilistic Revision

Revision in the possibilistic setting, as proposed by Dubois and Prade (1992), considers a possibility distribution π taking values on an ordinal scale ([0, 1], for simplicity)

and a new piece of information μ (we let E be the set of models of μ) that is totally certain ($N(\mu) = 1$).[17]

Distribution π can be extended to formulas via the possibility measure Π, that gives a preorder that is the dual of an epistemic entrenchment: $\alpha \leq_{EE} \beta$ if and only if $\Pi(\neg\alpha) \geq \Pi(\neg\beta)$ (Dubois and Prade 1991). The set of beliefs $Bel(\pi) = \{\phi : \Pi(\phi) > \Pi(\neg\phi)\}$ induced by a possibility measure π is deductively closed, and its models form the set $Mod(Bel(\pi)) = \{\omega \in \mathscr{W} \mid \pi(\omega) = 1\}$. In such a framework, possibilistic revision relies on an ordinal counterpart of conditioning:

$$\pi(\omega \mid_{min} \mu) = \begin{cases} 1 & \text{if } \pi(\omega) = \Pi(\mu) \text{ and } \omega \in Mod(\mu); \\ \pi(\omega) & \text{if } \pi(\omega) < \Pi(\mu) \text{ and } \omega \in Mod(\mu); \\ 0 & \text{if } \omega \notin Mod(\mu). \end{cases}$$

This possibilistic revision is in agreement with the AGM axioms, but since it considers the new piece of information as fully certain, countermodels of μ are considered impossible, while the possibilistic ordering among the models of μ is preserved. This operator satisfies revision properties **R*1**–**R*6**, **C1**, **C3**, **C4** but it violates **C2**. Extensions of this approach to when the input information is uncertain (i.e., of the form $0 < \Pi_I(\neg\mu) = a < \Pi_I(\mu) = 1$) are proposed in Dubois and Prade (1997), Benferhat et al. (2010b), adapting Jeffrey (1983)'s rule to the qualitative setting. This possibilistic counterpart of Jeffrey's rule can subsume numerous techniques of iterated revision (Benferhat et al. 2010b).

5.1.3 Spohn-Style Revision

In this approach initiated by Spohn (1988), and presented in a more extensive monograph (Spohn 2012), an epistemic state is represented by a ranking function denoted by κ. From the links existing between this representation and the one based on possibility theory, it is clear that the set of accepted beliefs $Bel(\kappa)$ induced by κ is $Mod(Bel(\kappa)) = \{\omega \in \mathscr{W} \mid \kappa(\omega) = 0\}$.

Conditioning by some uncertain piece of information (μ, m) (understood as a constraint $\kappa_I(\neg\mu) = m > 0$) is defined by:

$$\kappa_{(\mu,m)}(\omega) = \begin{cases} \kappa(\omega) - \kappa(\mu), & \text{if } \omega \in Mod(\mu); \\ \kappa(\omega) - \kappa(\neg\mu) + m, & \text{if } \omega \notin Mod(\mu). \end{cases}$$

This operation can be viewed as an infinitesimal version of Jeffrey's revision rule, if one interprets, as done by Spohn (1990), $\kappa_I(\neg\mu) = m$ as the infinitesimal probability $P_I(\neg\mu) = \epsilon^m$. Spohn conditioning (for $m = \infty$) is thus the infinitesimal counterpart of Bayesian conditioning and also the counterpart of possibilistic conditioning based on product:

[17] It stands for $N(Mod(\mu)) = 1$ where $Mod(\mu)$ is the set of models of formula μ.

$$\kappa(\omega|\mu) = \kappa(\omega) - \kappa(\mu),$$

which becomes $\frac{\pi(\omega)}{\Pi(\mu)}$ in possibility theory, after suitable rescaling.

Another revision operator called "ordinal" proposed by Spohn (1988) comes down to defining a new ranking function κ_μ^N as follows:

$$\kappa_\mu^N(\omega) = \begin{cases} \kappa(\omega) - \kappa(\mu), & \text{if } \omega \in Mod(\mu); \\ \kappa(\omega) + 1, & \text{if } \omega \notin Mod(\mu). \end{cases}$$

This operator is of the form $\kappa_{(\mu,m)}$, with $m = \kappa(\neg\mu) + 1$. It satisfies properties **R*1**–**R*6**, **C1**–**C4**. In possibility theory language it would write $\pi_\mu^N(\omega) = \frac{\pi(\omega)}{k}$ for $\omega \notin Mod(\mu)$ letting $\pi(\omega) = k^{-\kappa(\omega)}$.

Williams (1994) systematized these ranking function revision operations by proposing a more general operation called *transmutation*, of which Spohn conditioning is a particular case, and other specific change operators were proposed on this basis (Williams 1994; Williams et al. 1995; Nayak 1994; Papini 2001). Possibilistic variants of all Spohnian revision operations have been described in (Dubois and Prade 1997; Dubois et al. 1998; Benferhat et al. 2010b). More recently, a general discussion comparing Spohn theory and possibility theory can be found in (Dubois and Prade 2016).

5.1.4 Revision in the Theory of Evidence

In the theory of evidence, an epistemic state is defined by a mass function: $2^{\mathscr{W}} \to [0, 1]$, such that $\sum_{\emptyset \neq E \subseteq \mathscr{W}} m(E) = 1$, that can be seen as a family of consistent logical theories, each encoded by a formula ϕ_i weighted by $m(Mod[\phi_i]) > 0$. The revision by a new input μ proposed by Shafer (1976), and called Dempster's rule of conditioning is constructed as follows:

1. For each formula ϕ_i consistent with μ, transfer the mass $m(Mod[\phi_i])$ to the conjunction $\phi_i \wedge \mu$, then sum all masses thus assigned to formulas equivalent to the latter.
2. Delete all formulas ϕ_i inconsistent with μ.
3. Renormalize the new mass function, so that the sum of the masses still be 1.

It should be obvious to the reader that this change rule combines Bayes rule (the renormalisation step) with an AGM-style expansion operation (in the case when μ is consistent with ϕ_i). It generally verifies the success postulate, but it is not defined if μ is inconsistent with all ϕ_i's having positive mass. It is more conveniently expressed by means of the plausibility function $Pl(A) = \sum_{B \cap A \neq \emptyset} m(B)$ in the form of a generalized Bayesian conditioning:

$$Pl(A|E) = \frac{Pl(A \cap E)}{Pl(E)} \text{ if } E = Mod[\mu]$$

The extension Jeffrey's rule of revision to belief functions has been studied by Smets (1993) and Halpern (2003). See (Ma et al. 2011) for a unified view of the AGM revision and the revision of belief functions by belief functions.

5.2 Information Fusion

In this section, only existing approaches to uncertain information fusion are reviewed. Indeed, it does not cover the issue of preference merging, which is the topic of a large literature pertaining to multicriteria evaluation and voting theory, already accounted for in chapters "Multicriteria Decision Making" and "Collective Decision Making" of this volume. The oldest information fusion methods are once more probabilistic and date back to the 1960's. There are basically two approaches, a Bayesian one and a non-Bayesian one. Ten years later, the theory of evidence has proposed an original fusion rule. It actually adapts one proposed by Dempster in 1967, and has its origin in older works by Bernoulli and Lambert in the XVIIIth century. Finally, possibility theory offers a set-theoretic view and a panoply of fusion operations, one of which is to some extent compatible with the setting of evidence theory.

The idea that there can be a unique ideal information fusion method looks delusive. The way to merge information depends of the level of conflict between information sources, and of assumptions pertaining to their reliability. Three fusion modes can be envisaged (Dubois and Prade 1987, 1995):

- A conjunctive mode consisting in focusing on the possible worlds common to all information pieces. It makes sense only if all sources are consistent with one another and considered reliable.
- A disjunctive mode that does not take sides and is tolerant to conflict while incurring a possible loss in informativeness.
- A counting-based mode that is akin to a majority-style voting process, that favors pieces of information proposed by the largest number of sources. It presupposes sources are independent. This is the prevailing approach in statistics.

It is clear that in the cases where information sources are numerous, there is a good chance that the two first approaches, if taken stricto sensu, fail. The conjunctive approach due to inconsistency, and the disjunctive approach due to a lack of informativeness. In that case, the third approach looks like a decent compromise. However, the conjunctive and disjunctive approaches can be generalized by looking for maximal subgroups of consistent sources, yielding consistent pieces of information to be combined disjunctively. The latter approach is inspired by a method to exploit inconsistent sets of logical formulas due to Rescher and Manor (1970). In the case of two sources providing incomplete information items of the form $\omega \in E_1$ and $\omega \in E_2$ it comes down to the fusion rule that yields $\omega \in E_1 \cap E_2$ if $E_1 \cap E_2 \neq \emptyset$ and $\omega \in E_1 \cup E_2$ otherwise. See (Dubois 2011) for a postulate-based justification of this scheme: it is the only one that preserves symmetry, optimism and minimal commitment.

In contrast with the Boolean framework of merging Boolean theories in the previous section, there was no consensual axiomatic approach to numerical fusion. Walley (1982) discusses many natural properties a fusion operation should satisfy, in the very general setting of imprecise probabilities. Oussalah (2000) does the same in the setting of possibility theory and Smets (2007) discusses the well-foundedness of many belief function fusion rules.

An attempt to provide a unified principled approach to information fusion can be found in the recent paper (Dubois et al. 2016), which proposes the maximal consistent subset approach as verifying the proposed postulates. They have proposed eight properties that any fusion rule should obey, namely *unanimity* (preserve what all sources find possible, delete what all sources find impossible), *information monotonicity* (the more imprecise a set of consistent sources, the less precise the result of the fusion), *consistency enforcement* (justifying renormalisation), *optimism* (assume as many sources as possible are reliable), *fairness* (the result should retain a trace of the information supplied by each source), *insensitivity to vacuous information*, *symmetry* (all sources of equal reliability should play the same role), and *minimal commitment* (the result of the fusion cannot be arbitrarily precise). They also propose a formal framework for a general representation of information items covering logic and numerical approaches. The proposed postulates essentially justify fusion rules relying on maximal consistent subsets of sources.

In the following we essentially survey the main fusion rules proposed in the various uncertainty theories.

5.2.1 Probabilistic Methods

The Bayesian approach to probabilistic fusion (Genest and Zidek 1986) assumes that each source i is characterized by a likelihood function $P_i(\mu_i|\omega)$ equals to the probability that the source declares μ_i when the real state of the world is ω.[18] Moreover, one more piece of information must be available, namely a prior probability $p(\omega)$ on each possible state of the world. In the simplest case, it is also assumed that the k sources are independent and provide observations $\mu_1, \ldots, \mu_k$. The result of the fusion is a posterior probability of each world ω obtained by means of Bayes rule:

$$p(\omega|\mu_1, \ldots, \mu_k) = \frac{(\prod_{i=1}^{k} P_i(\mu_i|\omega)) \cdot p(\omega)}{\sum_{\omega' \in \Omega} (\prod_{i=1}^{k} P_i(\mu_i|\omega')) \cdot p(\omega')}$$

This approach, which is actually a revision operation of a prior information based on several inputs, can be generalized to dependent sources using Bayesian networks (chapter "Belief Graphical Models for Uncertainty Representation and Reasoning" of Volume 2). It is often used when information sources are sensing devices.

In the alternative approach to probabilistic fusion, developed for the merging of expert opinions by Cooke (1991), each source provides a probability distribution

[18] Notice that such probabilities are attached to sources.

$p_i(\omega)$ on possible worlds, and it is assumed that the relative weight α_i of each source has been estimated prior to merging (based on experimental tests). The proposed fusion is based on counting (interpreting α_i as a number of independent replications of source i). It comes down to computing a consensus probability distribution in the form of a weighted average:

$$p_+(\omega) = \sum_{i=1}^{n} \alpha_i \, p_i(w).$$

This merging rule is the only one that is stable via projection when the probability distributions p_i are multidimensional, because the set-function $\sum_{i=1}^{n} \alpha_i P_i$ is still a probability measure.

5.2.2 Possibilistic Fusion Rules

In the possibility theory framework, it is supposed that each of the k sources provides a possibility distribution π_i on possible worlds. Any of the three fusion modes can be used (Dubois and Prade 1987; Dubois et al. 1999):

- Under the conjunctive mode, it is possible to merge the possibility distributions using a t-norm t that generalises a logical conjunction $\pi_t(\omega) = t(\pi_1(\omega), \ldots, \pi_k(\omega))$ (these operations are associative and have 1 as an identity). The presence of a conflict requires a normalization step so as to recover a possibility distribution of the form:

$$\hat{\pi}_t(\omega) = \frac{t(\pi_1(\omega), \ldots, \pi_k(\omega))}{\max_{w' \in \mathscr{W}} t(\pi_1(\omega'), \ldots, \pi_k(\omega'))}$$

  Note that this form of fusion is idempotent, if the t-norm $t = \min$ is chosen (thus it does not presuppose the independence of sources), but associativity is then lost due to the renormalization factor. In contrast if the product t-norm is chosen, there is a reinforcement effect due to the fusion rule and associativity is preserved by renormalisation. The latter product-based fusion rule was the one used in the MYCIN expert system in the 1970's, and it is very close to the Bayesian fusion rule, interpreting possibility distributions as likelihood functions, and replacing the prior probability by a uniform possibility distribution.
- If the conflict between sources is too strong (the denominator $\max_{w' \in \mathscr{W}} t(\pi_1(\omega'), \ldots, \pi_k(\omega'))$ is too small), the fusion rule becomes numerically instable, and one may use a multivalued extension of a disjunction (a t-conorm), such as the maximum: $\pi_{\max}(\omega) = \max(\pi_1(\omega), \ldots, \pi_k(\omega))$, which presupposes that at least one source is reliable without knowing which one is.
- The counting mode comes down to computing a (possibly weighted) arithmetic mean between the $\pi_i(\omega)$'s (like its probabilistic variant), followed by a renormalisation. However the weighted average of possibility measures Π_i yields a belief function.

Possibilistic fusion applies to the case when the sources supply tolerance intervals for an ill-known deterministic value. More refined approaches exist, based on an assumption on the number of reliable sources (Dubois and Prade 1995), or performing the disjunction of partial results after conjunctive merging of information items supplied by maximal consistent subsets of sources (Destercke et al. 2009).

5.2.3 Information Fusion in the Theory of Evidence

Suppose now that information items supplied by sources take the form of mass functions m_i on $\mathscr{W}$, each corresponding to a belief function. Dempster rule of combination is a conjunctive method which in some sense extends the Bayesian fusion rule and proceeds as follows. We give it for two sources only as it is associative:

- For each pair of focal sets E of m_1 and F of m_2 (such that $m_1(E) > 0, m_2(F) > 0$), we perform an intersection $E \cap F$, if not empty, and the mass $m_1(E)m_2(F)$ is assigned to this intersection.
- We normalize the mass function so that the sum of the resulting masses is 1.

This method corresponds to the formula, for $C \neq \emptyset$:

$$\hat{m}(C) = \frac{\sum_{E \cap F = C} m_1(E)m_2(F)}{\sum_{E \cap F \neq \emptyset} m_1(E)m_2(F)}$$

The Bayesian fusion rule is retrieved on two sources if we combine three mass functions m_i, $i = 1, 2, 3$, one of which is a probability measure ($m_3(\{\omega\}) = p(\omega), \forall \omega \in \mathscr{W}$, and $m_3(E) = 0$ if E is not a singleton). In this case, $\hat{m}$ coincides with the probability measure obtained by Bayes rule, letting $P(\mu_i|\omega) = \sum_{\omega \in E} m_i(E) = Pl_i(\omega)$. Besides the conditioning rule of Dempster is retrieved if $m_2(E) = 1$ in the expression of Dempster combination rule above, which points out the fact that revision can be viewed as a weighted fusion between an uncertain piece of information and a certainly true one. Like all renormalized conjunctive fusion rules, Dempster combination becomes questionable if the renormalization factor $\sum_{E \cap F \neq \emptyset} m_1(E)m_2(F)$ is too small. This fusion rule is not defined if it is 0. Then other, non-conjunctive, fusion rules must be used, especially by changing the renormalization method (one may assign the complement to 1 of the renormalisation factor to the tautology $\mathscr{W}$) or replacing the conjunction in Dempster rule of combination by a disjunction (Dubois and Prade 1986). Many alternative fusion rules can be found in (Dubois and Prade 1988; Smets 2007).

5.3 *Semantic Fusion of Weighted Knowledge Bases*

Weighted fusion applies when formulas in logical bases do not have the same importance. The most qualitative view is to consider for each source or agent a totally preordered set of formulas, from the most to the least certain ones. This situation is often encoded by means of possibilistic logic (Dubois et al. 1994) or using ranking

functions (Spohn 1988; Konieczny 2009). In the possibilistic logic setting a profile $E = \{B_1, \ldots B_n\}$ is a set of n possibilistic knowledge bases, each of which being made of a finite set of weighted formulas of the form (φ_j, a_j) with $a_j \in [0, 1]$. The formula (φ_j, a_j) expresses the idea that the degree of certainty (or priority) associated to the belief (or constraint) represented by φ_j is at least a_j. Each base B_i induces a possibility distribution π_i on interpretations as follows (see the chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" in this volume):

$$\pi_i(\omega) = \begin{cases} 1 & \text{if } \forall (\varphi_j, a_j) \in B_i,\ \omega \models \varphi_j; \\ 1 - \max\{a_j \mid (\varphi_j, a_j) \in B_i \text{ and } \omega \not\models \varphi_j\} & \text{otherwise} \end{cases}$$

The idea captured by this equation is that an interpretation is all the less plausible as it violates at least one more certain formula. To each interpretation ω is assigned a vector collecting possibility degrees assigned to ω by each profile base, denoted by $v_E(\omega) = (\pi_1(\omega), \ldots, \pi_n(\omega))$.

The intuition behind the semantic approaches proposed for the fusion of possibilistic bases is to merge the component of this vector in order to obtain a unique possibility distribution, denoted by π_E. Considering the possibility distribution as encoding a faithful assignment, it is easy to express in this setting the properties of syncretic assignments described in Sect. 4, for instance:

$\forall \omega \in \mathscr{W},$ if $\forall B_i \in E,\ \pi_i(\omega) = 1$ then $\pi_E(\omega) = 1$ (property (**1**));
$\forall \omega, \omega' \in \mathscr{W},$ if $\forall B_i \in E,\ \pi_i(\omega) \geq \pi_i(\omega')$ then $\pi_E(\omega) \geq \pi_E(\omega')$ (property (**5**)).

Other properties can be requested, for instance associativity, often found in numerical fusion rules. One may also use the classification by fusion operation modes recalled above. In particular conjunctive operators consider as plausible only interpretations that are plausible for all sources (assumed to be reliable), and disjunctive operators consider plausible any interpretation that is plausible in the sense of at least one source (again supposed reliable). Kaci et al. (2000) and Benferhat et al. (2002) studied several merging operators of this type. They also studied the extension of the logical postulates for merging knowledge bases to possibilistic logic (Benferhat and Kaci 2003).

The main issue is how to encode a possibility distribution aggregation operation, defined on interpretations, by means of a syntactic aggregation of ordered or weighted bases. For instance, merging the possibility distributions associated to two possibilistic knowledge bases using the minimum operator comes down to performing a simple set union of the bases. See (Benferhat et al. 2002) for the syntactic encoding of other possibilistic fusion operations.

Note that the aggregation of possibility distributions and the aggregation of distance functions are two very similar approaches. One may argue that the logical approach to fusion of Sect. 4 implicitly relies on an ordered knowledge base (as induced by faithful and syncretic assignments), and in practice it uses a numerical aggregation function on distances that can be expressed by means of a possibilistic fusion rule. Generally this distance is language-dependent (e.g., based on Hamming

distance between interpretations). KP and possibilistic fusion methods are thus compatible. The main differences are that no formula is considered impossible in the sense of faithful assignments and that the result of the fusion operation in the KP method is a classical knowledge base, namely the layer of most certain formulas in the weighted base resulting from the numerical aggregation. The problem is that even if the aggregation function used is associative, this associativity is lost by selecting the most certain classical knowledge base at each step (Benferhat et al. 2002). This is because the syncretic assignment obtained by merging knowledge bases K_1 and K_2 does not coincide with the faithful assignment one constructs from the merged classical base K_{12} extracted from the syncretic assignment, prior to merging it with a third knowledge base.

Other approaches to the merging of weighted knowledge bases in the possibility theory setting were proposed (Qi et al. 2006b) and the reader can consult the survey of possibilistic fusion operations by Qi et al. (2010). In the setting of Spohn ranking functions, Meyer (2001) also defined a number of fusion operations. Some of them, unsurprisingly are the translation in terms of ranking functions of usual semantic fusion operations that can be expressed on models. But some of them look remote from what is expected from a fusion operation. Finally, one property that may be useful to have is *reversibility*, that is the capacity to retrieve pieces of information prior to merging. This is possible through a suitable encoding of weights using polynomials (Seinturier et al. 2006).

All the works mentioned above, that rely on a numerical encoding of weights come down to aggregating distance values, integers or possibility degrees (numerical or ordinal). They implicitly assume a common value scale. This assumption is problematic as it presupposes that the value scale used by one source is the same or has the same meaning for any other source. This is the so-called commensurateness assumption. This assumption is very natural, if for instance sources are sensors of the same kind. But if sources are autonomous agents, commensurateness is somewhat questionable.

In particular if weighted bases express preferences rather than relative certainty, we get closer to the setting of social choice and voting methods, where the commensurateness assumption is not taken for granted (Arrow 1963). The standard assumption in voting theory is that only a total order is supplied by each agent, that is, only ordering matters. The problem of merging ordered bases can then take advantage of results and methods in voting theory and the literature on social choice (Arrow 1963; Arrow et al. 2002). Maynard-Zhang and Lehmann (2003) studied the fusion of conflicting partially ordered belief bases. More recently, Benferhat et al. (2007, 2009) proposed merging methods for logical bases that do not appeal to majority voting, not to the commensurateness assumption. Of course, it leads to fusion methods that are much less committing than those obtained under the commensurateness assumption. One interesting issue would be to check whether fusion methods that drop commensurateness come close to known voting methods or not. See (Dubois et al. 2016) for a comparison of axioms for information fusion and voting.

6 Conclusion

Belief change is a very active topic in artificial intelligence. Most research works on revision and merging have been developed in propositional logic, in possibilistic logic or in probability theory. Some approaches have been proposed in other numerical frameworks, like belief functions. In fact, revision and merging problems arise in numerous fields that do not require a logical framework, like image processing, fusion of expert opinions in reliability, robotics, radar detection, relational databases, etc. (see (Appriou et al. 2001) for a survey of such applications). The extension of revision and merging to a wide range of contexts seems to be a very promising topic of research, as presented in Sect. 4.4.

Dropping the assumption of commensurability is certainly an interesting issue to explore following the works on the merging of partially pre-ordered beliefs bases (Benferhat et al. 2007, 2009). At a more fundamental level, there are several attempts to axiomatize revision and merging operators, in addition of those detailed in Sects. 2 and 4 (for instance, for revision, the axioms justifying the Jeffrey's probabilistic rule, and for merging, the properties listed a long time ago by Walley (1982) for imprecise probabilities, a framework covering propositional logic as well as uncertainty theories). An effort to unify the characterization of revision and merging rules is necessary (see (Dubois et al. 2016) for the latter).

Despite the increasing interest in artificial intelligence for this topic, very few implementations exist for the logic approach. One partial explanation is the algorithmic complexity of decision problems linked to belief change in logic, generally at least at the second level of polynomial hierarchy and beyond (Liberatore 1997; Eiter and Gottlob 1992; Konieczny et al. 2004). Nevertheless, in certain situations, the use of heuristics and of appropriate data structures gives an average practical complexity reasonable for problems that are theoretically untractable. For the implementation of revision operators, some examples are (Williams 1995) for transmutations, (Benferhat et al. 2001) for possibilistic revision, (Wurbel et al. 2000; Benferhat et al. 2010a) for removed sets revision. For the implementation of merging operators, some examples are the merging operators defined by Bloch and Lang (2002) with an implementation based on binary decision diagrams (Gorogiannis and Hunter 2008a), the CoBA platform for default based merging (Delgrande et al. 2007), for the removed sets merging (Hué et al. 2007) an implementation based on logic programs with a semantics stemming from stable models (Hué et al. 2008) and more recently a SAT based implementation of revision and merging (Konieczny et al. 2017a, b). All these implementations remain difficult to compare in the absence of benchmarks. So the construction of sets of benchmarks for belief change, in the same spirit as the benchmarks used for the SAT problem, might be helpful in the future.

References

Alchourrón CE, Makinson D (1985) On the logic of theory change: safe contraction. Stud Log 44:405–422

Alchourrón CE, Gärdenfors P, Makinson D (1985) On the logic of theory change: partial meet contraction and revision functions. J Symb Log 50:510–530

Alferes J, Leite JA, Pereira LM, Przymusinska H, Przymusinski TC (2000) Dynamic updates of non-monotonic knowledge bases. J Log Program 45(1–3):43–70

Alves MHF, Laurent D, Spyratos N (1998) Update rules in datalog programs. J Log Comput 8(6):745–775

Appriou A, Ayoun A, Benferhat S, Besnard P, Bloch I, Cholvy L, Cooke R, Cuppens F, Dubois D, Fargier H, Grabisch M, Hunter A, Kruse R, Lang J, Moral S, Prade H, Safiotti A, Smets P, Sossai C (2001) Fusion: general concepts and characteristics. Int J Intell Syst 16(10):1107–1134

Arrow K, Sen AK, Suzumura K (eds) (2002) Handbook of social choice and welfare, vol 1. North-Holland, New York

Arrow KJ (1963) Social choice and individual values, 2nd edn. Wiley, New York

Baader F, McGuinness DL, Nardi D, Patel-Schneider PF (2003) The description logic handbook: theory, implementation, and applications. Cambridge University Press, Cambridge

Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (2010) The description logic handbook: theory, implementation and applications, 2nd edn. Cambridge University Press, New York

Baral C, Kraus S, Minker J (1991) Combining multiple knowledge bases. IEEE Trans Knowl Data Eng 3(2):208–220

Baral C, Kraus S, Minker J, Subrahmanian VS (1992) Combining knowledge bases consisting of first-order theories. Comput Intell 8(1):45–71

Benferhat S, Kaci S (2003) Fusion of possibilistic knowledge bases from a postulate point of view. Int J Approx Reason 33(3):255–285

Benferhat S, Cayrol C, Dubois D, Lang J, Prade H (1993) Inconsistency management and prioritized syntax-based entailment. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'93), pp 640–645

Benferhat S, Dubois D, Prade H (1997) Some syntactic approaches to the handling of inconsistent knowledge bases: a comparative study, part 1: the flat case. Stud Log 58:17–45

Benferhat S, Dubois D, Prade H (2001) A computational model for belief change and fusing ordered belief bases. In: Rott H, Williams MA (eds) Frontiers in belief revision. Kluwer, Dordrecht, pp 109–134

Benferhat S, Dubois D, Kaci S, Prade H (2002) Possibilistic merging and distance-based fusion of propositional information. Ann Math Artif Intell 34(1–3):217–252

Benferhat S, Lagrue S, Papini O (2004) Reasoning with partially ordered information in a possibilistic logic framework. Fuzzy Sets Syst 144(1):25–41

Benferhat S, Lagrue S, Papini O (2005) Revision of partially ordered information: axiomatization, semantics and iteration. In: Proceedings of the international joint conference on artificial intelligence (IJCAI-05), pp 376–381

Benferhat S, Lagrue S, Rossit J (2007) An egalitarist fusion of incommensurable ranked belief bases under constraints. In: Proceedings of the national conference on artificial intelligence (AAAI'07), pp 367–372

Benferhat S, Lagrue S, Rossit J (2009) An analysis of sum-based incommensurable belief base merging. In: Proceedings of the international conference on scalable uncertainty management (SUM'08). Lecture notes in computer science, vol 5785. Springer, Berlin, pp 55–67

Benferhat S, Ben-Naim J, Papini O, Würbel E (2010a) An answer set programming encoding of prioritized removed sets revision: application to GIS. Appl Intell 32(1):60–87

Benferhat S, Dubois D, Prade H, Williams M (2010b) A framework for revising belief bases using possibilistic counterparts of Jeffrey's rule. Fundam. Inform. 99:147–168

Benferhat S, Bouraoui Z, Papini O, Würbel E (2017) Prioritized assertional-based removed sets revision of dl-lite belief bases. Ann Math Artif Intell 79(1–3):45–75

Besnard P, Schaub T (1996) A simple signed system for paraconsistent reasoning. In: Alferes JJ, Pereira LM, Orlowska E (eds) Proceedings of the European workshop on logics in artificial intelligence (JELIA'96). Lecture notes in computer science, vol 1126. Springer, Berlin, pp 404–416

Besnard P, Gregoire E, Ramon S (2009) A default logic patch for default logic. In: Proceedings of the European conference on symbolic and quantitative approaches to reasoning under uncertainty (ECSQARU'09). Lecture notes in computer science, vol 5590. Springer, Berlin, pp 578–589

Bloch I (1996) Information combination operators for data fusion: a comparative review with classification. IEEE Trans Syst Man Cybern A 26(1):52–67

Bloch I, Lang J (2002) Towards mathematical morpho-logics. In: Bouchon-Meunier B, Gutierrez-Rios J, Magdalena L, Yager RR (eds) Technologies for constructing intelligent systems, vol 2. Physica-Verlag GmbH, Heidelberg, pp 367–380

Booth R, Meyer T (2006) Admissible and restrained revision. J Artif Intell Res 26:127–151

Booth R, Meyer T, Varzinczak I, Wassermann R (2011) On the link between partial meet, kernel, and infra contraction and its application to Horn logic. J Artif Intell Res 42:31–53

Borgida A (1985) Language features for flexible handling of exceptions in information systems. ACM Trans Database Syst 10:563–603

Boutilier C (1993) Revision sequences and nested conditionals. In: Proceedings of international joint conference on artificial intelligence (IJCAI'93), pp 519–525

Boutilier C (1996) Iterated revision and minimal change of conditional beliefs. J Philos Log 25(3):263–305

Calvanese D, Kharlamov E, Nutt W, Zheleznyakov D (2010) Evolution of dl-lite knowledge bases. In: Proceedings of international semantic web conference (ISWC'10), pp 112–128

Chacón JL, Pino Pérez R (2006) Merging operators: beyond the finite case. Inf Fusion 7(1):41–60

Cholvy L (1998) Reasoning about merged information. In: Dubois D, Prade H (eds) Belief change, vol 3. Handbook of defeasible reasoning and uncertainty management systems. Kluwer, Dordrecht, pp 233–263

Cholvy L, Hunter T (1997) Fusion in logic: a brief overview. In: Proceedings of European conference on symbolic and quantitative approaches to reasoning under uncertainty (ECSQARU'97). Lecture notes in computer science, vol 1244. Springer, Berlin, pp 86–95

Condotta JF, Kaci S, Marquis P, Schwind N (2009a) Merging qualitative constraint networks in a piecewise fashion. In: Proceedings of international conference on tools for artificial intelligence (ICTAI'09). IEEE Computer Society, pp 605–608

Condotta JF, Kaci S, Marquis P, Schwind N (2009b) Merging qualitative constraints networks using propositional logic. In: 10th European conference on symbolic and quantitative approaches to reasoning with uncertainty (ECSQARU'09). Lecture notes in computer science, vol 5590. Springer, Berlin, pp 347–358

Cooke RM (1991) Experts in uncertainty. Oxford University Press, Oxford

Coste-Marquis S, Devred C, Konieczny S, Lagasquie-Schiex MC, Marquis P (2007) On the merging of Dung's argumentation systems. Artif Intell 171:740–753

Creignou N, Papini O, Rümmele S, Woltran S (2016) Belief merging within fragments of propositional logic. ACM Trans Comput Log 17(3):20:1–20:28

Dalal M (1988a) Investigations into a theory of knowledge base revision: preliminary report. In: Proceedings of national conference on artificial intelligence (AAAI'88), pp 475–479

Dalal M (1988b) Updates in propositional databases. Technical report, Rutgers University

Darwiche A, Pearl J (1994) On the logic of iterated belief revision. In: Proceedings of international conference on theoretical approaches to reasoning about knowledge (TARK'94). Morgan Kaufmann, pp 5–23

Darwiche A, Pearl J (1997) On the logic of iterated belief revision. Artif Intell 89:1–29

de Kleer J (1986) An assumption-based TMS. Artif Intell 28(2):127–162

de Kleer J (1990) Using crude probability estimates to guide diagnosis. Artif Intell 45:381–392

Delgrande J, Peppas P (2015) Belief revision in Horn theories. Artif Intell 218:1–22
Delgrande J, Wassermann R (2013) Horn clause contraction functions. J Artif Intell Res 48:475–511
Delgrande JP, Schaub T (2007) A consistency-based framework for merging knowledge bases. J Appl Log 5(3):459–477
Delgrande JP, Dubois D, Lang J (2006) Iterated revision as prioritized merging. In: Proceedings of international conference on principles of knowledge representation and reasoning (KR'06), pp 210–220
Delgrande JP, Liu D, Schaub T, Thiele S (2007) COBA 2.0: a consistency-based belief change system. In: Proceedings of European conference on symbolic and quantitative approaches to reasoning under uncertainty (ECSQARU'07). Lecture notes in computer science, vol 4724. Springer, Berlin, pp 78–90
Delgrande JP, Schaub T, Tompits H, Woltran S (2008) Belief revision of logic programs under answer set semantics. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'08), pp 411–421
Delgrande JP, Schaub T, Tompits H, Woltran S (2009) Merging logic programs under answer set semantics. In: Proceedings of the international conference on logic programming (ICLP'09). Lecture notes in computer science, vol 5649. Springer, Berlin, pp 160–174
Delgrande JP, Schaub T, Tompits H, Woltran S (2013) A model-theoretic approach to belief change in answer set programming. ACM Trans Comput Log 14(2):14:1–14:46
Delobelle J, Konieczny S, Vesic S (2015) On the aggregation of argumentation frameworks. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'15), pp 2911–2917
Delobelle J, Haret A, Konieczny S, Mailly J, Rossit J, Woltran S (2016) Merging of abstract argumentation frameworks. In: Principles of knowledge representation and reasoning. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'16), Cape Town, South Africa, 25–29 April 2016, pp 33–42
Destercke S, Dubois D, Chojnacki E (2009) Possibilistic information fusion using maximal coherent subsets. IEEE T Fuzzy Syst 17(1):79–92
Doyle J (1979) A truth maintenance system. Artif Intell 12(3):231–272
Dubois D (1986) Belief structures, possibility theory and decomposable confidence measures on finite sets. Comput Artif Intell (Bratislava) 5(5):403–416
Dubois D (2008) Three scenarios for the revision of epistemic states. J Log Comput 18(5):721–738
Dubois D (2011) Information fusion and revision in qualitative and quantitative settings. Steps towards a unified framework. In: Proceedings of the european conference on symbolic and quantitative approaches to reasoning under uncertainty (ECSQARU'11), Belfast. Lecture notes in artificial intelligence, vol 6717. Springer, Berlin, pp 1–18
Dubois D, Prade H (1986) A set-theoretic view of belief functions - logical operations and approximation by fuzzy sets. Int J Gen Syst 12(3):193–226
Dubois D, Prade H (1987) Une approche ensembliste de la combinaison d'informations imprécises ou incertaines. Revue d'Intelligence Artificielle 1:23–42
Dubois D, Prade H (1988) Representation and combination of uncertainty with belief functions and possibility measures. Comput Intell 4:244–264
Dubois D, Prade H (1991) Epistemic entrenchment and possibilistic logic. Artif Intell 50:223–239
Dubois D, Prade H (1992) Belief change and possibility theory. In: Gärdenfors P (ed) Belief revision. Cambridge University Press, Cambridge, pp 142–182
Dubois D, Prade H (1995) La fusion d'informations imprécises. Traitement du Signal 11:447–458
Dubois D, Prade H (1997) A synthetic view of belief revision with uncertain inputs in the framework of possibility theory. Int J Approx Reason 17(2–3):295–324
Dubois D, Prade H (2016) Qualitative and semi-quantitative modeling of uncertain knowledge - a discussion. In: Beierle C, Brewka G, Thimm M (eds) Computational models of rationality, essays dedicated to Gabriele Kern-Isberner on the occasion of her 60th birthday. College Publications, pp 280–296

Dubois D, Lang J, Prade H (1994) Possibilistic logic. In: Gabbay DM, Hogger CJ, Robinson JA (eds) Handbook of logic in artificial intelligence and logic programming, vol 3: nonmonotonic reasoning and uncertain reasoning. Oxford Science Publications, Oxford

Dubois D, Moral S, Prade H (1998) Belief change rules in ordinal and numerical uncertainty theories. In: Dubois D, Prade H (eds) Belief change. Kluwer, Dordrecht, pp 311–392

Dubois D, Prade H, Yager R (1999) Merging fuzzy information. In: Bezdek J, Dubois D, Prade H (eds) Fuzzy sets in approximate reasoning and information systems. Kluwer Academic Publishers, Norwell, pp 335–401

Dubois D, Fargier H, Prade H (2004) Ordinal and probabilistic representations of acceptance. J Artif Intell Res (JAIR) 22:23–56

Dubois D, Liu W, Ma J, Prade H (2016) The basic principles of uncertain information fusion. An organised review of merging rules in different representation frameworks. Inf Fusion 32:12–39

Dung PM (1995) On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif Intell 77:321–357

Eiter T, Gottlob G (1992) On the complexity of propositional knowledge base revision, updates, and counterfactuals. Artif Intell 57(2–3):227

Eiter T, Fink M, Sabbatini G, Tompits H (2002) On properties of update sequences based on causal rejection. TPLP 2(6):711–767

Everaere P, Konieczny S, Marquis P (2008) Conflict-based merging operators. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'08), pp 348–357

Everaere P, Konieczny S, Marquis P (2010) Disjunctive merging: Quota and Gmin operators. Artif Intell 174(12–13)

Fagin R, Ullman JD, Vardi MY (1983) On the semantics of updates in databases. In: Proceedings of ACM SIGACT-SIGMOD symposium on principles of database systems (PODS'83), pp 352–365

Fariñas del Cerro L, Herzig A (1996) Belief change and dependence. In: Proceedings of the international conference on theoretical approaches to reasoning about knowledge (TARK '96). Morgan Kaufmann, pp 147–161

Fermé E, Hansson S (2011a) AGM 25 years. J Philos Log 40:295–331

Fermé E, Hansson S (2011b) Editorial introduction: 25 years of AGM theory. J Philos Log 40:113–114

Fraassen BV (1981) A problem for relative information minimizers in probability kinematics. Br. J. Philos. Sci. 32:375–379

Freund M, Lehmann D (1994) Belief revision and rational inference. Technical report, TR-94-16, Institute of Computer Science, The Hebrew University of Jerusalem

Friedman N, Halpern J (1996) Belief revision: a critique. In: Proceedings of international conference on principles of knowledge representation and reasoning (KR'96), pp 421–431

Gärdenfors P (1988) Knowledge in flux. MIT Press, Cambridge

Gärdenfors P (1990) Belief revision and nonmonotonic logic: two sides of the same coin? In: Proceedings of the European conference on artificial intelligence (ECAI'90), pp 768–773

Gärdenfors P (1992) Belief revision: an introduction. In: Gärdenfors P (ed) Belief revision. Cambridge University Press, Cambridge, pp 1–28

Gärdenfors P (2011) Notes on the history of ideas behind AGM. J Philos Log 40:115–120

Genest C, Zidek J (1986) Combining probability distributions: a critique and an annoted bibliography. Stat Sci 1(1):114–135

Gorogiannis N, Hunter A (2008a) Implementing semantic merging operators using binary decision diagrams. Int J Approx Reason 49(1):234–251

Gorogiannis N, Hunter A (2008b) Merging first-order knowledge using dilation operators. In: Proceedings of international symposium on foundations of information and knowledge systems (FoIKS'08). Lecture notes in computer science, vol 4932. Springer, Berlin, pp 132–150

Grabisch M, Marichal J, Mesiar R, Pap E (2009) Aggregation functions. Cambridge University Press, Cambridge

Grove A (1988) Two modellings for theory change. J Philos Log 17:157–180

Halpern J (1997) Defining relative likelihood in partially-ordered preferential structures. J Artif Intell Res 7:1–24
Halpern JY (2003) Reasoning about uncertainty. The MIT Press, Cambridge
Hansson SO (1993) Reversing the Levi identity. J Philos Log 22:637–669
Hansson SO (1997) Semi-revision. J Appl Non-Cass Log 7:151–175
Hansson SO (1998) Revision of belief sets and belief bases. In: Dubois D, Prade H (eds) Belief change, vol 3. Handbook of defeasible reasoning and uncertainty management systems. Kluwer, Dordrecht, pp 17–75
Hansson SO (1999) A textbook of belief dynamics. Theory change and database updating. Kluwer, Dordrecht
Haret A, Rümmele S, Woltran S (2015) Merging in the Horn fragment. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'15), Buenos Aires, Argentina, 25–31 July 2015, pp 3041–3047
Harper WL (1975) Rational belief change, Popper functions and counterfactuals. Synthese 30:221–262
Hué J, Papini O, Würbel É (2007) Syntactic propositional belief bases fusion with removed sets. In: Proceedings of the European conference on symbolic and quantitative approaches to reasoning under uncertainty (ECSQARU'07). Lecture notes in computer science, vol 4724. Springer, Berlin, pp 66–77
Hué J, Papini O, Würbel E (2008) Removed sets fusion: performing off the shelf. In: Proceedings of the European conference on artificial intelligence (ECAI'08) (FIAI 178), pp 94–98
Hué J, Papini O, Würbel E (2009) Merging belief bases represented by logic programs. In: Proceedings of European conference on symbolic and quantitative approaches to reasoning under uncertainty (ECSQARU'09). Lecture notes in computer science, vol 5590. Springer, Berlin, pp 371–382
Jeffrey R (1983) The logic of decision, 2nd edn. Chicago University Press, Chicago
Jin Y, Thielscher M (2007) Iterated belief revision, revised. Artif Intell 171:1–18
Junker U, Brewka G (1989) Handling partially ordered defaults in TMS. In: Proceedings of international joint conference on artificial intelligence (IJCAI'89), pp 1043–1048
Kaci S, Benferhat S, Dubois D, Prade H (2000) A principled analysis of merging operations in possibilistic logic. In: Boutilier C, MGoldszmidt (eds) Proceedings of the conference on uncertainty in artificial intelligence (UAI'00). Morgan Kaufmann, pp 24–31
Katsuno H, Mendelzon AO (1991) Propositional knowledge base revision and minimal change. Artif Intell 52:263–294
Kern-Isberner G (2001) Conditionals in nonmonotonic reasoning and belief revision - considering conditionals as agents. Lecture notes in computer science, vol 2087. Springer, Berlin
Kharlamov E, Zheleznyakov D, Calvanese D (2013) Capturing model-based ontology evolution at the instance level: the case of dl-lite. J Comput Syst Sci 79(6)
Konieczny S (2000) On the difference between merging knowledge bases and combining them. In: Proceedings of international conference on principles of knowledge representation and reasoning (KR'00), pp 135–144
Konieczny S (2009) Using transfinite ordinal conditional functions. In: Proceedings of European conference on symbolic and quantitative approaches to reasoning under uncertainty (ECSQARU'09). Lecture notes in computer science, vol 5590. Springer, Berlin, pp 396–407
Konieczny S, Pino Pérez R (2000) A framework for iterated revision. J Appl Non-Class Log 10(3–4):339–367
Konieczny S, Pino Pérez R (2002a) Merging information under constraints: a logical framework. J Log Comput 12(5):773–808
Konieczny S, Pino Pérez R (2002b) Sur la représentation des états épistémiques et la révision itérée. In: Livet P (ed) Révision des croyances, Hermes, pp 181–202
Konieczny S, Pino Pérez R (2008) Improvement operators. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'08), pp 177–186
Konieczny S, Lang J, Marquis P (2004) DA^2 merging operators. Artif Intell 157(1–2):49–79

Konieczny S, Lang J, Marquis P (2005) Reasoning under inconsistency: the forgotten connective. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'05), Edinburgh, Scotland, UK, 30 July–5 August 2005, pp 484–489

Konieczny S, Lagniez J, Marquis P (2017a) Boosting distance-based revision using SAT encodings. In: Proceedings of the workshop, LORI, pp 480–496

Konieczny S, Lagniez J, Marquis P (2017b) SAT encodings for distance-based belief merging operators. In: Proceedings of the national conference on artificial intelligence (AAAI'17), pp 1163–1169

Krümpelmann P, Kern-Isberner G (2012) Belief base change operations for answer set programming. In: Proceedings of the european conference on logics in artificial intelligence (JELIA'12), pp 294–306

Lehmann D (1995) Belief revision, revised. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'95), pp 1534–1540

Levi I (1980) The enterprise of knowledge. MIT Press, Cambridge

Lewis D (1973) Counterfactuals. Basil Blackwell, Oxford

Liberatore P (1997) The complexity of iterated belief revision. In: Proceedings of international conference on database theory (ICDT'97). Lecture notes in computer science, vol 1186. Springer, Berlin, pp 276–290

Lin J (1996) Integration of weighted knowledge bases. Artif Intell 83(2):363–378

Lin J, Mendelzon AO (1998) Merging databases under constraints. Int J Coop Inf Syst 7(1):55–76

Lin J, Mendelzon AO (1999) Knowledge base merging by majority. In: Pareschi R, Fronhöfer B (eds) Dynamic worlds: from the frame problem to knowledge management. Applied logic series, vol 12. Kluwer, Dordrecht, pp 195–218

Ma J, Liu W, Benferhat S (2010) A belief revision framework for revising epistemic states with partial epistemic states. In: Proceedings of the national conference on artificial intelligence (AAAI'10), pp 633–637

Ma J, Liu W, Dubois D, Prade H (2011) Bridging Jeffrey's rule, AGM revision and dempster conditioning in the theory of evidence. Int J Artif Intell Tools 20:691–720

Makinson D (2003) Ways of doing logic: what was different about AGM 1985? J Log Comput 13:5–15

Marquis P, Schwind N (2014) Lost in translation: language independence in propositional logic - application to belief change. Artif Intell 206:1–24

Maynard-Zhang P, Lehmann D (2003) Representing and aggregating conflicting beliefs. J Artif Intell Res (JAIR) 19:155–203

Meyer T (2001) On the semantics of combination operations. J Appl Non-Class Log 11(1–2):59–84

Nayak A (1994) Iterated belief change based on epistemic entrenchment. Erkenntnis 41:353–390

Nebel B (1991) Belief revision and default reasoning: syntax-based approach. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'91), pp 417–427

Oussalah M (2000) Study of some algebraic properties of adaptative combination rules. Fuzzy Sets Syst 114:391–409

Papini O (1992) A complete revision function in propositional calculus. In: Neumann B (ed) Proceedings of the European conference on artificial intelligence (ECAI'92). Wiley, London, pp 339–343

Papini O (2001) Iterated revision operators stemming from the history of an agent's observations. In: Rott H, Williams MA (eds) Frontiers in belief revision. Kluwer, Dordrecht, pp 281–303

Pearl J (1988) Probabilistic Reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo (CA)

Qi G, Yang F (2008) A survey of revision approaches in description logics. In: Proceedings of the 21st international workshop on description logics (DL2008), Dresden, Germany, 13–16 May 2008

Qi G, Liu W, Bell DA (2006a) Knowledge base revision in description logics. In: Proceedings of the european conference on logics in artificial intelligence (JELIA'06), pp 386–398

Qi G, Liu W, Bell DA (2006b) Merging stratified knowledge bases under constraints. In: Proceedings of the national conference on artificial intelligence (AAAI'06), pp 281–286
Qi G, Liu W, Bell DA (2010) A comparison of merging operators in possibilistic logic. In: Proceedings of the international conference on knowledge science, engineering and management (KSEM'10). Lecture notes in computer science, vol 6291. Springer, Berlin, pp 39–50
Rescher N, Manor R (1970) On inference from inconsistent premises. Theory Decision 1:179–219
Revesz PZ (1993) On the semantics of theory change: arbitration between old and new information. In: Proceedings of the 12th ACM SIGACT-SIGMOD-SIGART symposium on principles of databases, pp 71–92
Revesz PZ (1997) On the semantics of arbitration. Int J Algebr Comput 7(2):133–160
Ribeiro MM, Wassermann R (2007) Base revision in description logics – preliminary results. In: Proceedings of WOD'07, Innsbruck, Austria
Rott H (2001) Change, choice and inference: a study of belief revision and nonmonotonic reasoning. Oxford logic guides. Oxford University Press, Oxford
Rott H (2009) Shifting priorities: simple representations for twenty-seven iterated theory change operators. In: Makinson D, Malinowski J, Wansing H (eds) Towards mathematical philosophy. Springer, Berlin, pp 269–296
Schockaert S, Prade H (2009) Merging conflicting propositional knowledge by similarity. In: Proceedings of the international conference on tools for artificial intelligence (ICTAI'09). IEEE Computer Society, pp 224–228
Seinturier J, Papini O, Drap P (2006) A reversible framework bases merging. In: Proceedings of the international workshop on non-monotonic reasoning (NMR'06), pp 490–496
Sérayet M, Drap P, Papini O (2011) Extending removed sets revision to partially preordered belief bases. Int J Approx Reason 52(1):110–126
Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
Smets P (1993) Jeffrey's rule of conditioning generalized to belief functions. In: Proceedings of the conference on uncertainty in artificial intelligence (UAI'93), pp 500–505
Smets P (2007) Analyzing the combination of conflicting belief functions. Inf Fusion 8(4):387–412
Sombé L (1994) A glance at revision and updating in knowledge bases. Int J Intell Syst 9:1–27
Spohn W (1988) Ordinal conditional functions: a dynamic theory of epistemic states. In: Harper WL, Skyrms B (eds) Causation in decision, belief change, and statistics, vol 2. D. Reidel, pp 105–134
Spohn W (1990) A general non-probabilistic theory of inductive reasoning. Uncertainty in artificial intelligence. Elsevier Science, pp 149–158
Spohn W (2012) The laws of belief: ranking theory and its philosophical applications. Oxford University Press, Oxford
Turner H (2003) Strong equivalence made easy: nested expressions and weight constraints. TPLP 3:609–622
van de Putte F (2013) Prime implicates and relevant belief revision. J Log Comput 23(1):109–119
Walley P (1982) The elicitation and aggregation of belief. Technical report, Department of Statistics, University of Warwick, Coventry, UK
Wang Z, Wang K, Topor RW (2010) A new approach to knowledge base revision in dl-lite. In: Proceedings of the national conference on artificial intelligence (AAA'10)
Williams MA (1994) Transmutations of knowledge systems. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'94), pp 619–629
Williams MA (1995) Iterated theory base change: a computational model. In: Proceedings of the international joint confernce on artificial intelligence (IJCAI'95), pp 1541–1550
Williams MA, Foo N, Pagnucco M, Sims B (1995) Determining explanations using transmutations. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'95), pp 822–830
Wurbel E, Jeansoulin R, Papini O (2000) Revision: an application in the framework of GIS. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'00), pp 505–518

Yahi S, Benferhat S, Lagrue S, Sérayet M, Papini O (2008) A lexicographic inference for partially preordered belief bases. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'08), pp 507–517

Zhang Y, Foo NY (1998) Updating logic programs. In: Proceedings of the thirteenth European conference on artificial intelligence (ECAI'98), pp 403–407

Zhuang Z, Pagnucco M (2014) Entrenchment-based Horn contraction. J Artif Intell Res (JAIR) 51:227–254

Zhuang Z, Pagnucco M, Zhang Y (2013) Definability of Horn revision from Horn contraction. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'13), pp 1205–1211

Zhuang Z, Delgrande JP, Nayak AC, Sattar A (2016a) A new approach for revising logic programs. In: Proceedings of international workshop on non-monotonic reasoning (NMR'16), pp 171–176

Zhuang Z, Delgrande JP, Nayak AC, Sattar A (2016b) Reconsidering AGM-style belief revision in the context of logic programs. In: Proceedings of the european conference on artificial intelligence (ECAI'16), pp 671–679

Zhuang Z, Wang Z, Wang K, Qi G (2016c) Dl-lite contraction and revision. J Artif Intell Res (JAIR) 56:329–378

Reasoning About Action and Change

Florence Dupin de Saint-Cyr, Andreas Herzig, Jérôme Lang and Pierre Marquis

Abstract This chapter presents the state of research concerning the formalisation of an agent reasoning about a dynamic system which can be partially observed and acted upon. We first define the basic concepts of the area: system states, ontic and epistemic actions, observations; then the basic reasoning processes: prediction, progression, regression, postdiction, filtering, abduction, and extrapolation. We then recall the classical action representation problems and show how these problems are solved in some standard frameworks. For space reasons, we focus on these major settings: the situation calculus, STRIPS and some propositional action languages, dynamic logic, and dynamic Bayesian networks. We finally address a special case of progression, namely belief update.

1 Introduction

In this chapter, we are interested in *formalizing the reasoning of a single agent* who can make *observations* on a *dynamic system* and considers *actions* to perform on it. Reasoning about action and change is among the first issues addressed within Artificial Intelligence (AI); especially, it was the subject of the seminal article by McCarthy and Hayes (1969). Research in this area has been very productive until the late 1990s. Among other things, solutions to the various problems to be faced

F. Dupin de Saint-Cyr (✉) · A. Herzig
IRIT-CNRS, Université Paul Sabatier, Toulouse, France
e-mail: bannay@irit.fr

A. Herzig
e-mail: herzig@irit.fr

J. Lang
CNRS, Université Paris-Dauphine, PSL Research University, LAMSADE, Paris, France
e-mail: lang@lamsade.dauphine.fr

P. Marquis
CRIL-CNRS, Université d'Artois and Institut Universitaire de France, Lens, France
e-mail: marquis@cril.univ-artois.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_15

when dealing with action representation were put forward and a classification of action languages according to their expressive power was undertaken. Moreover, much progress towards the automatization of reasoning about action and change was made, for example through the design and the evaluation of algorithms implementing the reasoning processes of the main action languages and the investigation of the computational complexity of such processes.

The reasons why an agent may wish to act in order to modify the current state of a dynamic system or to learn more about it are numerous. For example, the goal can be to change the system into a configuration that the agent prefers over the actual one (such as moving a robot from a location to another one), or even into an optimal configuration. Alternatively, the objective can be to ensure that a certain property of the dynamic system is maintained, or that its successive states do not deviate too much from a normal path. The latter is for example the case when one wants to supervise and control a physical system, such as a furnace or that of a patient in an intensive care unit. Such scenarios involve concepts (state, action, observation, etc.) and processes connecting them (planning, prediction, explanation, etc.).

By 'formalizing', we first mean *modeling the concepts and processes* that are considered in such scenarios (the purpose is to define them rigorously from a mathematical point of view) and then *representing* them (that is, specifying how the information are encoded) and *automating* (designing algorithms suited to the processes under consideration). Note that there are typically two main reasons for modeling a dynamic system: *controlling* it (see chapter "Planning in Artificial Intelligence" of Volume 2 about planning), and obtaining more information about it, for diagnosing it or supervising it (see chapter "Diagnosis and Supervision: Model-based Approaches" of Volume 1). Once modeled, the same concept can be associated with several representations. If the choice of a specific model typically depends on the available pieces of information and what one wants to do with them, the choice of a representation (suited to a given model) is based on other criteria, such as computational efficiency and succinctness.

2 Reasoning About Action: Models

2.1 Basic Concepts and the Corresponding Models

In this section, we define some mathematical notions corresponding to the key concepts considered in reasoning about action and change.

The *model* of a reasoning process on *a dynamic system* can be divided in two parts: the model of the system (with its own dynamics) and the model of the agent (including her knowledge about the system). Sandewall (1995) has developed a taxonomy of reasoning problems on dynamic systems, and the remainder of this chapter elaborates on it.

Throughout the chapter, we assume that *time* is discrete (which is a common assumption in Artificial Intelligence). The *horizon* of the process is the set $\mathscr{H}$ of relevant steps for controlling and observing it. It can be finite ($\mathscr{H} = \{0, \ldots, N\}$ with $N \in \mathbb{N}$) or infinite ($\mathscr{H} = \mathbb{N}$); a degenerate case of a finite horizon is when there is only one change step ($\mathscr{H} = \{0, 1\}$).

A *state* is the description of the system at a given time point. Unless stated otherwise, *the set of all states*, denoted by $\mathscr{S}$, will be assumed finite. A *state trajectory* is a sequence of elements of $\mathscr{S}$, indexed by elements of $\mathscr{H}$. The system states at the different time points of $\mathscr{H}$ may only be partially known by the agent.

The specification of a reasoning process on a dynamic system requires first the *beliefs* of the agent about the state of the system at different time points (including the initial time point) and about the general laws that govern the evolution of the system. Thus, we first have to choose a model for uncertain belief states. For space reasons, we will focus only on two uncertainty models in the following: the *binary model*, where belief states b are *non-empty subsets of* $\mathscr{S}$ and the *Bayesian model*, where belief states b are *probability distributions* on $\mathscr{S}$.[1]

Transitions from one state to another are triggered by *events*. These events generally change not only the state of the system, but also the beliefs of the agent. An *action* is a special event triggered by an agent. The agent has a model of each *action* available to her. The set of actions available to the agent is denoted by $\mathscr{A}$, and is supposed to be finite. The agent can also have a model for *exogenous events*, which are phenomena whose dynamics are similar to actions but which are not triggered by the agent. They are triggered by nature or possibly by other agents more or less well-identified (i.e., whose identity may be imperfectly known), and their occurrences are a priori not known by the agent. We distinguish between the *action type* α (defined very generally) from the action occurrence(s) at one or more time point(s): a given action may have no occurrence in an instance of a problem, or may have one or several occurrences. Actions have two types of effects: *ontic (physical) effects*, on the world, and *epistemic effects*, on the beliefs of the agent. Epistemic effects can either be caused by her projection of the physical effects of the performed action (for instance, if I know that the action "delete file F" has the effect that file F no longer appears on my computer, then, when I execute this action, the resulting belief state is such that I know that F no longer is on my computer) or from *observations* or any form of feedback (for instance, if after trying to turn the light on by flipping the switch, I observe that the light is off, then, in my new state of belief, I know that the bulb is broken or that the power is off).

Actions have generally two types of effects at once (as in the case of the "switch" action above). Some actions, referred to as *purely epistemic actions*, have only epistemic effects, and no effect on the state of the world; for example, measuring a temperature, or querying a database. Other actions, referred to as *purely ontic actions*

[1] There are many other uncertainty models that should be mentioned but will not be, for space reasons - they include ordinal models, where belief states and action effects are modeled as pre-orders over $\mathscr{S}$, possibilistic models that are similar in spirit to them, non-Bayesian probabilistic models, where a belief state is a family of probability distributions, etc. (see chapter "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility" of this volume).

have epistemic effects (it is hard to imagine actions without any epistemic effect, apart from the action "do nothing"), but these epistemic effects are the simple projection, by the agent, of what she knows about the ontic effects of the action (as for the action "delete file F" above). In other words, a purely ontic action gives *no feedback* to the agent: her belief state after the execution of such an action coincides with the belief state she could foresee before executing the action ("what you foresee is what you get"). Every action can be decomposed in a unique way into a purely ontic action and a purely epistemic action; without loss of generality, we can thus assume that each available action is either purely ontic or purely epistemic (and we will make such an assumption in the rest of the chapter, unless stated otherwise).

Let us start by describing *purely ontic actions*. The effects of a purely ontic action α are defined by a *transition system* between the states of the world, modeled as a *binary relation* $\boldsymbol{R}_\alpha$ over $\mathscr{S}$.

The simplest case is when actions are *deterministic and always executable*. In this case, the transition system of α is a mapping R_α from $\mathscr{S}$ to $\mathscr{S}$. An action has *conditional effects* if the resulting state after its execution depends on the state before its execution. For example, the action "switch off the light" may be regarded as deterministic and unconditional (if we assume that it always has the effect that the lamp is off after its execution). "Toggle the switch" can be considered as deterministic, and with conditional effects since its effects depend on the state ("on" or "off") of the light before the execution of the action.

More generally, actions are not always executable: there can be states s such that $R_\alpha(s) = \emptyset$; actions can also be *non-deterministic*: there are states s such that $R_\alpha(s)$ contains more than one element. For example, the action "delete file F" is not executable if the file does not exist; in this case, the modeler will define the effect of the action only for states where the file exists, and executing the action will be forbidden in the other cases. Another model would make advantage of an action with conditional effects, where the transition associated with the action would be associated with the identity relation in situations where file F does not exist, and would lead to states where the file is deleted otherwise.

In the non-deterministic case, the transition model chosen depends on the nature of the uncertainty one wants to deal with; with each initial state is associated a belief state on the subsequent states. Note that choosing a deterministic or a non-deterministic model for a system may depend on the knowledge and the goals of the modeler: the action "turn the computer off" can be considered as non-deterministic for an agent who is not a computer scientist (since it may happen that after the execution of the action the computer is still on) but as deterministic, with conditional effects, for an expert in computer science (since this expert will be able to determine the cases where the computer stays on after being turned off). Modeling a dynamic system as a transition system between states amounts to making the implicit assumption that the system is Markovian.[2] Such an assumption can be made without loss of generality by considering more complex states (encoding state trajectories). For the sake of brevity,

[2] A system is Markovian if the transition of the system to any given state depends only on the current state and not on the previous ones.

we will stick to the following two models: the binary non-deterministic model and the stochastic model.

In the binary non-deterministic model, the transition system of an action α is a mapping R_α from $\mathscr{S}$ to $2^{\mathscr{S}}$ (or to $2^{\mathscr{S}} \setminus \{\emptyset\}$, when α is always executable). For example, if the system states are $\mathscr{S} = \{c\_on, c\_stand\_by, c\_off\}$ (representing the activity of a computer: "on", "stand-by" or "off") then the action of "shutting down the computer" may be modeled as $R_{shut\_down}(c\_on) = \{c\_on, c\_off\}$, $R_{shut\_down}(c\_stand\_by) = R_{shut\_down}(c\_off) = \emptyset$ (meaning that one can "turn off the computer" only if it is "on", and in this case, it is not sure that the "shut down" action succeeds). Note that if α is a purely epistemic action, then $R_\alpha(s) = \{s\}$ for all s.

In the stochastic model (here, the Bayesian model for uncertainty), R_α is a stochastic matrix, i.e., a family of probability distributions $p(.|s, \alpha)$ for $s \in \mathscr{S}$, where $p(s'|s, \alpha)$ is the probability to obtain the state s' after the execution of α in s. In this model, it is possible to specify how much the "shut down" action succeeds; thus, R_{shut_down} could be represented by $p(c\_on \mid c\_on, shut\_down) = 0.1$, $p(c\_off \mid c\_on, shut\_down) = 0.9$, $p(c\_stand\_by \mid c\_on, shut\_down) = 0$.

Epistemic effects of actions are expressed in terms of feedback. The actions that the agent decides to execute do not depend directly on the system state (which may be unknown to the agent) but on the agent's beliefs (and in particular, on what has been observed earlier). Ideally, the current state and what the agent may observe coincide. In this case, the belief state of the agent is perfect, but this hypothesis reflects an ideal case and does not often hold. In order to define the epistemic effects of actions, an *observation space* $\boldsymbol{\Omega}$ can be introduced in the model. This space, unless otherwise indicated, is supposed finite. The observations are the feedback given by the system and each observation at a given time point from the horizon is a projection of a state (not necessarily totally observed) of the system. The observations are called *reliable* if this projection corresponds to the *actual state* of the system (unreliable observations can arise from faulty sensors, for example).

Taking observations into account concerns two distinct stages: the off-line stage when the decision policy is generated, and the on-line stage when it is exploited (i.e., when the plan is executed). During the off-line stage, the agent who generates the policy takes advantage of her knowledge about the observations which could be made at the on-line stage. During the on-line phase, the actions which are triggered by the agent typically depend on the observations which are effectively made. Note that nothing prevents from having two distinct agents (one who computes a decision policy and another one who executes it).

Two assumptions corresponding to two extreme cases are commonly made: when the system is *fully observable*, the observation space is identical to the state space: when she generates a decision policy, the agent knows that at the on-line stage, the actual state of the system will be known exactly at each time point; when the system is *non-observable*, the observation space is a singleton $\{o^*\}$, where o^* is a fictitious observation (the empty observation): the system gives no feedback.

When none of these two extreme assumptions hold, one faces the more general situation of *partial observability*, where observations and system states are constrained

by an observation-state *correlation structure*, the definition of which varies with the uncertainty model under consideration. In general, each state s and each *epistemic action* α correspond to a *belief state* $\boldsymbol{O_\alpha(s)}$ over the space of observations, representing the prior beliefs about the observations obtained when action α is performed in state s. In the binary model for uncertainty, a family of sets $\boldsymbol{O}_\alpha(s)$ for $s \in \mathscr{S}$ is thus considered, where $O_\alpha(s)$ *is a non-empty set of* Ω; so $o \in O_\alpha(s)$ reflects the fact that s is a state compatible with the observation o which results from the execution of α. If α is purely ontic, then $O_\alpha(s) = \{o^*\}$ for all s. In the Bayesian model for uncertainty, the *feedback* is modeled by a probability distribution $p(.|s, \alpha)$ on Ω where $\boldsymbol{p}(o|s, \alpha)$ *is the probability to observe* o *when* α *is executed in* s.

In this section, we assumed that only one action at a time can be executed. In some problems, it is natural to perform several actions in a *concurrent* way. This requires to be able to define the effects of combinations of actions; for reaching this goal, the same models as previously considered can be used, viewing every possible combination of actions as a specific action. A typical example (Thielscher 1995) is the one of a table that can be lifted by the right side or by the left side: the two actions performed in sequence and independently do not have the same effect as when they are executed simultaneously, especially when a glass of water is on the table!

2.2 Types of Reasoning and Their Implementations

Reasoning on a dynamic system requires to take account of a time horizon, the prior beliefs on the system (general laws of the domain and action effects), the occurrences of actions at some time points, and the observations at given time points (it is a simplified model – see (Sandewall 1995) for a more general one, where, in particular, the actions can have a duration). We are now going to approach some specific types of reasoning implying reasoning on a dynamic system, as well as their implementation by means of algorithms.

2.2.1 Prediction and Postdiction with Ontic Actions

Prediction (also called *projection*) consists in determining, according to one initial state of belief b and the description of a purely ontic action α, the new state of belief b' resulting from the application of α in b. The transformation of a state of belief into another one by an action is called *progression*; noted $b' = prog(b, \alpha)$. Of course, the formal definition of $prog$ depends on the nature of the space of the beliefs (static and dynamic), thus it depends on the chosen representation of uncertainty. In the simplest case (that of classical planning) where belief states are perfect, actions are deterministic and always achievable, each $prog(., \alpha)$ is a total function mapping a state to another one. In the binary nondeterministic model, a state s' is possible after the execution of α in the state of belief $b \subseteq \mathscr{S}$ if there exists a possible state s in the whole set of states corresponding to the initial belief b, such as s' is a possible result of

α in s, i.e. $prog(b, \alpha) = \bigcup_{s \in b} R_\alpha(s)$. In the probabilistic model, the obvious choice is obtained by identifying the model of the process to a Markov chain: $prog(b, \alpha)$ is the probability distribution b' on $\mathscr{S}$ defined by $b'(s') = \sum_{s \in \mathscr{S}} b(s)p(s'|s, \alpha)$ (where $p(.|s, \alpha)$ is the probability distribution associated with R_α).

The second type of reasoning is *postdiction*. It consists in determining, according to one final state of belief b' and the description of a purely ontic action α which has just been carried out, the state of belief b before the action was done. This transformation of a state of belief into another, is sometimes also called *regression* or *weak regression*; noted $b = reg_w(b', \alpha)$. The weak regression corresponds to the progression by the reverse action of α (noted α^{-1}), which transition system $R_{\alpha^{-1}}$ is the reciprocal relation of the relation R_α; thus it holds that $reg_w(b', \alpha) = prog(b', \alpha^{-1}) = \{s | R_\alpha(s) \cap b' \neq \emptyset\}$.

Postdiction must be distinguished from *goal regression*, also called *strong regression*, which is the reverse transformation of progression. It is defined only for the binary model [3]: given a belief state $b' \subseteq \mathscr{S}$ and a purely ontic action α, the aim is to find the belief state $b = reg_S(b', \alpha)$ such that $prog(b, \alpha) \subseteq b'$ and b is maximum for set inclusion; this belief state is the least informative state of belief (thus the least conjectural) which guarantees that the execution of α in it led to the goal b'.

Let us notice that $reg_S(b', \alpha) \subseteq reg_w(b', \alpha)$ with the particular case $reg_S(b', \alpha) = reg_w(b', \alpha)$ when α is deterministic.

Progression and regression are two key processes of reasoning for *planning* (see chapter "Planning in Artificial Intelligence" of Volume 2), which consists in determining the actions to carry out to make evolve the system as the agent wishes it (for example, get as close as possible to a reference trajectory in the case of the supervision, or to reach a goal state in the case of classical planning). On the other hand, postdiction has little interest for planning itself (because if b is obtained as a possible postdiction from b' with action α, it is not guaranteed that by carrying out the action α one would again obtain the state b', while strong regression guarantees it by definition).

2.2.2 Prediction and Postdiction with Epistemic Actions

The progression of a belief state by an epistemic action depends on the nature of the reasoning process. In the case of a supervision process or a diagnosis, the agent reasons online and thus has access to all the observations coming from the actions *feedback* during its reasoning; thus it is enough to define the progression of a belief state by an observation, which is related to *belief revision* (see chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume). In the binary model, the progression of one belief state $b \subseteq \mathscr{S}$ by an observation o after having carried out the action α is $b \cap S(o)$, where $S(o) = \{s | o \in O_\alpha(s)\}$; while in the Bayesian model, the revision of b by o is the probability distribution $b(.|S(O))$.

[3] In the probabilistic model, there may not exist a unique probability distribution b on $\mathscr{S}$ satisfying $b'(s') = \sum_{s \in \mathscr{S}} b(s).p(s'|s, \alpha)$, $b'(s')$ with $p(s'|s, \alpha)$ being known for all s, s' and α.

The *filtering* process consists in determining the new state of belief b', given an initial belief state b, an action α, and an observation o resulting from the execution of α. In the binary model, this new belief state is simply $prog(b, \alpha) \cap S(o)$. In the Bayesian model, the probability distribution b' obtained after having carried out α and observing o is $b'(s') = \frac{p(o|s',\alpha).\sum_{s\in S} b(s).p(s'|s,\alpha)}{\sum_{s''\in S}(p(o|s'',\alpha).\sum_{s\in S} b(s).p(s''|s,\alpha))}$; it is the formula expressing the revision of the beliefs by the *feedback* in the partially observable Markov decision processes (see chapter "Planning in Artificial Intelligence" of Volume 2).

In the case of a planning process, where the aim is to build an off line plan and to reason on its effects, the progression of a belief state by an epistemic action is in general not a unique belief state, but a set of such states (one for each possible observation, since the actual observation cannot be known off line). In the binary model, $prog(b, \alpha)$ is the set of belief states $\{b \cap S(o) \mid b \cap S(o) \neq \emptyset\}$ for o varying in Ω. For sake of shortness, we do not give details on regression by epistemic actions.

2.2.3 Event Abduction

The third type of reasoning is *event abduction*. It concerns reasoning on the event which took place between two successive time points t and $t + 1$, starting from the description of the possible events and from the description of the belief states at time t and $t + 1$.[4] If the event in question is exogenous, this reasoning is called *explanation*.

As for planning, progression and goal regression are two key processes for event abduction: in planning, one must choose the actions to be carried out to make the system evolve as desired starting from its current state; in event abduction, the objective is to determine which event α led the system to evolve as it did between t and $t + 1$ (even if this evolution was not desirable). In the binary model, to compute such α consists in searching among the possible events those satisfying $b' \subseteq prog(b, \alpha)$ (or equivalently $b \supseteq reg_F(b', \alpha)$).

2.2.4 Scenario Extrapolation

More generally, these types of reasoning, that we defined in a context where there is only one change stage (thus two time points), take place in situations where the horizon is unspecified, and where the input information is a complex *scenario* describing a partial trajectory of the system (at each time point, some information may be available about the occurrence of an action and/or an observation). In the typical case where no action was carried out and where the user wants to find the events (or more simply, the elementary changes) which occurred at each time point, the process is called *extrapolation*.

[4] A more complex abduction problem consists in reasoning not only on the event which took place, but also on the system states at time points t and $t + 1$, on which one wishes to obtain more precise beliefs.

Another situation is when one seeks to recognize some trajectories among a set of reference trajectories in order to predict the events that will occur and/or the states that the system will reach; this process is called *scenario monitoring* or *scenario recognition*.

The sequence of observations can also contain action occurrences (Dupin de Saint-Cyr 2008; Delgrande and Levesque 2012; Hunter and Delgrande 2015) (scenarios are also called narratives or histories) and the two previous tasks of completing or recognizing some trajectories can be done in this more complex context. These tasks involve both prediction, postdiction and event abduction in situation that can be pervaded with uncertainty (fallible knowledge, erroneous perception, exogenous actions, and failed actions).

A crucial aspect of the reasoning about change approaches in Artificial Intelligence is that they assign a prominent role to *inertia*: by default, the system tends to remain static, and the changes other than those which are directly caused by action occurrences are rare, this is why one seeks to minimize them. This assumption is crucial if one wants to reason about action in presence of uncertainty without losing too much information. Very often, reasoning about change amounts to *minimizing change*; we will come back on this subject when we will approach the languages for reasoning about action. Indeed, according to the way actions are represented, there exist numerous ways of carrying out the progression of a state or the regression of a formula (encoding a set of states) by an action. Concerning the temporal or dynamic logic representations, progression and regression can be computed via some formula transformations (in particular, conjunction and forgetting). The use of change minimization principles is often proposed as a means to solve the frame problem (cf. Sect. 3.1), but it seems henceforth admitted that it is rather necessary to set up processes which remove the solutions containing abnormal changes (not caused by actions) than processes which minimize them.

This idea to focus on abnormal changes rather than on maximising inertia is well in accordance with the approaches that reason on a world under continuous change where the agent should adapt its action model when a surprise occurs (discrepancy between what is observed and what was expected). This kind of research is more related to the domain of planning in a dynamic word and particularly in the context of goal driven autonomy agents (GDA) that must reason about partially observable domains with a partial knowledge about available actions (Molineaux and Aha 2014; Dannenhauer et al. 2016; Dannenhauer and Cox 2017).

3 Reasoning About Action: Languages

3.1 Problems Related to the Representation of Actions

In the majority of real problems, the system is naturally described by some variables, called *state variables*, that represent some features about some objects, etc. In this case, a state of the system corresponds to the description of a value for each one of the state variables, these values being able to change with the course of time. State variables are usually called *fluents*, a fluent describes a dynamic property of the system. Obviously, the number of possible states is exponential in the number of variables. The *explicit* description of the effects of the actions, which consists in specifying in extenso the functions R_α, becomes then unfeasible in practice, and is somewhat unnatural, because the user is obliged to describe the actions state by state. Similar considerations apply to the description of the correlations between states and observations and with the computation of the operations of progression, regression, etc.

However, there often exist much more economic and natural ways to represent the effects of the actions. For example, let us consider the action to "flip a switch" which causes the alternative "lighting on" or "off" of a bulb. If the representation of the problem requires to take into account the "on"/"off" states of 10 bulbs, then 1024 states of the system will have to be considered (all possible configurations of the 10 bulbs) in order to describe one flip action, whereas this action only causes the change of state of one particular bulb. To describe such action, one rather wants to be limited to indicate that it changes the state of this bulb and, implicitly, that it leaves the other bulbs in their current state.

Action languages were built precisely to this aim: obtaining representations of the effects of the actions which are both more economic (or more compact) and more natural. The problem of preventing the user from explicitly describing the fluents that are not modified by an action in the various possible contexts is known as *the frame problem* (McCarthy and Hayes 1969). It is indeed a problem involved in the choice of a representation of the actions (and not of a modeling problem, i.e. the problem does not rely on the choices of the fluents used to model the system but on the coding of actions in general).

In the same vein, one may face a problem that is the dual of the frame problem, known as *ramification problem* (Finger 1987) which is solved when the action language makes it possible to avoid describing explicitly all the fluents that an action modifies, directly or indirectly, in the various possible contexts. Following up on the previous example, each flip of the switch causes the lighting on/off of the associated bulb, then the room where the bulb is located becomes enlightened and consequently one can settle there for reading. This derived fact is a consequence of the execution of the action but it is not natural, when the action is described, to specify it directly: it results rather from a (static) law which expresses that when a room is lit, one can practice the reading there.

When one deals with action representation, the *qualification* problem (McCarthy 1977) is also often evoked; this problem expresses the incapacity to describe all the pre-conditions that guarantee to obtain the "normal" effect of an action. To deal with this problem, it is first necessary to circumscribe the world with the individuals and the properties explicitly present in the representation; for example, flipping the switch when the associated bulb is off will cause the lighting of this one only if the conflict between Bordures and Syldaves did not cause the destruction of the electric line feeding the house. From our point of view, this problem is not intrinsic within the action representation, it occurs more primarily as soon as the modeling phase starts and simply reflects the difference existing between a situation of the physical world and a representation of this one, which necessarily abstracts it. However, in order to give the pre-conditions of an action, this restriction to the situations that have a representation in the language does not remove the need for reviewing all the situations in which the action is carried out normally. Solving the qualification problem means being able to state the "natural" pre-conditions of an action without having to describe explicitly the list of all the values of the fluents which allow the action to normally take place.

Once actions are represented, it is necessary to build algorithms allowing the computation of the basic operations (progression, regression, etc). The choice of an action language thus depends, on the one hand, of its more or less natural aspect, on the other hand, of its compactness (or space efficiency), and finally, of the complexity of the basic operations when the actions are represented in this language (its computational efficiency).

There exist many action languages which were developed and studied by the community. They can be gathered in several families, according to the nature of the mathematical objects that they use (propositional or first order logic formulas, temporal or dynamic logic formulas, Bayesian networks, state automata, etc.). Giving an exhaustive panorama would be too long and little digest. We will thus only sketchily present the languages which received the most attention from the community, and which are sufficiently representative of the range of the existing languages. Each following sub-section approaches a particular language, or a family of languages, by briefly giving its specificities.

3.2 The Situation Calculus

From an historic perspective the *situation calculus* introduced by McCarthy and Hayes (1969) is the first formalism devoted to reasoning about actions. The definitions given by these authors enabled them to set the basic concepts (presented higher) on reasoning about change and action. The situation calculus is a typed first order logic language with equality, whose types are fluents, states (called situations), actions and objects. In order to simplify the presentation, here we only consider propositional fluents, which have one situation as single argument; we do not mention the objects of the world. Thus, $\neg P(S_0)$ express that the fluent P is false in the situation S_0. S_0

denotes the state of the system at the initial time point of the horizon. For situations and actions we need both variables (denoted respectively $s, \ldots$ and $x, \ldots$) and constants (denoted respectively $S, \ldots$ and $A, \ldots$). The function do applies to a situation and an action and returns a situation. Thus, the formula $\neg P(S_0) \wedge P(do(A_1, S_0))$ expresses that P is false in S_0 and true in $do(A_1, S_0)$, i.e. in the situation obtained by applying the action A_1 in S_0. The formula $\forall s \neg P(s)$ expresses that P is always false. The formula $\forall s((\forall x \neg P(do(x, s)) \leftrightarrow x = A_0)$ expresses that A_0 is the single action which guarantees to make P false in any state where it is applied.

McCarthy and Hayes set a general representation framework enabling them to represent actions by their pre-conditions and their effects (represented by logical formulas). Many approaches were then proposed in order to characterize the "good" consequences of these formulas. Initially, all the authors bet on *change minimization* in order to restrict the set of models so that the properties resulting from the inertia principle can be deduced without having to mention them explicitly. This was accomplished thanks to a second order logic formula, and various circumscription policies were studied for this purpose (the reader can refer to (Moinard 2000) for a review). McCarthy (1986) and then Hanks and McDermott (1986) used the circumscription of abnormality predicates (by considering that a fluent must persist unless otherwise explicitly indicated) within the framework of the situation calculus. However, there are some examples where circumscription does not give the expected result. One of most famous is the *Yale Shooting Problem* proposed by Hanks and McDermott: someone is alive in the initial situation, and one carry out successively the three actions "Load", "Wait" then "Shoot". The action "Shoot" is described by the formula: $\forall s$, (loaded $(s) \rightarrow$ (Abnormal (Alive, Shoot, $s) \wedge \neg$Alive (do(Shoot, s)))).[5] The fact that, by default, the fluents are persisting is described by the second order logic formula $\forall f, s, a, ((f(s) \wedge \neg$Abnormal $(f, a, s) \rightarrow f$ (do$(a, s))$.[6] The circumscription of the predicate *Abnormal* makes it possible to obtain a logical model in which the person is alive at the initial time point and dead (non alive) after the action "Shoot". However, another model is possible: the one where the rifle unloaded itself during "Waiting" and the person is still alive after "Shoot". Circumscribing the *Abnormal* predicate does not allow for preferring the first model to the second one because the two models have incomparable sets of abnormalities w.r.t. set inclusion (in the first model, it is "Alive" which is abnormal in the presence of the action "Shoot"; in the second one, it is "Loaded" which is abnormal w.r.t. "Wait"). Chronological ignorance, proposed by Shoham (1988) and consisting in preferring models where the changes occur the latest, allows one to obtain a satisfactory answer for this example. But this last ad hoc approach was challenged by other examples that it handles badly (Sandewall 1995; Friedman and Halpern 1994).

[5] If the rifle is loaded in the situation s then the fluent "Alive" is abnormal (i.e., non persistent) when the action "Shoot" takes place in s and the person will not be alive any more in the resulting situation.

[6] If the fluent is not abnormal with respect to an action then it keeps its value after the execution of this action.

Another solution suggested by Lifschitz and Rabinov (1989) is to impose that all the fluents that are modified by an action are systematically non inert when this action is carried out. This idea is close to the solution, proposed by Castilho et al. (1999), to use a dependence relation between an action and the atoms on which it may act. The reader can refer to Sandewall (1995) for an excellent synthesis of all these works.

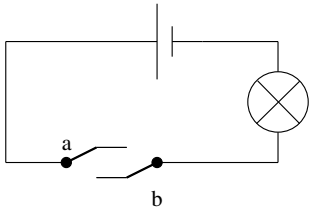
In short, approaches based on change minimization are based on *non-monotonic* logics and are very complex; they are not able to deduce all the intuitive consequences that are expected from a description of a set of actions and an initial situation.

The situation changed with the publication of what was called *Reiter's solution* to the frame problem (Reiter 1991). Reiter suggests a *monotonic* solution based on *successor state axioms* (SSA). These axioms must be given for each fluent P (which is equivalent to an assumption of complete information about the conditions of change of truth value of a fluent) and they have the following form:

$$\forall s, x \quad (P(do(x, s)) \leftrightarrow \gamma_P(x, s))$$

where $\gamma_P(x, s)$ is a formula which does not contain the function symbol do and which can only contain S_0 as situation constant. Thus, the SSA for P describes the conditions under which P is true after an action has been performed, in function of what was true before.

Let us consider the Toggle-switch example (Lifschitz 1990):
In a room, the light is on only if both switches are up or both down. Initially, the switch a is up and the switch b in down, the light is thus off, someone toggles the switch a.



The fluents are U_a ("the switch a is up") and U_b ("the switch b is up"). In this example, the SSA for fluent U_a can be written:

$$\forall s, x \quad U_a(do(x, s)) \leftrightarrow ((\neg U_a(s) \wedge x = T_a) \vee (U_a(s) \wedge x \neq T_a))$$

where T_a is the action to toggle the switch a, i.e., flip its position.

Reiter explains that using Successor State Axioms is a solution to the frame problem because one can reasonably expect the size of the set of SSA to be in the order of the number of fluents (which contrasts with the size of the explicit description of the frame axioms that would be in the order of the number of fluents set multiplied by the number of actions).

According to Reiter, quantification over actions is the key solution to the frame problem. As we will show in Sect. 3.4, the assumption of complete information about the conditions under which fluents change their truth value (translated by the $\leftrightarrow$ in the SSA) allows Reiter to deal with the frame problem in a satisfactory way.

The presence of an SSA for each fluent allows *for regressing* formulas: atoms of the form $P(do(\alpha, \sigma))$ (where α and σ are terms built with variables, constants and the function do) are replaced by the right member of the SSA for P, by applying first the suitable substitution; this process is reiterated until complete elimination of the function do. By construction, the formula thus obtained only relates on the initial state S_0.

For example, the formula

$$U_a(do(T_a, do(T_a, S_0)))$$

is first replaced by:

$$(\neg U_a(do(T_a, S_0)) \wedge T_a{=}T_a) \vee (U_a(do(T_a, S_0)) \wedge T_a{\neq}T_a)$$

which can be simplified into $\neg U_a(do(T_a, S_0))$. In a second step, this last formula is replaced by $\neg(\neg U_a(S_0) \wedge T_a{=}T_a) \vee (U_a(S_0) \wedge T_a{\neq}T_a)$ which can be simplified into $U_a(S_0)$. We have thus proven by regression that the switch is up after two executions of T_a if and only if it is up in the initial state S_0.

In order to decide whether the application of the action α in the state S_0 leads to a state in which ψ holds, it is enough to decide if the formula $\phi(S_0) \rightarrow \psi(do(\alpha, S_0))$ is valid. The regression of $\psi(do(\alpha, S_0))$ results in a formula $\psi'(S_0)$. If the argument S_0 is eliminated, we obtain the propositional formula $\phi \rightarrow \psi'$, whose validity can be checked by using a suitable prover.

This solution was combined with epistemic logic (Scherl and Levesque 2003), which gives a formalism close to dynamic logic, described in Sect. 3.4. Moreover the framework of the situation calculus with Successor State Axioms has been recently used by Batusov and Soutchanski (2018) for causal ascription.

3.3 Propositional Action Languages

A weak point of the approaches based on the situation calculus is the difficulty of their algorithmic implementation. For this reason, researchers have also developed approaches based on propositional logic, which can benefit from off-the-shelf ASP or SAT solvers (see chapters "Logic Programming" and "Reasoning with Propositional Logic: from SAT Solvers to Knowledge Compilation" of Volume 2).

In action languages based on propositional logic, action effects are represented by local rules specifying only the fluents that change, possibly together with the conditions under which they change. Let F be a finite set of fluents. The states of $\mathscr{S}$ are the propositional interpretations over F, that is, $\mathscr{S} = 2^F$.

The most basic action language is arguably *STRIPS* (Fikes and Nilsson 1971), where an action is represented by a precondition and its effects (see chapter "Planning in Artificial Intelligence" of Volume 2), a precondition being a conjunction of literals and an effect being a consistent set of literals.

To encode the light switch example, one may take as set of fluents $F = \{U_a, U_b\}$, where U_a (resp. U_b) is true (resp. false) when switch a (resp. b) is on (resp. off). An action with conditional effects such as T_a ('switch a') can be written as

$$(U_a \mapsto \neg U_a) \wedge (\neg U_a \mapsto U_a).$$

The right member l of each rule of form $c \mapsto l$ is a direct action effect, which applies if and only if the corresponding condition is satisfied in the state in which the action is performed. Thus, applying $c \mapsto l$ in state s leaves s unchanged if s does not satisfy c and enforces the truth of l otherwise, leaving other fluents unchanged. This applies to each rule.[7] Thus, applying T_a in state s leads to change the truth value of U_a in s, as we expect. Importantly, such an action description rule is not a classical logical formula, and in particular, $\mapsto$ is not material implication. Indeed, a STRIPS action α can be seen as a *constraint* linking the state of the world *before* it is performed and the state of the world *after* it has been performed. In particular, $c \mapsto l$ is not equivalent to $\neg l \mapsto \neg c$.[8]

One of the limits of the STRIPS language is the impossibility to express static laws. These laws are however needed for the ramification problem to be dealt with. For instance, in the previous example, one may want to introduce a new fluent L expressing that "the light is on". With standard STRIPS, integrating this new fluent would require to modify all actions by specifying what happens to L. This solution is not reasonable when the number of fluents is large. A way to cope with this lack of expressiveness consists in encoding actions with a set of *basic* fluents on which the available actions act directly (U_a and U_b in the example); the fluents that are not basic are called *derived* fluents. Progression is first computed as in classical STRIPS, and then there is one additional step so as to take the static laws into account and make some inferences on derived fluents. Thus, to compute the progression of state s by an action, one starts by projecting s on the basic fluents; then one performs the progression of this projection, and finally the obtained state is completed using the static laws. In the switch example, one may take as static law

$$((U_a \wedge U_b) \vee (\neg U_a \wedge \neg U_b)) \leftrightarrow L$$

[7] A pathological case is when the conditions of rules leading to complementary literals are conjointly satisfied in s; in such a case, the progression is undefined; this can reflect an error when specifying the representation of the action, or the fact that s is impossible (and in this case corresponds to an implicit static law).

[8] If they were equivalent, then the encoding of action "Shoot" by *Loaded* $\mapsto \neg$ *Alive* in the *Yale Shooting Problem* would be equivalent to *Alive* $\mapsto \neg$ *Loaded*, meaning that shooting on a living person results on the gun being magically unloaded (and the person staying alive...).

where U_a and U_b are basic and L is derived. The progression of state $\{U_a, \neg U_b, \neg L\}$ by action T_a is thus $\{\neg U_a, \neg U_b, L\}$.

There are four main problems with STRIPS: it does not allow for representing (a) non-determinism, (b) static causal relations between fluents (as discussed in the previous paragraph), (c) concurrent actions, and (d) epistemic actions. To cope with this lack of expressiveness, more sophisticated action languages have been developed, both in the planning community (with ADL (Pednault 1989) and PDDL (Ghallab et al. 1998)) and in the knowledge representation community. We will now focus on the languages stemming from the latter community.

In the 70s and 80s, the knowledge representation community used to think of actions as simple rules linking action preconditions and action effects. Subsequently, some researchers suggested that prediction could be computed using *minimization of change*, so as to impose that, by default, fluents that are not concerned by the action should persist (these fluents, of course, do not need to be specified in action effects, so as to cope with the frame problem). Then, since the 90 s, minimization of change was progressively replaced by the use of propositional languages based on *causal implication*. The solutions of Reiter (1991), Lifschitz and Rabinov (1989) and Castilho et al. (1999) for solving problems occurring with minimization of change consist in expressing dependencies between an action and its effects. This very principle has been implemented in works using *causal implication* (see chapter "A Glance at Causality Theories for Artificial Intelligence" of this volume), which is distinct from material implication since it is meant to express these dependencies.

Some approaches using causal implication make use of the situation calculus (Stein and Morgenstern 1994; Lin 1995). Others use the modality C (Geffner 1990; Giordano et al. 1998; Turner 1999) or equivalently, define a new connective $\Rightarrow$ (Giunchiglia et al. 2004) Yet others define influence relations between fluents (Thielscher 1997). The main feature of these approaches is that they distinguish the fact of being true from the reason for being true, and use this distinction for computing the expected effect of actions for prediction or planning.

We give now some details about the action language $\mathscr{A}$ proposed by Gelfond and Lifschitz (1993). In this language, an action is expressed by means of *conditional causal rules* of the form

$$\text{if } c \text{ then } \alpha \text{ causes } l,$$

where α is an action name, c a conjunction of literals (omitted when it is equivalent to $\top$), and l a literal. A set of causal rules defines a deterministic transition system between states. Thus, the action α defined by the causal rules

$$\text{if } p \wedge q \text{ then } \alpha \text{ causes } \neg p, \text{ if } \neg p \wedge q \text{ then } \alpha \text{ cause } p \text{ and } \alpha \text{ cause } q$$

corresponds to the transition system R_α defined by $R_\alpha(pq) = R_\alpha(\bar{p}\bar{q}) = \bar{p}q$ and $R_\alpha(\bar{p}q) = R_\alpha(p\bar{q}) = pq$. An action α described by such causal rules corresponds to a *propositional action theory* Σ_α, expressing α by means of propositional symbols F_t and F_{t+1}, with $F_t = \{f_t | f \in F\}$ and $F_{t+1} = \{f_{t+1} | f \in F\}$, where f_t represents

fluent f at time t, that is, before action α has been performed, and f_{t+1} represents f at time $t+1$, after action α has been performed. The causal rules are translated into Σ_α according to the following principle: fluent f is true at $t+1$ if and only if one of these two conditions holds: (a) it was true at t and the state at t does not satisfy any precondition of a causal rule whose conclusion is $\neg f$, or (b) it was false at t and the state at t satisfies the precondition of a causal rule whose conclusion is f. One finds here the principle at work in the situation calculus, which we called 'Reiter's solution' in Sect. 3.2.

Formally, let $\Gamma(f)$ (respectively $\Gamma(\neg f)$) the disjunction of all preconditions of rules whose conclusion is f (respectively $\neg f$); then Σ_α is the conjunction of all the formulas

$$f_{t+1} \leftrightarrow \Gamma(f)_t \vee (f_t \wedge \neg\Gamma(\neg f)_t)$$

for $f \in F$. Thus, the action theory Σ_α corresponding to the action α previously described by its causal rules is

$$\Sigma_\alpha = (p_{t+1} \leftrightarrow ((\neg p_t \wedge q_t) \vee (p_t \wedge \neg(p_t \wedge q_t)))) \wedge (q_{t+1} \leftrightarrow \top),$$

which simplifies into

$$\Sigma_\alpha = q_{t+1} \wedge (p_{t+1} \leftrightarrow (p_t \leftrightarrow \neg q_t)).$$

An extension of language $\mathscr{A}$ is language $\mathscr{C}$ (Giunchiglia and Lifschitz 1998), which allows for expressing executability conditions and static rules, independently of any action, such as

$$\mathsf{Outside} \wedge \neg\mathsf{Umbrella} \wedge \mathsf{Umbrella}\ \mathsf{causes}\ \neg\mathsf{Dry},$$

that are also taken into account in the corresponding action theory. For instance, consider action Go-out with a unique causal rule

$$\mathsf{Go\text{-}out}\ \mathsf{causes}\ \mathsf{Outside};$$

the corresponding action theory, taking into account the previous static rule, is

$$\begin{aligned}\Sigma_{\mathsf{Go\text{-}out}} =& \mathsf{Outside}_{t+1} \wedge (\mathsf{Umbrella}_{t+1} \leftrightarrow \mathsf{Umbrella}_t) \wedge (\mathsf{Rain}_{t+1} \leftrightarrow \mathsf{Rain}_t) \\ & \wedge (\mathsf{Dry}_{t+1} \leftrightarrow \mathsf{Dry}_t \wedge (\mathsf{Umbrella}_t \vee \neg\mathsf{Rain}_t)).\end{aligned}$$

Non-determinism can be expressed in several different ways, explored independently in different papers:

- by *complex effects*, such as

$$\alpha\ \mathsf{causes}(p \leftrightarrow q),$$

a choice that is at the heart of belief update, cf. Sect. 4;
- by *disjunction of effects*, which are similar to nondeterministic union in dynamic logic, cf. Sect. 3.4), such as

$$\mathsf{Toss\ causes\ Heads\ or\ causes\ \neg Heads};$$

- by recursive causal rules, which is a more technical solution that we will not discuss here.

Some action languages (such as language $\mathscr{C}$) also have *concurrency*, whereas others have *epistemic actions*, thus enabling the distinction between facts and knowledge; thus, the action of testing whether the fluent f is true or false is represented by the causal rule

$$\alpha\ \mathsf{causes}\ \mathbf{K}f\ \mathsf{or\ causes}\ \mathbf{K}\neg f,$$

where $\mathbf{K}$ is the knowledge modality of epistemic logic S5 (see in particular (Herzig et al. 2003)).

Progression and *regression* can be applied directly in these languages. A belief state, in the binary uncertainty model, is a nonempty set of states, and can thus be represented by a consistent propositional formula. Progression and regression map a consistent formula and an action to a formula (which is always consistent in the case of progression and weak regression). The progression of formula φ by action α consists first in taking the conjunction of φ_t (expressing that φ is true before the action) and Σ_α, and then in forgetting in $\varphi_t \wedge \Sigma_\alpha$ all variables f_t, i.e., in deriving the strongest logical consequence of $\varphi_t \wedge \Sigma_\alpha$ independent of the variables f_t (see for instance (Lang et al. 2003)). Weak regression is computed similarly: the weak regression of ψ by α is the result of forgetting the variables f_{t+1} in $\psi_{t+1} \wedge \Sigma_\alpha$. The strong regression of ψ by α is obtained by computing the minimal conditions guaranteeing that the application of α will lead to a state satisfying ψ. Thus, in the previous example, the progression of $\mathsf{Dry} \wedge \mathsf{Umbrella}$ by Go-out is (up to logical equivalence)

$$\mathsf{Outside} \wedge \mathsf{Umbrella} \wedge \mathsf{Dry},$$

and the progression of $\neg\mathsf{Umbrella}$ by Go-out is

$$\mathsf{Outside} \wedge \neg\mathsf{Umbrella} \wedge (\mathsf{Rain} \rightarrow \neg\mathsf{Dry}),$$

whereas the weak regression of $\mathsf{Dry} \wedge \mathsf{Rain}$ by Go-out is

$$\mathsf{Umbrella} \wedge \mathsf{Rain} \wedge \mathsf{Dry}.$$

3.4 Dynamic Logic

There are other possible ways of representing actions and dealing with the corresponding problems. *Dynamic logic* is a formalism initially known in theoretical computer science for reasoning about program execution. In addition to Boolean operators, its language contains *modal operators* of the form $[\alpha]$, where α is a program. The combination of such an operator with a formula results in a formula of the form $[\alpha]\phi$, read 'ϕ is true after every execution of α'. Instead of a program, one may assume that α is an event or an action. For instance, the action of toggling switch a can be described by the two effect laws $(\neg U_a \rightarrow [T_a]U_a)$ and $(U_a \rightarrow [T_a]\neg U_a)$.

In the context of dynamic logic, an important aspect of reasoning about actions that was dealt with first in (Herzig and Varzinczak 2007) concerns the consistency of a domain description. It has been shown that for expressive action languages, beyond logical consistency, a good domain description should be modular, in the sense that effect laws describing the actions should not allow for deriving new static laws. For instance, the static laws $P_1 \rightarrow [A]Q$, $P_2 \rightarrow [A]\neg Q$ and $\neg[A]\bot$ together imply the static law $\neg(P_1 \wedge P_2)$; if this law is not deductible from the other static laws and only them, then these effect laws should be considered problematic.

Unlike in situation calculus, states are not explicit in dynamic logic. Although dynamic logic does not allow either for quantifying over actions, which is a key feature of Reiter's solution for the frame problem, it has been shown that this solution can be implemented in dynamic logic for the rather general case of explicit SSAs (van Ditmarsch et al. 2011). In such SSAs, x must be the only action variable of $\gamma_P(x, s)$ and if an action constant A does not appear in $\gamma_P(x, s)$ then $\gamma_P(A, s)$ should not be equivalent to $P(s)$. These conditions are natural for a system satisfying inertia. An example of SSA not satisfying them would be $\forall s(\forall x P(do(x, s))) \leftrightarrow \neg P(s))$, which means that P is changed in every state (thus P is a non-inert fluent). Note that the formula $\gamma_{U_a}(x, s)$ in our example from Sect. 3.2 satisfies these conditions. In order to translate these SSAs in dynamic logic, one introduces *assignment actions* of the form $P := \phi$; such an assignment describes an action where P takes the truth value that ϕ had in the previous state. This allows for associating with each action constant A the following set of assignments:

$$\sigma_{SSA}(A) \;=\; \{P := simp(\gamma_P(A)) \;\mid\; P \text{ appears in } \gamma_P(x)\}$$

where $simp(\gamma_P(A))$ is obtained from $\gamma_P(x)$ by eliminating argument s, substituting x by A and simplifying the equalities. In our example from Sect. 3.2, after substituting x by T_a we obtain:

$$\sigma_{SSA}(T_a) \;=\; \{U_a := (\neg U_a \wedge T_a{=}T_a) \vee (U_a \wedge T_a{\neq}T_a)\}$$

which can then be simplified into

$$\sigma_{SSA}(T_a) \;=\; \{U_a := \neg U_a\}.$$

Each occurrence of an abstract action symbol A is replaced by the corresponding assignment. As shown in (van Ditmarsch et al. 2011), this constitutes a solution (in Reiter's sense) to the frame problem. Thus Reiter's solution is transferred to dynamic logic, without any need to quantify over actions. It is also shown that Reiter's solution can be combined with epistemic logic, thus bridging it with epistemico-dynamic logics (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" of this volume).

A recent work in dynamic logic is to investigate epistemic extensions that are suitable for conformant planning (Li et al. 2017) and more generally for multiagent epistemic planning (Aucher and Bolander 2013; Bolander et al. 2015; Cooper et al. 2016). An overview paper about combinations of logics of action with logics of knowledge and belief is (Herzig 2015).

3.5 Dynamic Bayesian Networks

A *dynamic Bayesian network* (Dean and Kanazawa 1989) is a Bayesian network (see chapter "Belief Graphical Models for Uncertainty Representation and Reasoning" of Volume 2) in which the variables exist in as many copies as there are time points: for any fluent f and time step t there is a fluent f_t. For each time step t, there exists a Bayesian network linking the variables corresponding to t. Moreover, between these 'instantaneous' networks, the only allowed edges are those that are directed from past to future. The temporal directed acyclic graph (DAG) given on Fig. 1, equipped with probabilities for each variable, at each time step, conditionally on the values of the parents of the variable, constitutes a dynamic Bayesian network.

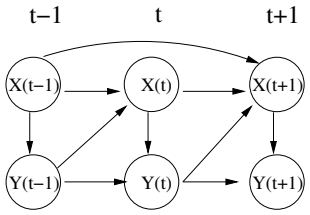


Fig. 1 The DAG of a dynamic Bayesian network

If the system is Markovian, in order to determine completely the behavior of the system it is enough to know the probability distribution for x_t and the conditional probability distribution for x_{t+1} given x_t. The Markovian assumption can reasonably be made for many classes of systems. A Markovian temporal DAG cannot admit an arc linking variables distant from more than one time step: by deleting the edge between x_{t-1} and x_{t+1} the diagram on the example below becomes Markovian, and a description restricted to time steps t and $t + 1$ suffices.

The truth value of a fluent f at a given time step can depend on its value at earlier time steps $t - \Delta$, which translates in probabilistic terms into the following "survival equation":

$$Pr(f_t) = Pr(f_t \mid f_{t-\Delta}).Pr(f_{t-\Delta}) + Pr(f_t \mid \neg f_{t-\Delta}).Pr(\neg f_{t-\Delta}).$$

The conditional probability $Pr(f_t \mid f_{t-\Delta})$ is called *survival function*. The survival function represents the tendency of propositions to persist given all events that can make them false. A classical survival function is: $Pr(f_t \mid f_{t-\Delta}) = \exp^{-\lambda.\Delta}$, which indicates that the probability that f persists decreases, from the last time step where f was observed to hold, at an exponential speed determined by λ.

If one has some information about events that can affect the truth value of the fluent, then the survival equation no longer fits. Generally, the probability that a proposition f is true in t is a function of:

- the probability $Pr(f_{t-\Delta})$ that it is true at $t - \Delta$
- the probability $Pr(do(f_t))$ of the occurrence of an event that makes f true at t
- and the probability $Pr(do(\neg f_t))$ of the occurrence of an event that makes f false at t.

From the standpoint of expressiveness, the interest of this class of probabilistic approaches for reasoning about change is that it allows for expressing numerical uncertainty on beliefs, observations, and causal laws (see also (Hanks and McDermott 1994) and by (Pearl 1988)). On the other hand, a problem is that it requires the specification of many prior probabilities, even if it is not always necessary to 'solve' the whole probabilistic network to determine the probability of a proposition: one can instead focus on a few key time steps (and key propositions). Note that the use of a dynamic possibilistic network allows one to reason without knowing precisely these probabilities (Heni et al. 2007).

4 Reasoning About Change: Update

Update is a research domain at the intersection of reasoning about actions (whence its presence in this chapter) and belief change (see chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume). It is a process for integrating into a belief base a modification of the state of the system that is explicitly specified by a propositional formula. More precisely, given a belief base K and a propositional formula α, the update of K by α is the progression of K by an action, or alternatively by an exogenous event whose occurrence is known and whose effect is α. Belief update is opposed to belief revision where a new piece of information about the system is integrated into a belief base about that system, under the hypothesis that the latter did not evolve. The distinction was clarified by Katsuno and Mendelzon (1991) and Winslett (1988), although update was studied before (Keller and Winslett

1985), partially by scholars from the database community (see chapter "Databases and Artificial Intelligence" of Volume 3).

The distinction between revision and update deserves to be clarified a bit more here. If the new piece of information ('the input') completes our beliefs about the world then it is not the world that has evolved, but only the agent's beliefs about the world. (This may be due to the questioning of an erroneous information about the world or a new piece of information about the characteristics of the world.) In that case we have to perform a *revision*.[9] Such a revision amounts to a simple addition (also called *expansion*) when the input is consistent with the beliefs; however, in case of inconsistency revision selects some beliefs that have to be rejected in order to restore consistency (see chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume). If the input characterizes an explicit evolution of the world (i.e., is the effect of an action or an exogenous event) then we speak of an *update*. The updated belief base describes the world after its evolution. The update therefore corresponds to a *progression*.

The difference can be illustrated by the following example (Morreau 1992). Suppose there is a basket containing either an apple or a banana. If we learn that in fact it does not contain bananas then our beliefs have to be revised and we deduce that the basket contains an apple. However, if we learn that the world has evolved in a way such that there is no banana in the basket any more (e.g. because somebody has performed the action of taking the banana out of the basket if it was there) then we have to update our beliefs, i.e., that now the basket is either empty or contains an apple.

Just as for revision, there does not exist a unique update operator that would suit all applications. It is therefore interesting to define criteria which determine which of these operators are 'rational', these criteria can be written under the form of rationality postulates. Paralleling Alchourrón, Gärdenfors and Makinson's postulates characterizing 'rational' revision operators (the so-called AGM postulates) (Alchourrón et al. 1985), Winslett (1990) was the first to define postulates for update operators. These postulates inspired Katsuno and Mendelzon (1991) who defined a new set of postulates. Similarly to the AGM postulates (see chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume), they are related to the existence of a set of preorder relations about the states of the system, where with each state there is associated a preorder. In (Katsuno and Mendelzon 1991), the authors implement an idea that had been put forward by Winslett: in order to update a belief base one may update each of the models of the base independently. Katsuno and Mendelzon's contribution is the idea that each model has to be updated towards the 'closest' models (in the sense of the preorder associated with the original model). According to Katsuno and Mendelzon, an update operator is a function

[9] As shown in (Friedman and Halpern 1999), revision remains relevant even if the initial belief state and the new formula do not refer to the same time point, as long as there is a syntactical distinction (via some time-stamping) between a fluent at a time point and the same fluent at another time point: what matters for revision is not that the world is static, but that the propositions that are used to describe the world are static. This also explains that belief extrapolation also corresponds to a revision process (Dupin de Saint-Cyr and Lang 2011).

$\diamond$ which given a formula $\mathcal{K}$ representing beliefs about the world and a formula ϕ representing the information about the evolution of the world, returns a new formula $\mathcal{K} \diamond \phi$. The postulates they propose in order to characterize the 'rational' operators $\diamond$ are the following:

U1 $\mathcal{K} \diamond \varphi$ implies φ.
U2 If $\mathcal{K}$ implies φ then $(\mathcal{K} \diamond \varphi)$ is equivalent to $\mathcal{K}$.
U3 If $\mathcal{K}$ and φ are satisfiable then $\mathcal{K} \diamond \varphi$ is satisfiable.
U4 If $\mathcal{K}_1$ is equivalent to $\mathcal{K}_2$ and φ_1 is equivalent to φ_2 then $\mathcal{K}_1 \diamond \varphi_1$ is equivalent to $\mathcal{K}_2 \diamond \varphi_2$.
U5 $(\mathcal{K} \diamond \varphi) \wedge \psi$ implies $\mathcal{K} \diamond (\varphi \wedge \psi)$.
U6 If $\mathcal{K} \diamond \varphi_1$ implies φ_2 and $\mathcal{K} \diamond \varphi_2$ implies φ_1 then $\mathcal{K} \diamond \varphi_1$ is equivalent to $\mathcal{K} \diamond \varphi_2$.
U7 If $\mathcal{K}$ is complete then $(\mathcal{K} \diamond \varphi_1) \wedge (\mathcal{K} \diamond \varphi_2)$ implies $\mathcal{K} \diamond (\varphi_1 \vee \varphi_2)$.
U8 $(\mathcal{K}_1 \vee \mathcal{K}_2) \diamond \varphi$ is equivalent to $(\mathcal{K}_1 \diamond \varphi) \vee (\mathcal{K}_2 \diamond \varphi)$.
U9 If $\mathcal{K}$ is complete and $(\mathcal{K} \diamond \varphi_1) \wedge \varphi_2$ is satisfiable then $\mathcal{K} \diamond (\varphi_1 \wedge \varphi_2)$ implies $(\mathcal{K} \diamond \varphi_1) \wedge \varphi_2$.

U1 stipulates that ϕ is a piece of information describing the world after its evolution (this is one of Winslett's postulates). U2 says that if ϕ was already true in all states of the system before the update then the system does not evolve. This is the postulate requiring that inertia has always to be preferred to spontaneous evolution. It is however not always desirable because it forbids the existence of transitory states, i.e., states within which the system may not stay because it immediately evolves towards other states. U3 expresses that a consistent representation of the system and of its evolution can always be updated in a consistent way (which was also one of Winslett's postulates). However, systems may exist where there is no transition between two states: for example, when $\varphi = dead$ and $\alpha = alive$ then one may wish the update to fail. So this postulate is not always desirable. U8 (also one of Winslett's postulates) means that the update is defined as a progression operator. U9 is a restriction of the converse of U5. We refer to (Dubois et al. 1995; Herzig and Rifi 1999) for more detailed critiques of these postulates.

The following representation theorem relates these postulates to the existence of a set of preorder relations between states of the world:

Theorem 1 (Katsuno, Mendelzon) $\diamond$ *satisfies U1, U2, U3, U4, U5, U8, U9*[10] *if and only if for every* $\omega \in \Omega$ *there is a total preorder* $\leq_\omega$ *such that*

(1) $\forall \omega' \in \Omega, \quad \omega <_\omega \omega'$ *(*$\leq_\omega$ *is "faithful");*
(2) $Mod(\mathcal{K} \diamond \varphi) = \bigcup_{\omega \models \mathcal{K}} \{\omega' \models \varphi$ *such that* $\forall \omega'' \models \varphi, \omega' \leq_\omega \omega''\}$.

Item (1) means that for each model ω of $\mathcal{K}$, the models of the update of ω by φ are the models of φ that are closest to ω w.r.t. $\leq_\omega$, and (2) means that the set of models of the update of $\mathcal{K}$ by φ is the union of the sets of models resulting from the update of each model of $\mathcal{K}$ by φ (which follows directly from postulate U8).

[10] If U6 and U7 are used instead of U9 then the theorem gives us a faithful preorder that is only partial.

Numerous update operators were proposed in the literature. Thanks to the above theorem they can be defined by associating with each state a faithful total preorder relation between states. In practice, such a set of preorders is a way of *minimizing change*. For example, Winslett defined a relation $\leq_\omega^{\text{PMA}}$ between states that she called 'Possible Models Approach', abbreviated PMA. It is based on the function $\text{diff}_{\text{PMA}}(\omega_1, \omega_2)$ (the set of variables whose value differs between the two states ω_1 and ω_2): $\omega_1 \leq_\omega^{\text{PMA}} \omega_2 \quad \Leftrightarrow_{def} \quad \text{diff}_{\text{PMA}}(\omega_1, \omega) \subseteq \text{diff}_{\text{PMA}}(\omega_2, \omega)$. This relation is faithful and therefore defines an update operator. The corresponding update operator can also be characterized in terms of independence (the logical consequences of $\mathscr{K}$ that are independent of φ persist) (Marquis 1994). In the examples of Morreau (1992), the initial beliefs are $\mathscr{K} = (banana \wedge \neg apple) \vee (apple \wedge \neg banana)$, so there are two models $\omega_1 = \{\neg apple, banana\}$ and $\omega_2 = \{apple, \neg banana\}$. When the agent then learns (φ) that somebody took the banana if it was there (update by $\neg banana$), the states of the system representing the information φ are ω_2 and $\omega_3 = \{\neg apple, \neg banana\}$. The updated base $\mathscr{K} \diamond_{\text{PMA}} \varphi$ can be computed by taking the union, for all models ω of $\mathscr{K}$, of the models of φ that are closest to ω. Here, the model of φ that is closest to ω_1 for the relation $\leq_\omega^{\text{PMA}}$ is ω_3; the model of φ closest to ω_2 is ω_2 itself (because $\leq_\omega^{\text{PMA}}$ is faithful). So the set of models of $\mathscr{K} \diamond_{\text{PMA}} A$ is $\{\omega_2, \omega_3\}$. This means that after the update there is either an apple in the basket or the basket is empty.

The PMA relation has been refined by assigning priorities to some fluents (Winslett 1988), which allows for handling fluents that do not persist in the same way. Other update operators go for increased expressiveness, e.g. the one proposed by Cordier and Siegel (1995) which allows for more or less prioritary transition constraints. These constraints take the form of pairs of formulas (φ, ψ) and are satisfied by a pair of models (ω, ω') when ω satisfies φ and ω' satisfies ψ. Then ω' is considered closer to ω than ω'' if the transition (ω, ω') violates less prioritary constraints than the transition (ω, ω'').

Updates *à la* Katsuno and Mendelzon (and in particular Winslett's PMA (Winslett 1988) but also Forbus' operator (Forbus 1989)) are built on minimization of change. However, minimization of change is not always desirable for updates. In particular, Herzig and Rifi (1999) have shown that the approaches building on minimization of change do not allow updates by disjunctions; more formally, an update operator satisfying the Katsuno-Mendelzon postulates cannot handle disjunctions correctly, the culprit being postulate U5.

For that reason, several scholars studied update operators that are not built on minimization. They in particular studied a family of update operators that is based on the concept of *dependence*. Such updates of a belief base β by a formula α consists in first forgetting in β "all information concerning α" (leaving the truth values of the variables that are not concerned by the update unchanged), and then adding α to the result. It remains to work out what "all information concerning α" means. Such a kind of relevance is induced by a dependence relation between formulas: α concerns β if and only if β depends on α. This approach is general because the notion of dependence between formulas can vary.

Most of the dependence-based approaches to update consider that the dependence relation is expressed first between formulas and propositional variables, and can then be extended to a dependence relation between formulas: α and β are dependent if and only if there is at least one variable on which α and β are dependent. Examples of such update operators can be found in (Herzig 1996; Doherty et al. 1998), (Herzig and Rifi 1999). This principle allows for remediating several counterintuitive aspects of minimization-based approaches and moreover is generally of lower computational complexity. A slight drawback is however that it is too little conservative: too much information of the initial base is forgotten. This can be counterbalanced by replacing the dependence relation between formulas and variables by a dependence relation between formulas and literals (Herzig et al. 2013).

As an update by a formula α can be viewed as a progression by a particular action whose effect is α ("to make α true") (see a discussion in (Lang 2007)), it makes sense to situate update w.r.t. propositional action languages. We start by observing that STRIPS is a particular case of both formalisms, corresponding to an update by conjunctions of literals. Axiom U8—which requires that the update of a set of models is the union of the update of the individual models—is exactly the definition of the progression of a belief state by an action. The two paradigms however differ in the variety of available actions: on the one hand, update offers the possibility of taking into account disjunctive effects (representing a unique but imperfectly known effect) and more generally effects consisting of arbitrary propositional formulas. On the other hand, action languages allow for conditional effects such as

$$(\textsf{if Heads then flip-coin causes} \neg\text{Tails}, \textsf{if} \neg\text{Heads then flip-coin causes Tails}),$$

nondeterministic effects such as

$$\text{Toss-coin causes Heads or causes} \neg\text{Heads},$$

concurrent effects such as

$$\begin{array}{c}\text{if } G \text{ and } D \text{ are actions consisting in lifting the left and the right side of a table and}\\ \text{if a glass of water is on the table then}\\ G \text{ causes Spilled},\ D \text{ causes Spilled},\ G \text{ concurrently with } D \text{ cause } \top\end{array}$$

as well as static causal rules allowing for ramifications. Approaches aiming at unifying the potentialities of various approaches are not numerous. Some update approaches take ramification into account by resorting to integrity constraints (Doherty et al. 1998) or allow for nondeterministic updates (Brewka and Hertzberg 1993), or conditional or concurrent updates (Herzig et al. 2001). However, an embedding of Winslett's and Forbus' update operator and of Dalal's revision operator into dynamic logic was recently provided in (Herzig 2014).

Update corresponding to the progression of an action, there exists a generalization (rightly called *generalized update*) that enables both revision and event abduction

(Boutilier 1998). Generalized update allows for example for handling the following scenario: an agent wakes up in the morning and believes that the lawn is dry just as it was when she went to sleep. She subsequently observes that the lawn is wet, which first of all leads to a revision of her beliefs, then to the abduction of an event (it rained), and finally to an update by the effects of the event (the road is wet, too). Belief extrapolation (Dupin de Saint-Cyr and Lang 2011) and other related formalisms such as (Berger et al. 1999) only handle the abduction of events. Finally, update can be viewed as an ordinal form of Lewis's *imaging* operator (Dubois and Prade 1993) as well as the predictive phase of the Kalman filter (Cossart and Tessier 1999; Benferhat et al. 2000).

Goldszmidt and Pearl (1992) were also interested by accounting for revision and update at the same time. They reason about a set of default rules that are of the causal kind, from which they deduce an order on the pairs of states of the world that they are filtering according to the input. If the last operator is a revision by φ then the pairs of states where the final state satisfies φ see their plausibility increased. In the case of an update by φ, one has to perform a revision by the dummy action $do(\varphi)$.

The contributions of Winslett, and of Katsuno and Mendelzon, are important for two reasons. First of all, they established a clear distinction between revising and updating a belief base. Second, they elaborated a set of postulates guaranteeing that a rational update is related to the existence of a set of preorders between the possible states of the system. Katsuno and Mendelzon, and Winslett, implicitly opted for the particular case where the fluents are by default inertial (they only change if an action or an event occurs that changes them). There is a further implicit hypothesis embodied by postulate U3 (Winslett's MB4): asserting that any update can be performed means that the input is always consistent with the possible evolution of the world.

Recently, belief change (including belief update) within the framework of fragments of propositional logic has gained attention. A propositional fragment simply is a subset of a propositional language which has some valuable properties (typically, from the computational side) but is not fully expressive w.r.t. propositional logic (some propositional formulas do not have any equivalent representation in the fragment). For instance, the Horn CNF fragment is the set of CNF formulas where each clause is Horn, i.e., it contains at most one positive literal. It is well-known that the satisfiability of any Horn CNF formula can be decided in linear time but that some propositional formulas (e.g. the clause $a \vee b$) cannot be turned into equivalent Horn CNF ones. Other fragments which are often considered are the Krom one (the set of all CNF formulas where each clause is binary) and the affine fragment (the set of all conjunctions of exclusive-or clauses), and each of them offers the same tractability property as the Horn one w.r.t. the satisfiability issue and the same limitations as to expressiveness. In order to preserve the tractability benefits, when a belief base from a given fragment has to be updated, it is expected that the updated base belongs to the same fragment. However, update operators satisfying all the Katsuno-Mendelzon postulates (especially Winslett's PMA and Forbus' operator) do not ensure this property. This calls for a notion of refinement of an update operator for a given fragment, which warrants that the result of any update is in the fragment when the initial base is

in the fragment as well. Any refined operator is required to approximate the behavior of the operator considered at start (especially, leading to the same updated base as it when this base fits in the fragment), the price to be paid being the loss of some rationality postulates. A constructive approach to such refinements of update operators has been introduced in (Creignou et al. 2015), and the Katsuno-Mendelzon postulates satisfied by the refined operators identified as well.

Several tentatives were also made to extend update to more expressive frameworks:

- *ASP*: Slota and Leite (2010) adapt Katsuno & Mendelzon's postulates to logic programs. They also define update operators for hybrid belief bases (Slota et al. 2011; Slota 2012). Such hybrid bases are made up of an ontology component, expressed in the language of the description logic ALCIO (ALC with inverse and nominals), and a rule component, expressed in the language of answer-set programming under the stable semantics (see chapter "Logic Programming" of Volume 2). Update operators are studied in particular for the strong equivalence semantics for ASP as well as for hybrid belief bases (Motik and Rosati 2010).
- *Belief states*: Lang et al. (2001) define update operators over epistemic states. In addition to beliefs, such epistemic states, represented by orders on worlds, allow for expressing the relative plausibility of beliefs. The authors extend the class of dependence-based update operators to epistemic states. Baral and Zhang (2005) generalize update so as to distinguish between facts and knowledge, as in the epistemic logic S5. Their process is called *knowledge update* and allows one to account for the effects of epistemic actions by *updating* the epistemic formulas describing the agent's beliefs. Such a framework takes the viewpoint of a modeler agent O who reasons about the belief state of another agent ag. For example, the update of an S5 model by $K_{ag}\varphi$ means that O updates her beliefs about ag's beliefs; the mental state of ag is seen by O as part of the external world, and the update by $K_{ag}\varphi$ corresponds to an action whose effect is to make $K_{ag}\varphi$ true (for example, the action of telling ag that φ is true).
- *Description logics* (see chapter "Reasoning with Ontologies" of this volume): Liu et al. (2011) update assertions of an "ABox" (that is, the factual component of the belief base). They highlight an expressiveness problem that arises in that framework: sometimes the expected result of an update cannot be encoded by an assertion of the basic description logic ALC. For example, the assertion

  $$\texttt{mary} : \texttt{Person} \sqcap \exists \texttt{child\_of}.\texttt{Person} \sqcap \forall \texttt{child\_of}.(\texttt{Person} \sqcap \texttt{Happy})$$

  expresses that every child of Mary is happy. If one updates the ABox containing this information by the fact that Peter becomes unhappy, i.e., by the assertion $\texttt{peter} : \texttt{Person} \sqcap \neg \texttt{Happy}$, then one has to take two possible cases into account: the case where Peter is among Mary's children and the case where he is not. Intuitively, the principle of minimal change requires that the result of the update is on the one hand the new piece of information (Peter is unhappy) and on the other hand the fact that every child of Mary either has the property of being happy or has the

property of being Peter. The latter assertion (i.e., Mary's children are either happy or called Peter) cannot be expressed in ALC: it requires an extension by object names. The extended logic (called ALCO) allows for writing

$$\texttt{mary} : \texttt{Person} \sqcap \exists \texttt{child\_of}.\texttt{Person} \sqcap \forall \texttt{child\_of}.(\texttt{Person} \sqcap (\texttt{Happy} \sqcup \{\texttt{pierre}\})).$$

It turns out that almost all description logics have similar expressiveness problems. If we allow for object names as concepts as done in ALCO, this wipes out the distinction between ABox assertions (which are about particular objects) and TBox concept inclusions which should not be about particular objects. This is unsatisfactory because the distinction is one of the very basic ideas of description logics.

- *Action descriptions*: Eiter et al. (2010) define a framework for minimal change of action descriptions that they call "action description updates".
- *Abstract argumentation*: researchers in that domain (see chapter "Argumentation and Inconsistency-Tolerant Reasoning" of this volume) are also interested by update and more generally change operators. An abstract argumentation system is represented by a graph whose vertices are arguments and whose arcs are attacks between arguments. Some authors including Boella et al. (2009), Cayrol et al. (2010), Liao et al. (2011), Booth et al. (2013), Coste-Marquis et al. (2014) are interested in the impact of a change (by adding / withdrawing arguments or attacks) on such systems. Baumann and Brewka (2010), Baumann (2012) introduced the notion of *enforcement*, which is very similar to the notion of update (cf. (Bisquert et al. 2013; Dupin de Saint-Cyr et al. 2016)) because the idea is to minimally modify an argumentation system in a way such that it satisfies a given goal (usually expressed in terms of arguments that should be accepted). They define a preference relation between argumentation systems, which is similar to preference relations between models in classical update.

5 Conclusion

Reasoning about action and change is one of the oldest topics in Artificial Intelligence. Since 1995, the topic is the subject of a biennial workshop *International Workshop on Nonmonotonic Reasoning, Action and Change (NRAC)* held in conjunction with the IJCAI (*International Joint Conference on Artificial Intelligence*) conference.

Several periods followed, during which the researchers were interested in conceiving several formal settings for modeling the important tasks related to reasoning about action: STRIPS first, then approaches based on minimal change, and then approaches based on *successor state axioms*. To this variety of formal settings corresponds a variety of languages for representing and reasoning about action: propositional logic, situation calculus, dynamic logic, graphical models (among others).

Reasoning about action and change has close connections with other areas of Artificial Intelligence, including non-monotonic reasoning, belief change, reasoning under uncertainty, planning (and in particular Markov Decision Processes); it also has links with control theory (more precisely, Kalman filtering and discrete event systems).

References

Alchourrón C, Gärdenfors P, Makinson D (1985) On the logic of theory change: partial meet contraction and revision functions. J Symb Log 50:510–530

Aucher G, Bolander T (2013) Undecidability in epistemic planning. In: Proceedings of the 23rd international joint conference on artificial intelligence (IJCAI/AAAI'13), pp 27–33

Baral C, Zhang Y (2005) Knowledge updates: semantics and complexity issues. Artif Intell 164(1–2):209–243

Batusov V, Soutchanski M (2018) Situation calculus semantics for actual causality. In: Proceedings of the 32nd national conference on artificial intelligence (AAAI'18), pp 1744–1752

Baumann R (2012) What does it take to enforce an argument? Minimal change in abstract argumentation. In: Proceedings of the 19th European conference on artificial intelligence (ECAI'12), pp 127–132

Baumann R, Brewka G (2010) Expanding argumentation frameworks: enforcing and monotonicity results. In: Proceedings of the 3rd international conference on computational models of argument (COMMA'10), pp 75–86

Benferhat S, Dubois D, Prade H (2000) Kalman-like filtering in a possibilistic setting. In: Proceedings of the 14th European conference on artificial intelligence (ECAI'00), pp 8–12

Berger S, Lehmann DJ, Schlechta K (1999) Preferred history semantics for iterated updates. J Log Comput 9(6):817–833

Bisquert P, Cayrol C, Dupin de Saint-Cyr F, Lagasquie-Schiex M-C (2013) Enforcement in argumentation is a kind of update. In: Proceedings of the 7th international conference on scalable uncertainty management (SUM'13), pp 30–43

Boella G, Kaci S, van der Torre L (2009) Dynamics in argumentation with single extensions: attack refinement and the grounded extension. In: Proceedings of the 8th international conference on autonomous agents and multiagent systems (AAMAS'09), pp 1213–1214

Bolander T, Jensen MH, Schwarzentruber F (2015) Complexity results in epistemic planning. In: Proceedings of the 24th international joint conference on artificial intelligence (IJCAI'15), pp 2791–2797

Booth R, Kaci S, Rienstra T, van der Torre L (2013) A logical theory about dynamics in abstract argumentation. In: Proceedings of the 7th International conference on scalable uncertainty management (SUM'13), pp 148–161

Boutilier C (1998) A unified model of qualitative belief change: a dynamical systems perspective. Artif Intell 98(1–2):281–316

Brewka G, Hertzberg J (1993) How to do things with worlds: on formalizing actions and plans. J Log Comput 3(5):517–532

Castilho M, Gasquet O, Herzig A (1999) Formalizing action and change in modal logic I: the frame problem. J Log Comput 9(5):701–735

Cayrol C, Dupin de Saint-Cyr F, Lagasquie-Schiex M-C (2010) Change in abstract argumentation frameworks: adding an argument. J Artif Intell Res 38:49–84

Cooper MC, Herzig A, Maffre F, Maris F, Régnier P (2016) A simple account of multi-agent epistemic planning. In: Proceedings of the 22nd European conference on artificial intelligence (ECAI'16), pp 193–201

Cordier M-O, Siegel P (1995) Prioritized transitions for updates. In: Proceedings of the 3rd European conference on symbolic and qualitative aspects of reasoning under uncertainty (ECSQARU'95), pp 142–150

Cossart C, Tessier C (1999) Filtering versus revision and update: Let us debate! In: Proceedings of the 5th European conference on symbolic and qualitative aspects of reasoning under uncertainty (ECSQARU'99), pp 116–127

Coste-Marquis S, Konieczny S, Mailly J-G, Marquis P (2014) On the revision of argumentation systems: minimal change of arguments status. In: Proceeding of the 14th international conference on principles of knowledge representation and reasoning (KR'14), pp 52–61

Creignou N, Ktari R, Papini O (2015) Belief update within propositional fragments. In: Proceedings of the European conference on symbolic and quantitative approaches to reasoning with uncertainty (ECSQARU'15), pp 165–174

Dannenhauer ZA, Cox MT (2017) Rationale-based visual planning monitors for cognitive systems. In: Proceedings of the 30th international Florida artificial intelligence research society conference (FLAIRS'17), pp 182–185

Dannenhauer D, Munoz-Avila H, Cox MT (2016) Informed expectations to guide gda agents in partially observable environments. In: Proceedings of the 25th international joint conference on artificial intelligence (IJCAI'16), pp 2493–2499

Dupin de Saint-Cyr F, Bisquert P, Cayrol C, Lagasquie-Schiex M-C (2016) Argumentation update in YALLA (Yet another logic language for argumentation). Int J Approx Reason 75:57–92

Dupin de Saint-Cyr F (2008) Scenario update applied to causal reasoning. In: Proceedings of the 11th international conference on principles of knowledge representation and reasoning (KR'08), pp 188–197

Dupin de Saint-Cyr F, Lang J (2011) Belief extrapolation (or how to reason about observations and unpredicted change). Artif Intell 175(2):760–790

Dean T, Kanazawa K (1989) A model for reasoning about persistence and causation. Comput Intell 5(2):142–150

Delgrande JP, Levesque HJ (2012) Belief revision with sensing and fallible actions. In: Proceedings of the 13th international conference on principles of knowledge representation and reasoning (KR'12), pp 148–157

Doherty P, Lukaszewicz W, Madalinska-Bugaj E (1998) The PMA and relativizing minimal change for action update. In: Proceedings of the 6th international conference on principles of knowledge representation and reasoning (KR'98), pp 258–269

Dubois D, Dupin de Saint-Cyr F, Prade H (1995) Update postulates without inertia. In: Proceedings of the 3rd European conference on symbolic and qualitative aspects of reasoning under uncertainty (ECSQARU'95), pp 162–170

Dubois D, Prade H (1993) Belief revision and updates in numerical formalisms: an overview, with new results for the possibilistic framework. In: Proceedings of the 13th international joint conference on artificial intelligence (IJCAI'93), pp 620–625

Eiter T, Erdem E, Fink M, Senko J (2010) Updating action domain descriptions. Artif Intell 174(15):1172–1221

Fikes R, Nilsson N (1971) STRIPS: a new approach to the application of theorem proving to problem solving. Artif Intell 2:189–208

Finger J (1987) Exploiting constraints in design synthesis. PhD thesis, Stanford University, Stanford

Forbus K (1989) Introducing actions into qualitative simulation. In: Proceedings of the 11th international joint conference on artificial intelligence (IJCAI'89), pp 1273–1278

Friedman N, Halpern JY (1999) Modeling belief in dynamic systems, part II: revision and update. J Artif Intell Res 10:117–167

Friedman N, Halpern J (1994) A knowledge based framework for belief change part II: revision and update. In: Proceedings of the 4th international conference on principles of knowledge representation and reasoning (KR'94), pp 190–201

Geffner H (1990) Causal theories for nonmonotonic reasoning. In: Proceedings of the 8th national conference on artificial intelligence (AAAI'90), pp 524–530

Gelfond M, Lifschitz V (1993) Representing action and change by logic programs. J Log Program 17:301–321
Ghallab M, Howe A, Knoblock C, McDermott D, Ram A, Veloso M, Weld C, Wilkins D (1998) The planning domain definition language. Technical report. AIPS-98 Planning Competition
Giordano L, Martelli A, Schwind C (1998) Dealing with concurrent actions in modal action logics. In: Proceedings of the 13th European conference on artificial intelligence (ECAI'98), pp 537–541
Giunchiglia E, Lee J, Lifschitz V, McCain N, Turner H (2004) Nonmonotonic causal theories. Artif Intell 153:49–104
Giunchiglia E, Lifschitz V (1998) An action language based on causal explanation: Preliminary report. In: Proceedings of the 15th national conference on artificial intelligence (AAAI'98), pp 623–630
Goldszmidt M, Pearl J (1992) Rank-based systems: a simple approach to belief revision, belief update, and reasoning about evidence and actions. In: Proceedings of the 3rd international conference on principles of knowledge representation and reasoning (KR'92), pp 661–672
Hanks S, McDermott D (1994) Modelling and uncertain world i: symbolic and probabilistic reasoning about change. Artif Intell 66:1–55
Hanks S, McDermott D (1986) Default reasoning, nonmonotonic logics, and the frame problem. In: Proceedings of the 5th national conference on artificial intelligence (AAAI'86), pp 328–333
Heni A, Ben Amor N, Benferhat S, Alimi A (2007) Dynamic possibilistic networks: representation and exact inference. In: Proceedings of the 4th IEEE international conference on computational intelligence for measurement systems and applications (CIMSA'07), pp 1–8
Herzig A (1996) The PMA revisited. In: Proceedings of the 5th international conference on principles of knowledge representation and reasoning (KR'96), pp 40–50
Herzig A (2014) Belief change operations: a short history of nearly everything, told in dynamic logic of propositional assignments. In: Proceedings of the 14th international conference on principles of knowledge representation and reasoning (KR'14), pp 141–150
Herzig A (2015) Logics of knowledge and action: critical analysis and challenges. Auton Agents Multi-Agent Syst 29(5):719–753
Herzig A, Rifi O (1999) Propositional belief base update and minimal change. Artif Intell 115(1):107–138
Herzig A, Varzinczak IJ (2007) Metatheory of actions: beyond consistency. Artif Intell 171:951–984
Herzig A, Lang J, Marquis P (2013) Propositional update operators based on formula/literal dependence. ACM Trans Comput Log 14(3):24:1–24:31
Herzig A, Lang J, Marquis P (2003) Action representation and partially observable planning using epistemic logic. In: Proceedings of the 18th international joint conference on artificial intelligence (IJCAI'03), pp 1067–1072
Herzig A, Lang J, Marquis P, Polacsek T (2001) Updates, actions, and planning. In: Proceedings of the 17th international joint conference on artificial intelligence (IJCAI'01), pp 119–124
Hunter A, Delgrande J (2015) Belief change with uncertain action histories. J Artif Intell Res 53:779–824
Katsuno H, Mendelzon A (1991) On the difference between updating a knowledge base and revising it. In: Proceedings of the 1st international joint conference on principles of knowledge representation and reasoning (KR'91), pp 387–394
Keller A, Winslett M (1985) On the use of an extended relational model to handle changing incomplete information. IEEE Trans Softw Eng SE-11:7, 620–633
Lang J (2007) Belief update revisited. In: Proceedings of the 20th international joint conference on artificial intelligence (IJCAI'07), pp 2517–2522
Lang J, Liberatore P, Marquis P (2003) Propositional independence: formula-variable independence and forgetting. J Artif Intell Res 18:391–443
Lang J, Marquis P, Williams M-A (2001) Updating epistemic states. In: Proceedings of the 14th Australian joint conference on artificial intelligence (AI'01), pp 297–308
Li Y, Yu Q, Wang Y (2017) More for free: a dynamic epistemic framework for conformant planning over transition systems. J Log Comput 27(8):2383–2410

Liao B, Jin L, Koons RC (2011) Dynamics of argumentation systems: a division-based method. Artif Intell 175(11):1790–1814
Lifschitz V (1990) Frames in the space of situations. Artif Intell 46:365–376
Lifschitz V, Rabinov A (1989) Things that change by themselves. In: Proceedings of the 11th international joint conference on artificial intelligence (IJCAI'89), pp 864–867
Lin F (1995) Embracing causality in specifying the indirect effects of actions. In: Proceedings of the 14th international joint conference on artificial intelligence (IJCAI'95), pp 1985–1991
Liu H, Lutz C, Milicic M, Wolter F (2011) Foundations of instance level updates in expressive description logics. Artif Intell 175(18):2170–2197
Marquis P (1994) Possible models approach via independency. In: Proceedings of the 11th European conference on artificial intelligence (ECAI'94), pp 336–340
McCarthy J (1977) Epistemological problems of artificial intelligence. In: Proceedings of the 5th international joint conference on artificial intelligence (IJCAI'77), pp 1038–1044
McCarthy J (1986) Applications of circumscription to formalizing common-sense knowledge. Artif Intell 28(1):1038–1044
McCarthy J, Hayes P (1969) Some philosophical problems from the standpoint of artificial intelligence. Mach Intell 4:463–502
Moinard Y (2000) Note about cardinality-based circumscription. Artif Intell 119(1–2):259–273
Molineaux M, Aha DW (2014) Learning unknown event models. In: Proceedings of the 28th national conference on artificial intelligence (AAAI'14), pp 395–401
Morreau M (1992) Planning from first principles. In: Belief revision. Cambridge University Press, Cambridge, pp 204–219
Motik B, Rosati R (2010) Reconciling description logics and rules. J Assoc Comput Mach 57(5)
Pearl J (1988) Embracing causality in formal reasoning. Artif Intell 35:259–271
Pednault E (1989) ADL: exploring the middle ground between STRIPS and the situation calculus. In: Proceedings of the 1st international conference on principles of knowledge representation and reasoning (KR'89), pp 324–332
Reiter R (1991) The frame problem in the situation calculus: a simple solution (sometimes) and a completeness result for goal regression. In: Artificial intelligence and mathematical theory of computation: papers in honor of John McCarthy. Academic, pp 359–380
Sandewall E (1995) Features and fluents. Oxford University Press, Oxford
Scherl R, Levesque HJ (2003) The frame problem and knowledge producing actions. Artif Intell 144(1–2):
Shoham Y (1988) Reasoning about change: time and causation from the standpoint of artificial intelligence. MIT Press, Cambridge
Slota M (2012) Updates of Hybrid Knowledge bases. Universidade Nove de Lisboa PhD thesis
Slota M, Leite J, Swift T (2011) Splitting and updating hybrid knowledge bases. Theory Pract Log Program 11(4–5):801–819
Slota M, Leite J (2010) On semantic update operators for answer-set programs. In: Proceedings of the 19th European conference on artificial intelligence (ECAI'10), pp 957–962
Stein L, Morgenstern L (1994) Motivated action theory: a formal theory of causal reasoning. Artif Intell 71(1):1–42
Thielscher M (1995) The logic of dynamic systems. In: Proceedings of the 14th international joint conference on artificial intelligence (IJCAI'95), pp 1956–1962
Thielscher M (1997) Ramification and causality. Artif Intell 89(12):317364
Turner H (1999) A logic of universal causation. Artif Intell 113(1–2):87–123
van Ditmarsch H, Herzig A, de Lima T (2011) From situation calculus to dynamic epistemic logic. J Log Comput 21(2):179–204
Winslett M (1988) Reasoning about action using a possible models approach. In: Proceedings of the 7th national conference on artificial intelligence (AAAI'88), pp 89–93
Winslett M (1990) Updating logical databases. Cambridge University Press, Cambridge

Multicriteria Decision Making

Christophe Gonzales and Patrice Perny

Abstract This chapter aims to present the main models used for preference aggregation and decision support in a unified framework.After recalling the definition of a multicriteria decision making problem, we distinguish two approaches for preference aggregation:*the compare then aggregate* approach (denoted CA) and the *aggregate then compare* approach (denoted AC). We first present some procedures allowing the construction of an overall preference relation (e.g., a dominance or concordance relation) from several binary relations. Then we consider the AC approach and present some scalarizing functions allowing the definition of an overall score from partial numerical evaluations. In particular we review the min, Tchebycheff, OWA, WOWA aggregators and Choquet and Sugeno integrals.

1 Introduction

Taking into account multiple and conflicting points of view in the analysis of preferences and studying the properties of preference aggregation procedures is quite old. Long before the birth of multicriteria optimization, collective decision-making problems and aggregation of preferences were addressed through the theory of voting, as can be seen in the writings of Borda (1781) and Condorcet (1785); these topics remained active until today, giving birth to the theory of social choice (Arrow 1951; Sen 1986a). In economics, the account of multiple criteria to explain rational behaviors dates back to the 1900s, notably with the works of Pareto (1906). The consideration of multiple objectives in mathematical programming was introduced in the middle of the 20th century with the goal-programming (Charnes et al. 1955). This work was then developed in the 1970s under the impetus of Geoffrion (see for example Geoffrion et al. 1973). The first international conference dedicated to

C. Gonzales
Aix Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
e-mail: christophe.gonzales@univ-amu.fr

P. Perny (✉)
Sorbonne Université, CNRS, LIP6, Paris, France
e-mail: patrice.perny@lip6.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_16

multicriteria decision-making took place in 1973 in South Carolina. This activity continued in Europe with the first Euro Working Group on multi-criteria decision-making led by Bernard Roy who appeared in 1975. The work on multiobjective combinatorial optimization followed a few years later and the first international conference on the subject (MOPGP) was established in 1994. In Artificial intelligence, taking into account multiple objectives in problem solving also appears in the same period, in particular in multiobjective state-space search (Stewart and White III 1991; Laumanns et al. 2002), multicriteria planning problems (White 1982; Gbor et al. 1998), and problems of satisfaction of flexible constraints (Schiex et al. 1995). Concerning modeling and aggregation of preferences, the work in AI is distinguished by the emphasis on qualitative models and, more recently, on the aspects of representation and automatic learning of preferences. The latter two aspects go beyond the scope of this chapter, which focuses on the main models used for preference aggregation.

2 Multicriteria Decision Problems

A *multicriteria* decision problem is characterized by the explicit consideration of several objectives to be optimized simultaneously in the analysis of preferences, the comparison of solutions and the determination of the optimal solution(s). To formally introduce a multicriteria decision problem, we first define a set A of alternatives (potential actions, feasible solutions, candidates) that can be given explicitly (for example by listing the solutions considered) or implicitly (for example by specifying a set of constraints or properties that the solutions must satisfy). In all cases, the A set defines the solutions on which one wishes to make the decision analysis, that is to say that they are the only ones available, the only ones realizable or the only ones admissible. We then introduce a finite set of criteria $N = \{1, \ldots, n\}$ taking the form of functions $f_i, i \in N$, modeling the objectives of the decision maker. For every $x \in A$ and every $i \in N$, we call *performance* of x on criterion i the quantity $f_i(x)$ reflecting the value of a with respect to criterion i. This formalism also applies to problems of collective decision-making. In such problems, N represents a set of agents and the performance $f_i(x)$ then represents the utility of the solution x from the point of view of agent i. The criteria f_i are defined on A and respectively valued in an ordered set $X_i, i \in N$.

The set $\mathscr{X} = X_1 \times \cdots \times X_n$ which constitutes a new description space of the alternatives, is called *the space of criteria*. To simplify the notations, we will set $x_i = f_i(x)$ for all $x \in A$ and all $i \in N$. Any solution x is therefore represented in $\mathscr{X}$ by a vector $(x_1, \ldots, x_n)$. We will assume that $X_i \subseteq \mathbb{R}$ for all $i \in N$ and therefore that $\mathscr{X} \subseteq \mathbb{R}^n$. To simplify the presentation, it will also be assumed that functions f_i are to be maximized (this is not restrictive because we can change the sign of evaluations to pass from minimization to maximization). Sometimes performances have no cardinal meaning and only their order matters to state preferences. In other cases they may represent a cardinal utility (uniquely defined up to a positive affine

transformation) or even an absolute evaluation reflecting the intrinsic attractiveness of a solution on each of the criteria considered.

Multicriteria decision problems vary depending on the question asked. We can distinguish *choice* problems where we try to determine the best solutions, *ranking* problems where the aim is to order, at least partially, the solutions according to their relative merit, or *sorting* problems where one seeks to assign the solutions to predefined categories according to their intrinsic value. To summarize, a multicriteria decision problem can always be formally characterized by a triple of the form:

$$(A, \{f_1, \ldots, f_n\}, Q)$$

where $Q \in \{\text{choice, ranking, sorting}\}$ is the question formulated. The choice problem is the one most often encountered in the theory and practice of multicriteria analysis. It aims to find a solution that optimizes the different criteria as well as possible, or to find a subset of solutions, as small as possible, containing the best solutions. Formulated in this way, multicriteria optimization is an ill-posed problem. Indeed, due to the potentially conflicting nature of the criteria, there is generally no solution optimizing all the criteria simultaneously. The only natural preference weak-order that can be built on A without adding preferential information to the description of the problem is in fact the so-called *weak Pareto dominance* relation denoted $\succsim_P$ and is defined as follows:

$$x \succsim_P y \text{ iff } [\forall i \in N,\ x_i \geq y_i]$$

Relation $\succsim_P$ is a weak partial order on A (i.e. a reflexive and antisymmetric binary relation, transitive but not complete) and it generally leaves many pairs of solutions incomparable. It is enough that a solution x is better than a solution y on one criterion and that it is the opposite on another criterion so that we can no longer compare them. To circumvent this difficulty, one generally seeks to construct a richer and more discriminating preference relation on the set of alternatives. This relation will be denoted here $\succsim$ with the convention that $x \succsim y$ means that x is judged at least as good as y given the performance vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$. Obviously one will generally be interested in constructing a transitive preference relation $\succsim$ such that $x \succsim_P y \Rightarrow x \succsim y$ for all $x, y \in A$, thus refining the weak Pareto dominance.

We can also define a strict preference relation $\succ$ as the asymmetric part of $\succsim$. We obtain $x \succ y$ iff $[x \succsim y \text{ and not}(y \succsim x)]$. For example, the asymmetric part of relation $\succsim_P$ is the so-called *Pareto dominance* denoted $\succ_P$. We have:

$$x \succ_P y \text{ iff } \begin{cases} \forall i \in N,\ x_i \geq y_i \\ \exists k \in N,\ x_k > y_k \end{cases}$$

Given a strict preference relation $\succ$ defined on A, the non-dominated solutions of A are formally defined as follows: $ND(A, \succ) = \{x \in A : \forall y \in A, \text{non}(y \succ x)\}$. This set is non-empty as soon as $\succ$ is transitive. For example, the set $ND(A, \succ_P)$ is never empty; this is the set of Pareto-optimal solutions also known as the Pareto set.

If, as suggested above, we work with a preference relation $\succ$ which refines the Pareto dominance $\succ_P$, the set $ND(A, \succ)$ will therefore be a subset of the Pareto set.

The overall weak preference relation $\succsim$ must be constructed from the n criteria representing the different points of view considered relevant in the analysis, taking into account how the decision-maker wants to define the resultant of potentially conflicting criteria. This is the multicriteria aggregation phase that we present in the next section.

3 Preference Aggregation

The *preference aggregation* problem consists of synthesizing information that reflects different aspects or points of view (e.g., performance indices, utilities, preferences), sometimes conflicting, on a same set of alternatives. It is critically important in many procedures used for assessment, comparison or classification in multicriteria decision support. Whether it is a problem of choice, ranking or sorting, the central question is always a problem of comparison. Thus, in a problem of choice, the identification of the best candidate needs to be able to compare it to all others; in a ranking problem, we need to compare any pair of alternatives; in sorting problems, assigning a solution to a category is often done by comparing the solution to a reference vector. In a multicriteria decision problem, the comparison of two solutions is performed on the basis of their respective performance vectors. For this purpose, one necessarily resorts to an *aggregation rule* to construct the overall preference relation $\succsim$. Aggregation rules can formally be introduced as follows:

Definition 1 An *aggregation rule* is a function that defines the preference $x \succsim y$ for any pair of alternatives (x, y) in $A \times A$ from performance vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ as follows:

$$x \succsim y \text{ iff } h(x_1, \ldots, x_n, y_1, \ldots, y_n) \geq 0 \tag{1}$$

where h is a real-valued function defined on $\mathbb{R}^{2n}$, non-decreasing in the n first arguments and non-increasing in the n last arguments such that $h(x, x) \geq 0$ for all $x \in \mathbb{R}^n$ (which enforces the reflexivity of $\succsim$ and the compatibility with the weak Pareto-dominance).

The h function which tests the preference of x over y performs on the one hand *the aggregation* of performances x_i and y_i, $i = 1, \ldots, N$ and, on the other hand, the *comparison* of solutions x and y through their performance vectors. Generally, these two steps (aggregation and comparison) are clearly distinguished and h is then defined as the combination of an aggregation function (also known as a scalarizing function) $\psi : \mathbb{R}^n \to \mathbb{R}$ allowing the synthesis of a vector of n performances in one scalar, and a performance comparison function $\phi : \mathbb{R}^2 \to \mathbb{R}$ which compares two performances. We can thus distinguish two different modes of operation giving rise to two distinct approaches to multicriteria aggregation.

The "Aggregate Then Compare" Approach (AC)

It consists of summarizing the value of any solution x by an overall score $\psi(x)$ calculated from its performance vector. This score is intended to summarize the overall value of the action and serves as a basis for the multicriteria comparison of solutions. This way of evaluating and comparing vectors of grades is very widespread, for example in the academic world where the comparison of two students is based on the average of their marks. The general form of the decision rules under the AC approach is as follows:

$$x \succsim y \text{ iff } \phi(\psi(x_1, \ldots, x_n), \psi(y_1, \ldots, y_n)) \geq 0 \tag{2}$$

where ψ is a non-decreasing function of its arguments. On the other hand, function ϕ allows the comparison of x and y on the basis of $\psi(x)$ and $\psi(y)$. The most common choice for ϕ is $\phi(x, y) = x - y$. In this case, $x \succsim y$ holds when $\psi(x) \geq \psi(y)$.

Example 1 The Nash product often used in Game Theory is an aggregation function leading to the following preference relation:

$$x \succsim y \text{ iff } \prod_{i=1}^{n} x_i \geq \prod_{i=1}^{n} y_i$$

This is clearly an instance of the AC approach where ψ is the product function and $\phi(x, y) = x - y$.

The "Compare Then Aggregate" Approach (CA)

It consists of comparing, criterion by criterion, the performances of the alternatives and then to aggregate these comparisons. Thus, for each pair (x, y) and each criterion i, one can define a binary index of partial preference $\phi_i(x, y)$ where ϕ_i is an increasing function of x_i, decreasing of y_i. The preference $x \succsim y$ is then defined by aggregating partial preference indices. Formally, we have:

$$x \succsim y \text{ iff } \psi(\phi_1(x, y), \ldots, \phi_n(x, y)) \geq 0 \tag{3}$$

Generally in this approach, each ϕ_i function is used to compare the performances of two alternatives on the same criterion (criterion i). There are however a few exceptions in the case of criteria sharing the same valuation scale (e.g., the Lorenz dominance relation introduced later in the chapter). In this latter case, ϕ_i is used to compare two performances associated with different criteria. In the CA approach, one can use the same aggregation functions ψ as for the AC approach but it is used for the aggregation of partial preference indices $\phi_i(x, y)$ and not for aggregating the performances themselves. It must therefore be assumed that one can compare quantities of type $\phi_i(x, y)$ and $\phi_k(x, y)$ for $i \neq k$ but it is not necessary to assume that we can compare the performances of an alternative on different criteria. If $\phi_i(x, y)$ depends only on x_i and y_i, then it defines a preference relation on A restricted to

criterion i. For example, we can consider preorders constructed from performances by setting $x \succsim_i y$ if and only if $\phi_i(x, y) \geq 0$ with:

$$\phi_i(x, y) = \begin{cases} 1 \text{ if } x_i > y_i \\ 0 \text{ if } x_i = y_i \\ -1 \text{ if } x_i < y_i \end{cases} \tag{4}$$

It is also possible to use, for ϕ_i, functions with thresholds such as $\phi_i(x, y) = x - y + q_i$ where q_i is a positive quantity representing an indifference threshold, i.e., the biggest difference that is not significant of a preference. In this case, if we set $x \succsim y$ iff $\phi_i(x, y) \geq 0$, we obtain a semi-order structure[1] well known in preference modeling. Other choices are possible for ϕ_i, leading to more general ordinal structures defined from numerical evaluations, see Roubens and Vincke (1985), Pirlot and Vincke (1997). Alternatively, the $\phi_i(x, y)$ indices can be used to represent preference intensities, monotonically increasing with differences of type $x_i - y_i$; this amounts to defining a fuzzy preference relation $\succsim_i$ for every criterion $i \in N$ (Perny and Roy 1992; Fodor and Roubens 1994). In all these cases, we see that the CA approach amounts to construct n preference relations (one per criterion) and then to aggregate these relations. By way of illustration, let us give the following example:

Example 2 The lexicographic aggregation is characterized by the following definition of the overall preference:

$$x \succ y \text{ iff } \exists k \in N, \begin{cases} x_k > y_k \\ \forall i < k, x_i = y_i \end{cases}$$

this is clearly an instance of the CA approach with $\psi(z_1, \ldots, z_n) = \sum_{i=1}^{n} 2^{n-i} z_i$ and ϕ_i defined as in Eq. (4).

Comparison of AC and CA

Both approaches AC and CA are represented in the following diagram showing the two possible paths to decide whether x is preferred to y from two performance vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$:

$$\begin{array}{ccc} (x_1, \ldots, x_n), (y_1, \ldots, y_n) & \xrightarrow{\psi} & \psi(x), \ \psi(y) \\ \Big\downarrow{\scriptstyle \phi_i} & & \Big\downarrow{\scriptstyle \phi} \\ \phi_1(x, y), \ldots, \phi_n(x, y) & \xrightarrow{\psi} & x \succsim y? \end{array}$$

The fundamental difference between AC and CA lies in the fact that the aggregation does not involve the same objects due to the inversion of the order of execution

[1] We recall that a semi-order is a complete, Ferrers and semi-transitive binary relation (see Pirlot and Vincke 1997). In a semi-order $\succsim_i$ defined with threshold q_i, we have $x \succ_i y$ if $x_i - y_i > q_i$ and $x \sim_i y$ if $|x_i - y_i| \leq q_i$.

of the aggregation and comparison operations. In AC, function ψ aggregates the performances x_i on the one hand and the performances y_i on the other hand whereas in CA we aggregate the preference indices $\phi_i(x, y)$. It should be emphasized that this distinction concerns the process of computing h in Eq. (1) and thus remains essentially formal. There are indeed some aggregation rules whose function h can be written either in one or the other of the two forms distinguished. For example, this is the case of preferences defined from linear aggregations of performances as follows:

$$x \succsim y \text{ iff } \sum_{i=1}^{n} x_i \geq \sum_{i=1}^{n} y_i \tag{5}$$

In this case, one naturally recognizes an instance of the AC approach:

$$\begin{aligned} h(x_1, \ldots, x_n,\ y_1, \ldots, y_n) &= \phi(\psi(x_1, \ldots,\ x_n),\ \psi(y_1, \ldots,\ y_n)) \\ \psi(z_1, \ldots, z_n) &= \sum_{j=1}^{n} z_j \\ \phi(u, v) &= u - v \end{aligned}$$

Nevertheless, using the same ψ function, one may also consider that it is an instance of the CA approach by setting:

$$\begin{aligned} h(x_1, \ldots,\ x_n, y_1, \ldots, y_n) &= \psi(\phi_1(x, y), \ldots, \phi_n(x, y)) \\ \phi_i(x, y) &= x_i - y_i \end{aligned}$$

Despite this non-empty intersection due to some singular cases, the distinction between the AC and CA approaches is important for structuring the field of multicriteria aggregation rules (but also for transposing them into the field of decision under uncertainty) and to differentiate between the main benefits and disadvantages of each of these approaches.

The choice of the AC approach based on the construction of a "scalarizing" function ψ often seduces by its operational simplicity and its intuitive appearance. When the preferential information is sufficiently rich to enable the construction of the overall evaluation function ψ, the multicriteria decision problem reduces to the problem of optimizing function ψ. However, it should be stressed that the AC approach requires a particularly rich information. We need to know how to commensurate partial evaluations on different criteria, what is the importance of every criteria and group of criteria, and how they interact in the definition of preferences. Moreover, the preference $x \succsim y$ is often defined using $\phi(\psi(x), \psi(y)) = \psi(x) - \psi(y)$ which simply amounts to comparing the two values $\psi(x)$ and $\psi(y)$. This mode of comparison therefore presupposes a priori that all solutions are reducible to scalar values and are comparable, and that preferences are transitive. From a descriptive point of view, this hypothesis is often debatable, given the necessarily imperfect information available, the heterogeneity of the performances and also the existence possible

conflicts between the criteria considered, making it difficult to compare certain pairs of solutions.

The CA approach, on the other hand, is particularly suited to multicriteria aggregation when there is insufficient information to compare the performances issued from distinct criteria and/or when certain criteria are quantitative and others are qualitative, or when we want to relax the assumption that all the alternatives are comparable. A binary relation $\succsim_i$ is constructed using ϕ_i for every $i \in N$ and then we have to aggregate the $(\succsim_1, \ldots, \succsim_n)$. This ordinal aggregation problem is generally quite difficult to solve and a number of negative results have shown the prescriptive limits of decision rules based on an ordinal aggregation (for a compilation of the main results see Sen 1986b). In particular, a method of aggregation defined by Eqs. (3) and (4) does not guarantee any transitivity property for $\succsim$ except in some very restrictive particular cases such as the lexicographic preferences introduced in Example 2. If we consider, for example, the majority rule that appears naturally as an instance of the CA approach obtained by setting $\psi(z_1, \ldots, z_n) = \sum_{i=1}^{n} z_i$ and $\phi_i(x, y)$ defined as in Eq. (4), a non-transitive preference is easily obtained as shown in the following example:

Example 3 Let us consider a problem with three criteria to be maximized and three alternatives $x = (3, 1, 2)$, $y = (2, 3, 1)$ and $z = (1, 2, 3)$. We have: $\phi_1(x, y) = \phi_3(x, y) = 1, \phi_2(x, y) = -1$ and therefore $x \succ y$. On the other hand we have: $\phi_1(y, z) = \phi_2(y, z) = 1, \phi_3(y, z) = -1$ and therefore $y \succ z$. Finally we have: $\phi_2(z, x) = \phi_3(z, x) = 1, \phi_1(z, x) = -1$ and therefore $z \succ x$. This intransitivity of the strict majority rule is well known in voting theory under the name of Condorcet paradox.

Despite the descriptive appeal of the majority rule, this lack of transitivity is rather problematic for determining a ranking of solutions or even a choice and an additional exploitation phase is then necessary. For this reason, the CA approach is mainly used for decision problems involving only a finite and small set of alternatives. The next section is intended to introduce some standard decision models falling in the CA approach and, when necessary, some techniques for making a choice or ranking from a non-transitive relation constructed with this approach. We will review the main models of the AC approach in the following section.

4 Decision Models in the CA Approach

Below, we introduce different binary relations $\succsim$ to compare the solutions of A on the basis of their performance vectors. Let us first introduce *dominance* relations and then *outranking* relations resulting from a concordance rule.

4.1 Dominance Relations

We have previously introduced the notions of weak Pareto dominance which is a basic preference relation. It is not very discriminating but can be enriched in various ways. We now present some richer dominance relations often used in decision theory. Most of these dominance relations are obvious instances of the CA approach. We will therefore not specify the functions ψ and ϕ_i characteristic of the CA approach, except in the few cases where membership to CA is less straightforward.

Oligarchic Dominance

An *Oligarchic weak-dominance* is a transitive preference relation that concentrates the decisive power on a subset of criteria $O \subseteq N$, namely the *Oligarchy*. It is formally defined as follows:

$$x \succsim y \text{ iff } \forall i \in O, x_i \geq y_i \tag{6}$$

When $O = N$ we obtain the weak Pareto dominance introduced before. When O only contains some of the criteria, the dominance defined by Eq. (6) is all the more discriminating than the cardinal of O is reduced. When O is reduced to a singleton we obtain a *dictatorial* aggregation rule. A refinement of this dictatorial procedure is given by the lexicographic procedure introduced in Example 2.

ε-Dominance

An interesting weakening of the weak Pareto dominance is the *ε-dominance* defined as follows:

$$x \succsim_\varepsilon y \text{ iff } \forall i \in N, (1+\varepsilon)x_i \geq y_i$$

for some arbitrarily small $\varepsilon > 0$. This relation is not transitive but it enables to "cover" the entire set of feasible alternatives with fewer solutions than needed for the weak Pareto dominance, as shown in the following example:

Example 4 Consider a tri-criteria problem with 4 feasible solutions $x = (10, 10, 10)$, $y = (11, 5, 10), z = (10, 2, 11), w = (4, 10, 3)$. In this example, the solutions x, y, z are Pareto-optimal while w is Pareto-dominated by x. If we consider $\varepsilon = 0.1$ then we can check that $x \succsim_\varepsilon y, x \succsim_\varepsilon z$ and $x \succsim_\varepsilon w$. We thus observe that x is at least as good as all the other solutions and covers by itself all the solutions under consideration. This can be an argument for choosing this solution. One can even check here that x strictly ε-dominates the solutions y, z, w, i.e., none of the following preferences $y \succsim_\varepsilon x, z \succsim_\varepsilon x$ and $w \succsim_\varepsilon x$ holds.

More generally, the notion of covering can be introduced as follows:

Definition 2 For all $\varepsilon > 0$ a subset $B \subseteq A$ of solutions is said to be a ε-covering of A when: $\forall x \in A, \exists y \in B, y \succsim_\varepsilon x$.

In general, for a given ε, several ε-coverings exist of different cardinalities, the most interesting being those that are minimal w.r.t. inclusion. The interest of this

concept is particularly evident in multicriteria decision problems with a large number of potential solutions. For example, let us consider the multicriteria shortest path problem; in this problem the set of Pareto-optimal solutions can grow exponentially with the number of vertices of the graph. This is well illustrated by the family of bi-criteria graphs introduced in Hansen (1980). In this family, the graph admitting $2n+1$ vertices includes 2^n Pareto-optimal solution-paths with distinct cost vectors of the form $(c, 2^n - 1 - c), c \in \{0, \ldots, 2^n - 1\}$. It then becomes impossible and useless to propose them all to the decision-maker. Instead, we can perform a selection of solutions that ε-covers the entire Pareto set. Under the assumption that the criteria values are positive integers and bounded by a quantity K, Papadimitriou and Yannakakis (2000) have indeed shown that for any number of criteria $n \geq 2$ and every set of alternatives A, there exists a ε-covering of A whose size is bounded from above by $\lceil \log K / \log(1+\varepsilon) \rceil^{n-1}$. This quantity remains polynomial in the size of the problem, when the number of criteria is fixed. For example, in the family of graphs considered by Hansen in (1980) the instance with 33 nodes ($n = 16$) leads to $2^{16} = 65536$ Pareto-optimal paths with distinct cost vectors of the form $(c, 2^{16} - 1 - c), c \in \{0, \ldots, 2^{16} - 1\}$. If we choose $\varepsilon = 0.1$ a covering of the Pareto-optimal cost vectors exists with at most 117 elements; this number decreases to 61 if $\varepsilon = 0.2$. For more details on the potential use of the ε-dominance and the associated covering concepts see Papadimitriou and Yannakakis (2000), Diakonikolas and Yannakakis (2008), Perny and Spanjaard (2008), Bazgan et al. (2009).

Lorenz Dominance

Among the Pareto-optimal solutions, not all are of equal interest to the decision-maker. Some are fairly balanced and show comparable levels of performance on each of the criteria while others alternate excellent performance and very bad points. In multicriteria analysis, the decision-maker often prefers a balanced solution, which does not favor a criterion at the expense of others. A similar principle appears in multi-agent decision problems found with the notion of equity; in this case the criteria measure individual utilities (see chapter "Collective Decision Making" of this volume). Formally, the idea of equity in the aggregation of preferences can be described by the following axiom, known as the "transfer principle" based on Pigou-Dalton transfers reducing inequalities (Shorrocks 1983; Moulin 1988):

Transfer principle. Let $x \in \mathbb{R}^n_+$ such that $x_i > x_j$ for $i, j \in N$. Then, for any ε such that $0 < \varepsilon \leq x_i - x_j$, $x - \varepsilon e_i + \varepsilon e_j \succ x$ where e_i (resp. e_j) is the vector whose ith (resp. jth) component equals 1, all the other components being equal to 0.

This axiom captures the notion of equity as follows: if $x_i > x_j$, a mean-preserving shift of performance improving x_j at the expense of performance x_i produces a better distribution of criteria satisfaction indices and therefore a better solution. Thus the vector $x = (10, 10)$ is preferred to the vector $y = (14, 6)$ because there is a size 4 transfer from y to x. Note that if the transfer was too large, for example $\varepsilon = 9$, the inequality would be increased rather than decreased. This is the reason why the ε amplitude of the transfer should remain lower than $x_i - x_j$.

The transfer principle is a mean-preserving operation since one quantity is removed from one component and added to another. It does not exist between two vectors whose average performances are not the same. However, it becomes more powerful when combined with the Pareto principle requiring that we strictly prefer a solution x to a solution y when $x \succ_P y$. For example, if one wishes to compare vectors $(11, 11)$ and $(12, 9)$, one can notice that $(11, 11)$ Pareto-dominates $(11, 10)$ and that $(11, 10)$ is deduced from the vector $(12, 9)$ by a Pigou-Dalton transfer of amplitude 1. Due to the Pareto principle and to the transfer principle, one has on the one hand $(11, 11) \succ (11, 10)$ and, on the other hand, $(11, 10) \succ (12, 9)$, whence, by transitivity, $(11, 11) \succ (12, 9)$. The vectors that can be compared by combining the Pareto principle and the transfer principle can be characterized using generalized Lorenz vectors and the notion of generalized Lorenz dominance (see Marshall and Olkin 1979; Shorrocks 1983):

Definition 3 For all $x \in \mathbb{R}^n_+$, the *generalized Lorenz vector* associated to x is defined by: $L(x) = (x_{\sigma(1)}, x_{\sigma(1)} + x_{\sigma(2)}, \ldots, x_{\sigma(1)} + x_{\sigma(2)} + \cdots + x_{\sigma(n)})$ where σ represents the permutation which sorts the components of x by increasing order. Thus $x_{\sigma(i)}$ represents the ith smallest component of x.

Definition 4 The generalized Lorenz dominance is a binary relation defined on $\mathbb{R}^n_+$ by: $\forall x, y \in \mathbb{R}^n_+,\ x \succsim_L y$ iff $L(x) \succsim_P L(y)$. The asymmetric part of this relation is defined by $x \succ_L y$ iff $L(x) \succ_P L(y)$.

The x vector Lorenz-dominates the y vector if $L(x)$ Pareto-dominates $L(y)$. To show that $\succsim_L$ is indeed an instance of the CA approach, it is sufficient to consider that:

$$\phi_i(x, y) = \begin{cases} 1 \text{ if } \sum_{j=1}^{i} x_{\sigma(j)} \geq \sum_{j=1}^{i} y_{\sigma(j)} \\ 0 \text{ otherwise} \end{cases}$$

and $\psi(z_1, \ldots, z_n) = \sum_{i=1}^{n} z_i - n$. The notion of Lorenz dominance was initially introduced to compare vectors with the same average (e.g., for comparing various income distributions over a population). The generalized version introduced above is more adapted to the context of multicriteria optimization because it makes it possible to compare vectors of performances that do not have the same average. The link between generalized Lorenz dominance and the transfer principle appears with the following result (Chong 1976):

Theorem 1 *For all pairs of vectors $x, y \in \mathbb{R}^n_+$, if $x \succ_P y$, or if x is obtained from y using a Pigou-Dalton transfer, the $x \succ_L y$. Conversely, if $x \succ_L y$, then there exists a sequence of Pigou-Dalton transfers and/or Pareto improvements allowing to pass from y to x.*

This result establishes the generalized Lorenz dominance as the minimal relation w.r.t. inclusion which simultaneously satisfies the Pareto principle and the principle of transfer. This dominance is transitive. To illustrate the use of Lorenz dominance to compare the vectors considered above, one can observe that $L(11, 10) =$

$(10, 21)$ while $L(12, 9) = (9, 21)$. We thus have $(11, 10) \succ_L (12, 9)$ since $(10, 21) \succ_P (9, 21)$.

A consequence of the preceding theorem is that if $x \succ_P y$ then $x \succ_L y$, which shows that the Lorenz dominance is potentially more discriminating than the Pareto dominance. It follows that $ND(\mathscr{X}, \succ_L) \subseteq ND(\mathscr{X}, \succ_P)$, that is to say that the non-dominated solutions in Lorenz's sense are Pareto-optimal solutions. Apart from a few specific cases, there are generally significantly fewer Lorenz-optimal than Pareto-optimal solutions. The Lorenz dominance thus appears as a natural refinement of the Pareto dominance allowing to remove unfair elements in the Pareto set.

Weighted Lorenz Dominance

The Lorenz dominance deals symmetrically with all components of the vectors that are compared. The $L(x)$ vector indeed remains invariant by permutation of the components of x and consequently the preference $x \succ_L y$ is not affected by a permutation of the components of x or y . This characteristic seems natural when one wishes to assign the same importance to all criteria or agents. On the other hand, if we want to give more weight to some of the criteria, we should consider a weighted extension of the Lorenz dominance. A first idea that naturally comes to mind to assign different weights to components is to duplicate them in proportion to the weights of the criteria (we assume here that the weights are rational numbers). Thus, if we want to compare the vectors $x = (10, 5, 15)$ and $y = (10, 12, 8)$ given that the criteria have weights defined by the vector $p = (3/6, 1/6, 2/6)$ we can consider the extended vectors $\tilde{x} = (10, 10, 10, 5, 15, 15)$ and $\tilde{y} = (10, 10, 10, 12, 8, 8)$ and test if $\tilde{x} \succ_L \tilde{y}$ or $\tilde{y} \succ_L \tilde{x}$. This is not the case here since the Lorenz vectors $(5, 15, 25, 35, 50, 65)$ and $(8, 16, 26, 36, 46, 68)$ are incomparable in terms of Pareto dominance. Here, the fact that criterion 3 is twice more important than criterion 1 did not allow us to prefer y although y distributes performance more equally than x.

A more elaborate way of proposing a weighted extension of the Lorenz dominance without having to duplicate the components (nor assuming that weights are rational numbers) is to associate to each vector x a cumulative function $F_x(z)$ which indicates the weight of the coalition formed by the criteria whose performance does not exceed threshold z. Denoting v the function which gives the weight of a subset of criteria, we have: $F_x(z) = v(\{i \in N, x_i \leq z\})$. We also consider the left inverse of F_x, denoted $\check{F}_x$ which reads $\check{F}_x(p) = \inf\{z \, in \, \mathbb{R} | F_x(z) \geq p\}$ for $p \in [0, 1]$. This quantity represents a kind of quantile function; it represents the minimum level z from which there exists a coalition of criteria satisfied at level z or more and which is of weight greater than or equal to p. Both F_x and $\check{F}_x$ are stepwise functions. We then define from F_x, F_y or $\check{F}_x$, $\check{F}_y$ the second order stochastic dominance by one of the following formulas which are known to be equivalent:

$$x \succsim_2 y \text{ iff } \forall z \in \mathbb{R}, F_x^2(z) \leq F_y^2(z) \qquad \text{with } F_x^2(z) = \int_{-\infty}^{z} F_x(t)dt \quad (7)$$

$$x \succsim_2 y \text{ iff } \forall p \in [0, 1], \check{F}_x^2(p) \geq \check{F}_y^2(p) \quad \text{with } \check{F}_x^2(p) = \int_{0}^{p} \check{F}_x(t)dt \quad (8)$$

This dominance is transitive and coincides with the second order stochastic dominance that will be introduced in chapter "Decision under Uncertainty" of this volume (just reinterpreting the v function as a probability measure and using Eq. (7)). In the case where the criteria are equally weighted, this dominance $\succsim_2$ reduces to the Lorenz dominance. In fact, when functions F_x, F_y (resp. $\check{F}_x$, $\check{F}_y$) are piecewise linear, testing $\succsim_2$ amounts to comparing the curves at break points in terms of weak Pareto dominance. Note that in the case of equally weighted components, the n break points are in k/n for $k = 1, \ldots, n$. We can then show that $n\check{F}_x^2(k/n) = L_k(x)$ (see Shorrocks 1983; Muliere and Scarsini 1989). This explains why $\succsim_2$ is equivalent to comparing the components of the Lorenz vectors $L_k(x)$ and $L_k(y)$ for all $k \in N$ and therefore, in this case, $\succsim_2$ is nothing else but the Lorenz dominance $\succsim_L$. For this reason $\succsim_2$ can be seen as a generalization of Lorenz dominance in the case of weighted criteria.

4.2 Concordance Relations

The concordance relations are preference relations that are not necessarily transitive, resulting from aggregation rules inspired by the majority voting rules (rules of concordance). In such rules, for each pair of solutions (x, y), we count the number of criteria in favor of x and y respectively, and it is based on this count to decide whether x is better than y. If the criteria do not all have the same weight, we can more generally evaluate the weight of the coalition of criteria in favor of x and against y. This so-called "concordant" coalition is a widely used notion in ELECTRE type methods (see Roy 1985; Roy and Bouyssou 1993; Vincke 1992). There are many variants of these rules, of which we give here some typical examples by assuming that the indices $\phi_i(x, y)$ are constructed as in Eq. (4):

Absolute Concordance

$$\forall (x, y) \in A \times A, \quad c(x, y) = v(\{i \in N : \phi_i(x, y) > 0\}) \tag{9}$$

$$x \succsim y \text{ iff } c(x, y) \geq s \tag{10}$$

where v is a set function defined on 2^N and valued in $[0, 1]$ and $s \in]0,\ 1]$ is an acceptance threshold named *concordance threshold*. The most standard instance of this family of rules is the absolute majority rule obtained for $s = (n + 1)/2$ and $v(E) = |E|$ for all $E \subseteq N$. When we wish to weight the criteria, we can define $v(E) = \sum_{i \in E} w_i$ where w_i represents the weight of criterion i. We may also resort to more general definitions such as $v(E) = \psi(w_1, \ldots, w_n)$ where ψ is an aggregation function.

Relative Concordance

$$x \succsim y \text{ iff } c(x, y) \geq c(y, x) \tag{11}$$

where $c(x, y)$ is defined by Eq. (9). This relation is generally not transitive. However, some instances of this rule are transitive on any set of alternatives. In fact, the transitivity appears for very specific definitions of the notion of importance of criteria. This point has been widely studied in the literature on decision theory, see e.g., Dubois et al. (2003) for a reference in AI.

Absolute Concordance with Veto

$$x \succsim y \quad \text{iff} \quad \begin{cases} c(x, y) \geq s \\ \forall i \in N, y_i - x_i \leq v_i \end{cases} \tag{12}$$

where v_i is the veto threshold that can be defined as the biggest difference of performance $y_i - x_i$ that can be imagined on criterion i and which is still compatible with the preference of x over y. If $y_i - x_i$ exceeds the veto threshold on some criterion i, x cannot be preferred nor be indifferent to y. This condition aims to prevent any compensation phenomenon when comparing two alternatives with very contrasted profiles. It also prevents to compensate a strong weakness with multiple weakly positive points. This principle of non-veto is presented in an absolute concordance rule but could also be inserted in the relative concordance rule. The reader may refer to Roy and Bouyssou (1993), Perny (1998) for more details on this point.

Concordance Rules with Reference Points

Let $p \in \mathbb{R}^n$ be a performance vector used as a reference point to assess and compare the alternatives. A concordance relation with reference point is defined by:

$$x \succsim y \ \text{ iff } \ c(x, p) \geq c(y, p) \tag{13}$$

where $c(x, y)$ is defined by Eq. (9). Using the same notations, we can also introduce the following relation:

$$x \succsim y \ \text{ iff } \ c(p, y) \geq c(p, x). \tag{14}$$

Note that, contrary to the standard concordance relations introduced before (see Eqs. 10–12), the concordance relations with reference point are naturally transitive, which facilitates their use for choice and ranking problems. One can find in Perny and Rolland (2006), Rolland (2013), Bouyssou and Marchant (2013) other examples of concordance rules with reference points, as well as some axiomatic analysis concerning these rules.

When a non-transitive concordance relation is used, the candidates cannot be directly ordered and it is difficult to determine an optimal choice. To overcome the problem, many methods for determining a winner or ranking the alternatives from a non-transitive strict preference relation $\succ$ have been proposed. Here are some examples:

The Net Flow Rule

Rank the candidates according to the Net Flow Score defined as follows:

$$\phi(x) = |\{y \in \mathscr{X} : x \succ y\}| - |\{y \in \mathscr{X} : y \succ x\}|$$

For a choice problem, select the alternatives maximizing the net flow.

Schwartz's Rule

Calculate $\succ^*$ the transitive closure of relation $\succ$. Then define a new strict preference relation $\succ_S$ as follows:

$$x \succ_S y \text{ iff } [x \succ^* y \text{ and not}(y \succ^* x)]$$

By construction relation $\succ_S$ is transitive since it is the asymmetric part of a transitive relation. For a choice problem, select the solutions of $ND(\mathscr{X}, \succ_S)$.

Decision Rules Based on Traces

The traces of a relation $\succ$ are defined by:

$$x \succ^+ y \text{ iff } \forall z \in \mathscr{X} \setminus \{x, y\},\ (y \succ z \Rightarrow x \succ z)$$
$$x \succ^- y \text{ iff } \forall z \in \mathscr{X} \setminus \{x, y\},\ (z \succ x \Rightarrow z \succ y)$$

Both relations $\succ^+$ and $\succ^-$ are transitive, and therefore their intersection too. They can therefore be used to partially order the solutions or to define a set of non-dominated elements, for example by calculating $ND(\mathscr{X}, \succ^+)$ or $ND(\mathscr{X}, \succ^-)$.

5 Decision Models in the AC Approach

5.1 The Weighted Mean

The decision model based on the weighted mean leads to the following definition of preferences:

$$x \succsim y \text{ iff } \sum_{i=1}^{n} w_i x_i \geq \sum_{i=1}^{n} w_i y_i$$

This model is probably the one that most quickly comes to mind when one aggregates performances. Yet it is often unsatisfactory because it provides no control on whether the optimal solutions are balanced or not. By way of illustration, let us consider the following example:

Example 5 A company wants to recruit a technical sales computer. Candidates must complete two interviews, one for the technical skills of the individual, the other

intended to evaluate the business skills. Consider a situation with 4 candidates that received the following grades: $x = (18, 5)$, $y = (4, 19)$, $z = (11, 11)$, $w = (9, 7)$. Candidate w who is Pareto-dominated by candidate z is quickly disqualified. The candidates x and y, who have a significant weakness on one of the two expected skills (score less than or equal to 5), do not seem to be suitable either. As a result, the z candidate seems to be the best compromise between technical and commercial skills. However, it can easily be verified that, whatever the weight vector (w_1, w_2) used, the candidate z will not be the one with the best weighted average, although he is Pareto-optimal. This is due to the fact that (11, 11) lies within the convex hull of the points x, y, z, w in the criterion space, whereas only the points on the boundaries of this convex hull can be obtained by optimizing a weighted sum of the performances.

The above example shows that when a weighted sum is used, we take the risk of eliminating some Pareto-solutions a priori, even before having chosen the weights of the criteria, although such solutions could achieve interesting compromises between the criteria. These well-known limits of the weighted sum justify the interest in other aggregators. A possible generalization of weighted means is provided by quasi-arithmetic means defined by:

$$\psi(x) = f^{-1}\left(\sum_{i=1}^{n} w_i f(x_i)\right)$$

where $f(x)$ is a strictly monotonic function. For instance, the weighted geometric mean is obtained for $f(x) = \ln(x)$, the weighted dual geometric mean when $f(x) = \ln(1 - x)$, the harmonic mean when $f(x) = 1/x$ and the weighted L_p norm for $f(x) = x^p$, $p \in \mathbb{N}$. The next section introduces a more powerful aggregator to explore various types of compromise solutions in the Pareto set.

5.2 The Weighted Tchebycheff Norm

One way to define preferences by a scalarizing function is to measure the distance to a reference point $p \in \mathbb{R}^n$ representing a target performance vector. The idea is to try to be as close as possible to the target on each of the criteria. The quality of a solution can then be defined as its distance to the target in the sense of the Tchebycheff norm (a.k.a. infinite norm).

Let $\lambda \in \mathbb{R}^n_+$ be a weighting vector used in combination with the Tchebycheff norm on the one hand to normalize criterion values when they are expressed on different scales, and on the other hand, to control the importance attached to the different criteria so as to generate compromise solutions with a bias reflecting the value system of the decision maker:

$$\psi(x) = ||\lambda(x - p)||_\infty = \max_{i \in N} \lambda_i |x_i - p_i|$$

A good choice for the reference point p is the *ideal point* $\alpha \in \mathbb{R}^n$ defined by $\alpha_i = \sup_{x \in \mathscr{X}} x_i$ which provides an upper bound of the set of Pareto-optimal vectors. Then, minimizing the Tchebycheff distance amounts to projecting the ideal point on the Pareto set, in a direction controlled by the weights λ_i. Usually the weights λ_i are defined as follows:

$$\lambda_i = \frac{w_i}{\alpha_i - \beta_i} \text{ with } \beta_i = \inf_{x \in \mathscr{X}^*} \{x_i\} \text{ and } \mathscr{X}^* = \{x \in \mathscr{X} : \exists i \in N, x_i = \alpha_i\}$$

The components α_i are obtained by single objective optimization on each component separately; this makes it possible to compute $\mathscr{X}^*$ and then components β_i. The optimization of the parameterized function ψ guarantees that, for any Pareto-optimal solution x, there exists a weighting vector w such that x will be part of the ψ-optimal solutions (Wierzbicki 1986) (in fact to avoid pathological cases and fully benefit from this property, it is better not to define α as the ideal point but as a neighbor point strictly above the ideal on every component). We thus correct the observed defect of the weighted sum since any Pareto-optimal solution can now be obtained by a minimization of ψ with the proper parameters. On the other hand, the optimization of function ψ does not quite guarantee the Pareto-optimality of the solutions obtained because of a drowning effect induced by the maximum. Assume indeed that the reference point is $p = (20, 20)$ and that the two feasible solutions are $x = (4, 2)$ and $y = (18, 2)$ we have $\psi(x) = \psi(y)$. Thus, x could be selected as the best choice while it is Pareto-dominated. To avoid this problem, we introduce an additional term, the weighted sum of the deviations from the ideal point multiplied by an arbitrarily small quantity $\epsilon > 0$; this weighted sum comes to play the role of a second criterion considered lexicographically after that of Tchebycheff to discriminate between equivalent solutions in terms of distance to the ideal point. We therefore arrive at the following aggregation function to be minimized:

$$t(x) = \max_{i \in N} \frac{w_i(\alpha_i - x_i)}{\alpha_i - \beta_i} + \varepsilon \sum_{i=1}^{n} \frac{w_i(\alpha_i - x_i)}{\alpha_i - \beta_i} \tag{15}$$

By minimizing function t defined by Eq. (15), we make sure to generate only Pareto-optimal solutions. Moreover, if ε is chosen to be small enough, the practical possibility of reaching any Pareto-optimal solution is preserved (Wierzbicki 1986). This dual quality justifies the use of this aggregator in optimization to explore the Pareto-optimal solutions in various directions controlled by the w vector. It is therefore widely used in interactive exploration methods (Steuer and Choo 1983; Steuer 1986; Wierzbicki 1999). This aggregator can of course be used to define a preference over the set of solutions (by proximity to the ideal point α) by setting: $x \succsim y$ iff $t(x) \leq t(y)$. An application of this decision model to multiobjective state space search is proposed in Galand and Perny (2006).

5.3 The Ordered Weighted Average (OWA)

The Ordered Weighted Average (Yager 1988) in an aggregation function enabling to weight the performances x_i according to their rank, once reordered with permutation σ such that $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$. Formally, the OWA aggregation is defined by:

$$OWA(x) = \sum_{i=1}^{n} w_i x_{\sigma(i)} \tag{16}$$

OWA is a symmetric function because the weights do not relate to the components of x but to those of the reordered vector. This family includes the minimum, maximum, median operations and all order statistics[2] as special cases, by using a weighting vector whose all but one components are zero, the remaining one being 1. It is also widely used in fair optimization as a linear extension of the Lorenz dominance introduced in the previous section. Indeed, noting that $x_{\sigma(i)} = L_i(x) - L_{i-1}(x)$ for $i > 1$, we have:

$$OWA(x) = \sum_{i=1}^{n-1} (w_i - w_{i+1}) L_i(x) + w_n L_n(x) \tag{17}$$

One can see that, if coefficients w_i are positive and chosen to decrease when i increases, OWA is a linear combination with positive coefficients of the components of the Lorenz vector. Therefore the weak order defined by:

$$x \succsim y \text{ iff } OWA(x) \geq OWA(y)$$

is a linear extension of Lorenz dominance, that is, $x \succsim_L y \Rightarrow OWA(x) \geq OWA(y)$. Thus OWA used with strictly decreasing weights w_i is an aggregator allowing to promote balanced solutions. Indeed, due to Eq. (17), an OWA-optimal solution is necessarily optimal in Lorenz's sense and there is therefore no Pigou-Dalton transfer allowing to reduce inequalities (due to Theorem 1). Another way to present the treatment of inequalities by an OWA is to consider Eq. (16) and note that by choosing decreasing weights, one assigns the greatest weight to the least satisfied criterion, then a little less importance to the second least satisfied criterion and so on. Of course, comparing or sorting performances from several criteria only makes sense when they are expressed on the same scale (if not, they must first be re-encoded using utility functions). To give an example of the use of OWA, if one wishes to compare the vectors $x = (10, 5, 15)$ and $y = (10, 12, 8)$ using an OWA with $w = (3/6, 2/6, 1/6)$, we obtain $OWA(x) = 50/6 = 8.33$ while $OWA(y) = 52/6 = 8.66$, therefore $y \succ x$. The OWA weights are used to control the attitude of the Decision Maker towards fairness. They can be elicited from preference statements provided by the Decision Maker (see e.g., Benabbou et al. 2015; Bourdache and Perny 2017).

[2]The order statistic of rank k of a sample of values is equal to the kth smallest value.

The OWA operator is widely used in social choice theory as a measure of inequality under the name of the generalized Gini social evaluation function (Weymark 1981). It is also used to aggregate fuzzy set membership functions (see Yager 1988). In artificial intelligence, it often appears in fair optimization problems or allocation problems of indivisible goods (Bouveret and Lang 2005; Golden and Perny 2010; Lesca and Perny 2010; Lesca et al. 2018), and also in voting rules (Goldsmith et al. 2014; Elkind and Ismaili 2015; Skowron et al. 2016; García-Lapresta and Martínez-Panero 2017). Note that, although OWA is not a linear function of criterion values, the optimization of an OWA function can be done by linear programming (provided that the criteria and the constraints defining the admissible solutions are linear in the decision variables), for more details see Ogryczak and Sliwinski (2003), Chassein and Goerigk (2015).

5.4 The Weighted OWA (WOWA)

As pointed out in the previous subsection, one characteristic of the OWA is to be a symmetric aggregation function. This property, which seems natural when the criteria represent individual points of view in a collective decision problem, may not be desired in multicriteria decision problems, particularly when certain criteria are considered more important than others. We then consider now a weighted extension of the OWA aggregator, the initial weights involved in the OWA definition being only used to control the importance attached to good and bad performances. This weighted OWA is know in the literature under the name of WOWA (Torra 1997); it uses a vector $p \in \mathbb{R}^n$ of criteria weights and takes the following form:

$$\begin{aligned} WOWA(x) &= \sum_{i=1}^{n} \left[x_{\sigma(i)} - x_{\sigma(i-1)}\right] \varphi\left(\sum_{k=i}^{n} p_{\sigma(k)}\right) \\ &= \sum_{i=1}^{n} \left[\varphi\left(\sum_{k=i}^{n} p_{\sigma(k)}\right) - \varphi\left(\sum_{k=i+1}^{n} p_{\sigma(k)}\right)\right] x_{\sigma(i)} \end{aligned}$$

where σ is the permutation reordering the components of x by increasing order, i.e., $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$; function φ is strictly increasing and such that $\varphi(0) = 0$. The induced preference is then defined by: $x \succsim y$ iff $WOWA(x) \geq WOWA(y)$.

This formulation is known as *Yaari's model* in decision under risk because it has been initially introduced and axiomatically justified in this context (Yaari 1987) (the weights p_i being interpreted as the probabilities of the states of nature, see the RDU model in chapter "Decision under Uncertainty" of this volume. Its importation into a multicriteria decision-making context is more recent and due to Torra (1997) who arrives at an identical formulation starting from an OWA. The specificity of the construction proposed by Torra lies in the definition of the φ function. It is constructed

from the weights $(w_1, \ldots, w_n)$ of an OWA so that $\varphi(i/N) = \sum_{k=1}^{i} w_{n-k+1}$, which allows to weight the criteria while controlling the importance given to good and bad performances as in an OWA. The WOWA is therefore constructed from the two weighting vectors p and w. By way of illustration, consider the following example:

Example 6 If we want to compare vectors $x = (10, 5, 15)$ and $y = (10, 12, 8)$ with a WOWA characterized by weights $w = (3/6, 2/6, 1/6)$ and criteria weights $p = (3/6, 1/6, 2/6)$, we use a piecewise linear φ function taking the following values at the key points: $\varphi(0) = 0, \varphi(1/3) = 1/6, \varphi(2/3) = 1/2, \varphi(1) = 1$. These values can be completed by linear interpolation to obtain:

x	0	1/6	2/6	3/6	4/6	5/6	1
$\varphi(x)$	0	1/12	1/6	2/6	1/2	3/4	1

Then we get:

$$WOWA(x) = 5 + (10 - 5)\varphi(5/6) + (15 - 10)\varphi(2/6) = 9.58$$
$$WOWA(y) = 8 + (10 - 8)\varphi(4/6) + (12 - 10)\varphi(1/6) = 9.16$$

Therefore x is preferred to y. Here the fact that the third criterion is more important than the second gives an advantage to x which is sufficient to compensate the inegalitarian side of this solution. We can verify that it would suffice to be more demanding on the equity requirement by choosing the weighting vector $w = (0.8, 0.25, 0.05)$ so that the preference is reversed in favor of y. In this case, we would indeed have:

$$\varphi(0) = 0, \varphi(1/3) = 0.05, \varphi(2/3) = 0.3, \varphi(1) = 1$$

and, by completing using linear interpolation:

x	0	1/6	2/6	3/6	4/6	5/6	1
$\varphi(x)$	0	0.025	0.05	0.175	0.3	0.65	1

In this case we obtain:

$$WOWA(x) = 5 + (10 - 5)\varphi(5/6) + (15 - 10)\varphi(2/6) = 8.5$$
$$WOWA(y) = 8 + (10 - 8)\varphi(4/6) + (12 - 10)\varphi(1/6) = 8.65$$

and this time we get y is preferred to x.

This example clearly shows how the two vectors of weights interact, one to control the weights of the criteria and the other to control the fairness requirement. It may also be noted that if weights w_i decrease when i increases (to favor solutions that equitably share performance among criteria) then the φ function is convex and the function WOWA is concave. In this case, it can be proven that WOWA is monotone increasing with dominance $\succsim_2$ introduced in Eq. (7), which means that:

$x \succsim_2 y \Rightarrow WOWA(x) \geq WOWA(y)$. Thus, we obtain a weighted version of the result concerning the monotony of OWA with respect to Lorenz dominance.

In practice, nothing prohibits the use of non-decreasing weights and we then obtain an aggregator that offers a diversity of behaviors in the aggregation. For example, if we use weights w_i increasing with i, we will have a concave φ and a convex WOWA. In maximization, we give a premium to the solutions with an imbalanced profile alternating good and bad performances. We will return to the control of WOWA in the more general framework of the Choquet integral. We can also notice that if $\phi(x) = x$ then WOWA reduces to a simple weighted average with weights p_i. Moreover, if $p_i = 1/n$ then WOWA reduces to an OWA, which is quite natural. Finally, it should be noted that WOWAs are very useful for equitable optimization when we want to associate weights with agents (exogenous rights), especially since when the w_i weights are decreasing when i increases, the optimization of WOWA can be done easily by linear programming using reformulations close to those needed to linearize an OWA (see Ogryczak and Sliwinski 2007 for more details).

Note that function φ does not necessarily have to be constructed from vectors p and w, it can be directly defined as a convex function to convey an idea of fairness (in maximization, the more function φ is convex the greater the requirement of fairness). On the contrary a concave φ function would exhibit a preference for contrasted profiles. In minimization problems, this is just the opposite and fairness is modeled by a concave φ function. The elicitation of φ can be performed incrementally using preference queries on specific pairs of alternatives, see e.g., Perny et al. (2016).

5.5 *The Choquet Integral*

The Choquet integral is one of the most sophisticated scalarizing function used for multicriteria aggregation (Choquet 1953; Grabisch 1996; Marichal 2000a; Marichal and Roubens 2000; Grabisch and Labreuche 2008; Grabisch et al. 2009). It includes both weighted sums, OWA and WOWA as special cases. It is defined from a set function, namely the *capacity* allowing to assign a weight to any subset of criteria $E \subseteq N$. More precisely, a capacity is defined as follows:

Definition 5 A capacity is a set function $v : N \rightarrow [0, 1]$ such that $v(\emptyset) = 0, v(N) = 1$ and $\forall A, B \subseteq N, A \subseteq B \Rightarrow v(A) \leq v(B)$.

The capacity is said to be:

- *concave* or *sub-modular* if $v(A \cup B) + v(A \cap B) \leq v(A) + v(B)$ for all $A, B \subseteq N$,
- *additive* if $v(A \cup B) + v(A \cap B) = v(A) + v(B)$ for all $A, B \subseteq N$,
- *convex* or *super-modular* if $v(A \cup B) + v(A \cap B) \geq v(A) + v(B)$ for all $A, B \subseteq N$.

For any given capacity v, the Choquet integral of a vector $x \in \mathbb{R}^n$ is defined by:

$$\begin{aligned} C_v(x) &= \sum_{i=1}^{n} \left[x_{\sigma(i)} - x_{\sigma(i-1)}\right] v(X_{\sigma(i)}) \\ &= \sum_{i=1}^{n} \left[v(X_{\sigma(i)}) - v(X_{\sigma(i+1)})\right] x_{\sigma(i)} \end{aligned}$$

where σ is the permutation reordering the components of x by increasing order, i.e., $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$, $x_{\sigma(0)} = 0$, $X_{\sigma(i)} = \{\sigma(i), \sigma(i+1), \ldots, \sigma(n)\}$ for $i = 1, \ldots, n$, and $X_{\sigma(n+1)} = \emptyset$. The preference relation associated to the Choquet integral is therefore defined by: $x \succsim y$ iff $C_v(x) \geq C_v(y)$.

We remark that WOWA is only a special case of Choquet integral in which the capacity v is defined by $v(E) = \varphi(\sum_{i \in E} p_i)$ for a weighting vector $(p_1, \ldots, p_n)$. It can be shown that v is convex (resp. concave) when φ is convex (resp. concave). The Choquet integral C_v can account for various behaviors depending on the choice of the capacity. When using an additive capacity, i.e., $v(E) = \sum_{i \in E} p_i$, the Choquet integral is reduced to the weighted sum and does not offer particular descriptive possibilities. However, we can describe much richer classes of preferences with concave, convex or other more general capacities. For example, if we use a convex capacity in maximization, the well-balanced profiles will be favored, and this will be the opposite if we choose a concave capacity, as shown by the following example:

Example 7 Let us consider an example with 3 criteria, i.e., $N = 1, 2, 3$ and two capacities v_1 and v_2 defined in the following table:

	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
v_1	0	0.2	0.1	0.3	0.45	0.5	0.65	1
v_2	0	0.35	0.5	0.55	0.7	0.9	0.8	1

One can easily check that v_1 is convex and v_2 is concave. Coming back to the comparison of vectors $x = (10, 5, 15)$ and $y = (10, 12, 8)$ we have:

$$\begin{aligned} C_{v_1}(x) &= 5 + (10-5)v_1(\{1,3\}) + (15-10)v_1(\{3\}) = 9 \\ C_{v_1}(y) &= 8 + (10-8)v_1(\{1,2\}) + (12-10)v_1(\{2\}) = 9.1 \\ C_{v_2}(x) &= 5 + (10-5)v_2(\{1,3\}) + (15-10)v_2(\{3\}) = 12.25 \\ C_{v_2}(y) &= 8 + (10-8)v_2(\{1,2\}) + (12-10)v_2(\{2\}) = 10.4 \end{aligned}$$

One can see that with v_1 solution y is preferred to x whereas with v_2 solution x is preferred to y.

More precisely, the use of a convex capacity in a Choquet integral conveys an idea of equity due to the following property (Chateauneuf and Tallon 1999):

Proposition 1 *If v is convex then $\forall x^1, x^2, \ldots, x^p \in \mathbb{R}^n$, $\forall k \in \{1, 2, \ldots, p\}$ and $\forall i \in \{1, 2, \ldots, p\}, \lambda_i > 0$ such that $\sum_{i=1}^{p} \lambda_i = 1$ we have:*

$$C_v(x^1) = C_v(x^2) = \cdots = C_v(x^p) \Rightarrow C_v\left(\sum_{i=1}^{p} \lambda_i x^i\right) \geq C_v(x^k).$$

This proposition means that if several solutions x^i, $i = 1, \ldots, p$ are indifferent for the decision maker, she we will prefer a solution whose performance vector is a convex combination of the x^i's to all these solutions. For example, if one is indifferent between two performance vectors (0, 20) and (20, 0), it is expected that a solution such as (10, 10) which corresponds to the average of the two preceding vectors is preferable. Obviously, the reverse preference is obtained with a concave capacity. The Choquet integral for a convex capacity is a concave function and conversely the Choquet integral for a concave capacity is a convex function (Lovász 1983). Another useful formulation of the Choquet integral is to express it as a function of the Möbius masses associated with capacity v. These masses are defined in this way:

Definition 6 To any capacity v defined on 2^N one can associate another set function on 2^N named *Möbius inverse* and defined by:

$$\forall A \subseteq N, m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} v(B) \tag{18}$$

Then, v can be recovered from its Möbius inverse m as follows:

$$\forall A \subseteq N, v(A) = \sum_{B \subseteq A} m(B) \tag{19}$$

Using the Möbius inverse m associated with v, the Choquet integral can be rewritten as follows:

$$C_v(x) = \sum_{B \subseteq N} m(B) \min_{i \in B} x_i \tag{20}$$

This highlights another interpretation of the Choquet integral as a linear aggregator in a new multidimensional space of size 2^n where n is the initial number of criteria. The components of a vector in this space correspond to quantities $\min_{i \in B} x_i$ for all subsets $B \subset N$. Whether we use the initial formulation of the Choquet integral or the one that involves the Möbius masses, we may be concerned about the presence of 2^n parameters to characterize the importance of the criteria and their interaction. Fortunately, in many practical cases, there is no need to consider all these coefficients, we can resort to *k-additive* capacities for some $k < n$, where k-additivity is defined as follows (Grabisch 1996; Dubois and Prade 1997):

Definition 7 A capacity v is said to be k-additive if its Möbius inverse equals zero for any subset $A \subseteq N$ such that $|A| > k$, and if $m(A) \neq 0$ for some A such that $|A| = k$.

If $k = 1$ we obtain an additive capacity. The $k-$additive capacities for small values of k greater than 1 are very useful in practice because they offer sufficient expressiveness to model positive or negative interactions between criteria while involving a fairly small number of parameters. For example, when $k = 2$, the capacity is completely characterized by only $(n^2 + n)/2$ coefficients (a Möbius mass for each singleton and each pair), which enables to model the following interactions between pairs of criteria:

- positive interaction: $m(\{i, j\}) > 0$ and therefore $v(\{i, j\}) > v(\{i\}) + v(\{j\})$
- no interaction: $m(\{i, j\}) = 0$ and therefore $v(\{i, j\}) = v(\{i\}) + v(\{j\})$
- negative interaction: $m(\{i, j\}) < 0$ and therefore $v(\{i, j\}) < v(\{i\}) + v(\{j\})$

Moreover, with a 2-additive capacity one obtains from Eq. (20) a very compact expression for the Choquet integral:

$$C_v(x) = \sum_i m_i x_i + \sum_{i>j} m_{ij} \min\{x_i, x_j\}$$

As with OWA and WOWA, the search for a solution maximizing $C_v(x)$ can be performed by linear programming in the case where v is convex (Lesca and Perny 2010). In the general case, it is more delicate but some efficient linearizations exist for some class of Möbius representations (Lesca et al. 2013). For more details on Choquet integrals, interaction indices and set functions in multicriteria analysis, the reader should refer to Grabisch (1996, 2016), Grabisch et al. (2009). For the elicitation of the capacity in the Choquet integral, the reader should refer to Grabisch et al. (1995), Marichal and Roubens (2000), Fallah Tehrani et al. (2012), Hüllermeier and Fallah Tehrani (2013), Benabbou et al. (2017).

The Choquet integral is used in various domains of artificial intelligence. For example, in machine learning, the use of Choquet integrals provides higher predictive capacities than linear models, while offering measures for quantifying the importance of individual predictor variables and the interaction between groups of variables (Fallah Tehrani et al. 2012). Moreover, in recommender systems (Beliakov et al. 2015), the advantage provided by Choquet integrals is to allow positive and negative synergies between criteria, with enhanced descriptive and prescriptive possibilities. Similarly, in multiagent decision making (Dubus et al. 2009), the Choquet integral is used to aggregate individual preferences using a possibly non-additive measure of the importance of agent coalitions, which allows one to model various notions of social welfare. In information fusion (Torra and Narukawa 2007), the use of the Choquet integral allows one to model positive or negative reinforcements among sets of observations. Finally, in multiobjective state-space search (Galand and Perny 2007), the use of Choquet integrals allows one to find compromise solutions that could not be obtained using linear aggregators.

5.6 The Sugeno Integral

The Sugeno integral (Sugeno 1974; Dubois et al. 1998; Marichal 2000b; Dubois et al. 2001a; Grabisch and Labreuche 2008; Couceiro et al. 2012) can be seen as a qualitative counterpart of the Choquet integral. In some cases, performance and capacity are expressed on a common ordinal scale. In the presence of such information, one cannot reasonably use the previous criteria which call for the cardinal properties of performance and importance indices (weight, capacity). A natural alternative is then to consider the Sugeno integral which reads:

$$S_v(x) = \max_{i \in N} \min\{x_{\sigma(i)}, v(X_{\sigma(i)})\}$$

where σ is the permutation reordering the components of x by increasing order, i.e., $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$, $X_{\sigma(i)} = \{\sigma(i), \sigma(i+1), \ldots, \sigma(n)\}$ for $i = 1, \ldots, n$. The resulting overall preference relation is therefore:

$$x \succsim y \text{ iff } S_v(x) \geq S_v(y).$$

This general aggregator has been introduced by Sugeno (1974) in fuzzy sets theory and imported into decision theory under uncertainty where its use was axiomatically justified (Dubois et al. 1998). The Sugeno integral can also be used in multicriteria decision making because the proposed axioms can easily be transposed. When v is a measure of possibility over N defined by $v(A) = \max\{\pi_i, i \in A\}$, $(\pi_1, \ldots, \pi_n)$ playing the role of ordinal weights (positive coefficients such that $\max\{\pi_i, i \in A\} = 1$), the Sugeno integral is nothing else but a *weighted maximum* defined by:

$$wmax(x) = \max_{i \in N} \min\{x_i, \pi_i\}$$

When v is a necessity measure defined on N by $v(A) = 1 - \max\{\pi_i, i \notin A\}$, $(\pi_1, \ldots, \pi_n)$ playing the role of ordinal possibilistic weights, the Sugeno integral takes the particular form of a *weighted minimum* defined by:

$$wmin(x) = \min_{i \in N} \max\{x_i, 1 - \pi_i\}$$

The weighted max operator reflects an optimistic view which consists of valuing the existence of at least one good performance on an important criterion. The weighted min reflects a more pessimistic view which consists of assessing the extent to which no important criterion exists on which the alternative under consideration performs poorly. These two models as well as the Sugeno integral have been studied in depth in a decision-making framework, see e.g., Dubois and Prade (1995), Dubois et al. (2001b).

6 Conclusion

The models presented in this chapter provide an overview of the main aggregators used to take multiple points of view into account, whether in multi-criteria decision-making or collective decision-making (see also chapter "Collective Decision Making" of this volume). Most of them are widely used in artificial intelligence. In multicriteria and collective decision-making, the CA approach is very present through ordinal methods of aggregation derived from social choice theory and voting procedures. However, the AC approach remains the most widespread, whether in multi-criteria decision-making to determine specific trade-offs in the Pareto set (by optimizing a scalarizing function), or in collective decision-making to determine a fair Pareto-optimal solution. Concerning this second approach, the reader wishing to obtain technical complements on the aggregation functions and their properties may refer to Grabisch et al. (2009).

Whether they fall under the CA or AC approach, the models presented in this chapter can also be used in decision-making under uncertainty, when trying to evaluate and compare acts in the sense of Savage. If we consider situations in which uncertainty is represented by a finite set of possible states $S = \{s_1, \ldots, s_n\}$, it appears that the different states act as different criteria to evaluate the possible acts. This explains why several models introduced in this chapter may also be used for decision under uncertainty. These criteria can even be generalized in the case of a continuous set of states of nature. The next chapter (chapter "Decision under Uncertainty" of this volume) is precisely intended to present in the most general case the decision models used under uncertainty and risk.

The main theoretical axes that still need to be developed in multicriteria analysis relate to the axiomatic justification of existing models (the results to characterize the preferences that can be represented by a particular model do not always exist in multicriteria analysis, even if neighboring results sometimes exist for decision making under uncertainty), the development of decision models with increased descriptive power in the presence of rich information, and the development of ordinal or partial aggregation methods in the presence of poor information. From a more operational point of view, the main challenges of multicriteria decision theory are to propose efficient methods for eliciting or learning the parameters of the models they propose (for example for recommendation systems) and, on the other hand, to develop efficient algorithms for determining preferred solutions in combinatorial problems (preference-based search). The combinatorial nature of the space of feasible solutions precludes the use of any explicit enumeration method to compare solutions. The search for preferred solutions necessarily involves the development of implicit enumeration methods, but the optimization problems that need to be solved are all the more difficult as the models are sophisticated. Decision theory, by producing models that are always richer to account for various decision-making behaviors, is therefore a source of permanent challenges for computer scientists. These aspects of elicitation and computation are widely studied in artificial intelligence and are the subject of numerous recent contributions to algorithmic decision theory.

References

Arrow K (1951) Social choice and individual values. Wiley, New York

Bazgan C, Hugot H, Vanderpooten D (2009) Implementing an efficient FPTAS for the 0–1 multi-objective knapsack problem. Eur J Oper Res 198(1):47–56

Beliakov G, Calvo T, James S (2015) Aggregation functions for recommender systems. Recommender systems handbook, pp 777–808

Benabbou N, Gonzales C, Perny P, Viappiani P (2015) Minimax regret approaches for preference elicitation with rank-dependent aggregators. EURO J Decis Process 3(1–2):29–64

Benabbou N, Perny P, Viappiani P (2017) Incremental elicitation of choquet capacities for multicriteria choice, ranking and sorting problems. Artif Intell J 246:152–180

Borda J-C (1781) Mémoire sur les élections au scrutin. Comptes rendus de l'Acadmie des sciences. Translated by Alfred de Grazia as Mathematical derivation of a election system. Isis, vol 44, pp 4251

Bourdache N, Perny P (2017) Anytime algorithms for adaptive robust optimization with OWA and WOWA. In: 5th international conference on algorithmic decision theory (ADT 2017). Lecture notes in computer science, vol 10576. Springer, Luxembourg, pp 93–107

Bouveret S, Lang J (2005) Efficiency and envy-freeness in fair division of indivisible goods. In: Proceedings of international joint conference on artificial intelligence (IJCAI'05)

Bouyssou D, Marchant T (2013) Multiattribute preference models with reference points. Eur J Oper Res 229(2):470–481

Charnes A, Cooper W, Ferguson R (1955) Optimal estimation of executive compensation by linear programming. Manag Sci 1:138–151

Chassein A, Goerigk M (2015) Alternative formulations for the ordered weighted averaging objective. Inf Process Lett 115(6):604–608

Chateauneuf A, Tallon J-M (1999) Diversification, convex preferences and non-empty core in the Choquet expected utility model. Econ Theory 19(3):509–523

Chong KM (1976) An introduction theorem for rearrangements. Can J Math 28:154–160

Choquet G (1953) Theory of capacities. Annales de l'Institut Fourier 5:131–295

Couceiro M, Dubois D, Prade H, Waldhauser T (2012) Decision-making with Sugeno integrals: DMU versus MCDM. In: Proceedings of European conference on artificial intelligence (ECAI'12), pp 288–293

de Condorcet M (1785) Essai sur lapplication de lanalyse la probabilit des dcisions rendues la pluralit des voix. Imprimerie Royale, Paris

Diakonikolas I, Yannakakis M (2008) Succinct approximate convex Pareto curves. In: Proceedings of ACM-SIAM symposium on discrete algorithms (SODA'08), pp 74–83

Dubois D, Fargier H, Perny P (2003) Qualitative decision theory with preference relations and comparative uncertainty: an axiomatic approach. Artif Intell 148(1):219–260

Dubois D, Marichal J, Prade H, Roubens M, Sabbadin R (2001a) The use of the discrete Sugeno integral in decision-making: a survey. Int J Uncertain, Fuzziness Knowl-Based Syst 9(5):539–561

Dubois D, Prade H (1995) Possibility theory as a basis of qualitative decision theory. In: Proceedings of international joint conference on artificial intelligence (IJCAI'95), pp 1924–1930

Dubois D, Prade H (1997) K-order additive discrete fuzzy measures and their representation. Fuzzy Sets Syst 92:167–189

Dubois D, Prade H, Sabbadin R (1998) Qualitative decision theory with Sugeno integrals. In: Proceedings conference on uncertainty in artificial intelligence (UAI'98), pp 121–128

Dubois D, Prade H, Sabbadin R (2001b) Decision-theoretic foundations of qualitative possibility theory. Eur J Oper Res 128:459–478

Dubus J, Gonzales C, Perny P (2009) Choquet optimization using GAI networks for multiagent/multicriteria decision-making. In: Proceedings of the international conference on algorithmic decision theory, pp 377–389

Elkind E, Ismaili A (2015) Owa-based extensions of the Chamberlin-Courant rule. In: Proceedings of the 4th international conference on algorithmic decision theory, pp 486–502

Fallah Tehrani A, Cheng W, Dembczynski K, Hüllermeier E (2012) Learning monotone nonlinear models using the Choquet integral. Mach Learn 89(1–2):183–211
Fodor J, Roubens M (1994) Fuzzy preference modelling and multicriteria decision support. Kluwer Academic, Cambridge
Galand L, Perny P (2006) Search for compromise solutions in multiobjective state space graphs. In: Proceeding of the 17th European conference on artificial intelligence, pp 93–97
Galand L, Perny P (2007) Search for Choquet-optimal paths under uncertainty. In: UAI 2007, Proceedings of the twenty-third conference on uncertainty in artificial intelligence, Vancouver, BC, Canada, 19–22 July 2007, pp 125–132
García-Lapresta JL, Martínez-Panero M (2017) Positional voting rules generated by aggregation functions and the role of duplication. Int J Intell Syst 32(9):926–946
Geoffrion A, Dyer J, Feinberg A (1973) An interactive approach for multicriteria optimization with an application to the operation of an academic department. Manag Sci 19:357–369
Golden B, Perny P (2010) Infinite order Lorenz dominance for fair multiagent optimization. In: Proceedings of international conference on autonomous agents and multiagent systems (AAMAS'10), pp 383–390
Goldsmith J, Lang J, Mattei N, Perny P (2014) Voting with rank dependent scoring rules. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence, pp 698–704
Grabisch M (1996) The application of fuzzy integrals in multicriteria decision making. Eur J Oper Res 89(3):445–456
Grabisch M (2016) Set functions, games and capacities in decision making. Springer, Berlin
Grabisch M, Labreuche C (2008) A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. 4OR 6(1):1–44
Grabisch M, Marichal J, Mesiar R (2009) Aggregation functions. Cambridge University, Cambridge
Grabisch M, Nguyen H, Walker E (1995) Fundamentals of uncertainty calculi, with applications. Encyclopedia of mathematics and its applications. Kluwer Academic, Cambridge
Gbor Z, Kalmr Z, Szepesvri C (1998) Multicriteria reinforcement learning. In: Proceedings of international conference of machine learning (ICML'98)
Hansen P (1980) Bicriterion path problems. In: Fandel G, Gal T (eds) International conference on multiple criteria decision making (MCDM'80)
Hüllermeier E, Fallah Tehrani A (2013) Efficient learning of classifiers based on the 2-additive Choquet integral. Computational intelligence in intelligent data analysis. Studies in computational intelligence, vol 445, pp 17–29
Laumanns M, Thiele L, Deb K, Zitzler E (2002) Combining convergence and diversity in evolutionary multiobjective optimization. Evol Comput 10(3):263–282
Lesca J, Minoux M, Perny P (2013) Compact versus noncompact LP formulations for minimizing convex Choquet integrals. Discret Appl Math 161(1–2):184–199
Lesca J, Minoux M, Perny P (2018) The fair OWA one-to-one assignment problem: NP-hardness and polynomial time special cases. To appear in algorithmica
Lesca J, Perny P (2010) LP solvable models for multiagent fair allocation problems. In: Proceedings of European conference on artificial intelligence (ECAI'10), pp 387–392
Lovász L (1983) Submodular functions and convexity. In: Bachem A, Grötschel M, Korte B (eds) Mathematical programming, the state of the art, pp 235–257
Marichal J-L (2000a) An axiomatic approach of the discrete Choquet integral as a tool to aggregate interacting criteria. IEEE Trans Fuzzy Syst 8(6):800–807
Marichal J-L (2000b) On Sugeno integral as an aggregation function. Fuzzy Sets Syst 114(3)
Marichal J-L, Roubens M (2000) Determination of weights of interacting criteria from a reference set. Eur J Oper Res 124(3):641–650
Marshall W, Olkin I (1979) Inequalities: theory of majorization and its applications. Academic, London
Moulin H (1988) Axioms of cooperative decision making. Monograph of the econometric society. Cambridge University, Cambridge

Muliere P, Scarsini M (1989) A note on stochastic dominance and inequality measures. J Econ Theory 49:314–323
Ogryczak W, Sliwinski T (2003) On solving linear programs with the ordered weighted averaging objective. Eur J Oper Res 148(1):80–91
Ogryczak W, Sliwinski T (2007) On optimization of the importance weighted OWA aggregation of multiple criteria. In: Proceedings of international conference on computational science and its applications (ICCSA'7). Lecture notes in computer science, vol 4705, pp 804–817
Papadimitriou C, Yannakakis M (2000) On the approximability of trade-offs and optimal access of web sources. In: Proceedings of IEEE symposium on foundations of computer science (FOCS'00), pp 86–92
Pareto V (1906) Manuale di Economia Politica. Piccola Biblioteca Scientifica, Milan. Traduit en anglais par Ann S. Schwier (1971) Manual of political economy. MacMillan, London
Perny P (1998) Multicriteria filtering methods based on concordance and non-discordance principles. Ann Oper Res 80:137–165
Perny P, Rolland A (2006) Reference-dependent qualitative models for decision making under uncertainty. In: Proceedings of European conference on artificial intelligence (ECAI'06), pp 422–426
Perny P, Roy B (1992) The use of fuzzy outranking relations in preference modelling. Fuzzy Sets Syst 49:33–53
Perny P, Spanjaard O (2008) Near admissible algorithms for multi objective search. In: Proceedings of European conference on artificial intelligence (ECAI'08), pp 490–494
Perny P, Viappiani P, Boukhatem A (2016) Incremental preference elicitation for decision making under risk with the rank-dependent utility model. In: Proceedings of UAI'16, pp 597–606
Pirlot M, Vincke P (1997) Semiorders properties, representations, applications. Kluwer Academic, Cambridge
Rolland A (2013) Reference-based preferences aggregation procedures in multi-criteria decision making. Eur J Oper Res 225(3):479–486
Roubens M, Vincke P (1985) Preference modelling. Springer, Berlin
Roy B (1985) Mthodologie multicritre d'aide la dcision. Economica
Roy B, Bouyssou D (1993) Méthodologie multicritre d'aide la dcision : mthodes et cas. Economica
Schiex T, Fargier H, Verfaillie G (1995) Valued constraint satisfaction problems: hard and easy problems. In: Proceedings of international joint conference on artificial intelligence (IJCAI'95), pp 631–639
Sen A (1986a) Social choice theory. In: Arrow K, Intriligator M (eds) Handbook of mathematical economics, vol 3. North-Holland, Amsterdam, pp 1073–1181
Sen AK (1986b) Social choice theory. In: Arrow K, Intrilligator M (eds) Handbook of mathematical economics. Elsevier Sciences, North-Holland, pp 1073–1181
Shorrocks A (1983) Ranking income distributions. Economica 50:3–17
Skowron P, Faliszewski P, Lang J (2016) Finding a collective set of items: from proportional multirepresentation to group recommendation. Artif Intell J 241:191–216
Steuer R, Choo E-U (1983) An interactive weighted Tchebycheff procedure for multiple objective programming. Math Program 26:326–344
Steuer RE (1986) Multiple criteria optimization: theory, computation and application. Wiley, New York
Stewart BS, White III CC (1991) Multiobjective A*. J Assoc Comput Mach 38(4):775–814
Sugeno M (1974) Theory of fuzzy integrals and its applications. PhD thesis, Tokyo Institute of Technology
Torra V (1997) The weighted OWA operator. Int J Intell Syst 12:153–166
Torra V, Narukawa Y (2007) Modeling decisions - information fusion and aggregation operators. Springer, Berlin
Vincke P (1992) Multicriteria decision aid. Wiley, New Jersey
Weymark J (1981) Generalized Gini inequality indices. Math Soc Sci 1:409–430

White D (1982) Multi-objective infinite-horizon discounted Markov decision processes. J Math Anal Appl 89:639–647
Wierzbicki A (1986) On the completeness and constructiveness of parametric characterizations to vector optimization problems. OR Spektrum 8:73–87
Wierzbicki AP (1999) Reference point approaches. In: Gal T, Stewart T, Hanne T (eds) Multicriteria decision making: advances in MCDM models, algorithmes, and applications. Kluwer Academic, Cambridge
Yaari M (1987) The dual theory of choice under risk. Econometrica 55:95–115
Yager R (1988) On ordered weighted averaging aggregation operators in multi criteria decision making. In: IEEE transaction on systems man and cybernetics, vol 18, pp 183–190

Decision Under Uncertainty

Christophe Gonzales and Patrice Perny

Abstract The goal of this chapter is to provide a general introduction to decision making under uncertainty. The mathematical foundations of the most popular models used in artificial intelligence are described, notably the *Expected Utility model* (EU), but also new decision making models, like *Rank Dependent Utility* (RDU), which significantly extend the descriptive power of EU. Decision making under uncertainty naturally involves risks when decisions are made. The notion of risk is formalized as well as the attitude of agents w.r.t. risk. For this purpose, probabilities are often exploited to model uncertainties. But there exist situations in which agents do not have sufficient knowledge or data available to determine these probability distributions. In this case, more general models of uncertainty are needed and this chapter describes some of them, notably *belief functions*. Finally, in most artificial intelligence problems, sequences of decisions need be made and, to get an optimal sequence, decisions must not be considered separately but as a whole. We thus study at the end of this chapter models of sequential decision making under uncertainty, notably the most widely used graphical models.

1 Introduction

Uncertainty and, more generally, decision making under uncertainty are central in artificial intelligence (AI). Indeed, even though AI addresses a wide range of problems, most of them involve to some extent uncertainties. This is the case, for instance, in diagnosis (Franklin et al. 1991; Jensen et al. 2001), prediction (Conati et al. 1997; Horvitz et al. 1998), robotics (Argall et al. 2009), planning (Puterman 1994), machine learning and image processing (Doucet and Johansen 2011). *Decision Theory* pro-

C. Gonzales (✉)
Aix Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
e-mail: christophe.gonzales@univ-amu.fr

P. Perny
Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6,
75005 Paris, France
e-mail: Patrice.Perny@lip6.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_17

poses sophisticated mathematical tools to reason in such contexts. These tools can be roughly divided into two classes: *decision support systems* (or *decision aiding systems*) and *automatic decision systems*. The goal of the former is to help, to guide, human agents to make their decisions, especially when those are very complex (e.g., when many conflicting "decision criteria" need be taken into account). As such, they prove to be very useful in critical domains like, e.g., in medical decision making (Franklin et al. 1991; Bleichrodt 1996), in space shuttle flying assistance (Horvitz and Barry 1995) and in strategic applications like choosing the location of a future airport (Keeney and Raiffa 1993). As for *automatic decision systems*, they are designed to enable non-human agents (robots or software) to choose the best actions to reach their goals. They prove to be useful, for instance, in robotics (Argall et al. 2009), in missiles or drones control (Dasgupta 2006), in serious games (Sordoni et al. 2010).

The decision theoretic tools developed in both decision aiding systems and automatic decision systems rely on mathematical models for representing the agents' preferences. Those enable, for instance, to justify why the majority of agents who are asked which envelope they would prefer to get between two envelopes A and B containing 100 € and 200 € respectively answer they prefer B. Actually, Decision "choose envelope B" yields the consequence of winning 200 €, which is often judged more preferable than the consequence of winning only 100 €. Unfortunately, in real-world applications, decisions are not so simple especially because, when decisions are made, their consequences are to some extent uncertain, i.e., they depend on the occurrences of some events that are still uncertain at the time the decision is made. For instance, when a physician must determine the best treatment to prescribe to his patient, his diagnosis does not allow him to know for sure the exact illness of the patient. Therefore, to take into account all the parameters involved in the decision making, decision models under uncertainty combine two components: a "preference model" and a "representation of uncertainties".

Models and algorithms originating from the field of "decision theory under uncertainty" are widely used in artificial intelligence. This is essentially due to two reasons: (i) these models have strong mathematical foundations; and (ii) their axiomatic justifications rely on rationality arguments with which everybody seems to agree. As a consequence, the conclusions reached by these algorithms can be justified using essentially common sense arguments. The rationality justification incited the majority of the artificial intelligence community to adopt the "*expected utility*" (EU) criterion as *its* decision criterion. In Sect. 2, we describe this criterion and we focus on its axiomatic foundation. The notion of risk and how it translates in the EU model are also investigated. Finally, this section ends with a discussion on the descriptive limits of this model. This naturally calls for other models that can go beyond these limits: in Sect. 4, the emphasis is made on representing uncertainties outside the probabilistic framework while, in Sect. 3, the linear model of preferences itself is questioned. Finally, Sect. 5, addresses sequential decision problems, especially their representations and the issues raised by non-linear models.

2 The Expected Utility Criterion (EU)

Let $\mathscr{D}$ denote the set of decisions that can be made by an agent. In the rest of this chapter, we assume that the agents have well defined preferences on $\mathscr{D}$ and we denote by $\succsim_{\mathscr{D}}$ their preference relation. Thus, $d_1 \succsim_{\mathscr{D}} d_2$ means that the agent prefers Decision d_1 to Decision d_2 or she is indifferent between the two decisions. Strict preference is denoted as usual by $\succ_{\mathscr{D}}$. As we have seen in the introduction, when making decisions under *certainty*, preferring d_1 to d_2 amounts to prefer the consequence yielded by decision d_1 to that yielded by d_2. Let $\mathscr{X}$ denote the space of all the possible consequences. Preference relation $\succsim_{\mathscr{D}}$ over $\mathscr{D}$ is thus induced by preference relation $\succsim_{\mathscr{X}}$ over $\mathscr{X}$ as follows: $d_1 \succsim_{\mathscr{D}} d_2$ if and only if $x(d_1) \succsim_{\mathscr{X}} x(d_2)$, where $x(d)$ represents the consequence of decision d.

Under uncertainty, i.e., when the consequence of a decision is not fully known when the decision is made, the equivalence between $\succsim_{\mathscr{D}}$ and $\succsim_{\mathscr{X}}$ does not exist anymore. However, in this case, it is reasonable to assume that preference relation $\succsim_{\mathscr{D}}$ takes into account not only the agent's preferences over the consequences of the decisions but also her attitude w.r.t. the uncertainty over the fulfillment of these consequences. As an example, when asked to choose between an envelope A containing 100 € and an envelope B randomly chosen among a heap of 100 envelopes in which 97 contain 1 € and 3 contain 1000 €, most of people prefer envelope A because the probability of getting 1000 € with envelope B is too low. From this simple example, we can deduce that the agent translates the uncertainty over the 100 envelopes into an uncertainty over the amount of money contained in envelope B, i.e., on the consequence yielded by decision d_2, and the decision is made taking into account the latter. Before investigating further how agents make their decisions, we need to define more precisely the notion of uncertainty from the agent's perspective. Similarly to probability theory, we need to define what are an event and an elementary event: an event is a set of possible results of a random experiment (above, the choice of an envelope) and an elementary event, which is called a *state of nature* in decision theory, corresponds to only one possible result. Thus, if the envelopes are numbered from 1 to 100, the fact that the envelope chosen is the 3rd one is a state of nature whereas the fact that it has an even number is a (non-elementary) event. Let $\mathscr{S}$ and $\mathscr{A} = 2^{\mathscr{S}}$ denote the set of the states of natures and the set of events respectively. The above example of the envelopes suggests that probabilities are an attractive representation of the uncertainties with which the agent has to cope. This is the very representation exploited in the model presented in this section. Note however that this is not the only possible choice, as we will see later.

In the EU model, uncertainties are represented by probabilities and the decision criterion is simply the maximum of the expectation of the satisfaction provided by the decisions (a.k.a. a *utility*). Let $u : \mathscr{X} \mapsto \mathbb{R}$ be a function such that $x \succsim_{\mathscr{X}} y$ if and only if $u(x) \geq u(y)$. Such a function is called a *utility function* or, for short, a *utility*. A utility function therefore assigns to each consequence a real number such that the preferred the consequence, the higher the number. The utility expectation criterion has been popularized by Daniel Bernoulli in the 18th century (Bernoulli

1738), although a letter by Gabriel Cramer to Nicolas Bernoulli seems to establish that Cramer proposed this criterion earlier. But its modern axiomatic foundations are due, on one hand to von Neumann and Morgenstern (1944) and, on the other hand, to Savage (1954). These two axiomatics differ essentially by the fact that the former assume the existence of a probability distribution on $(\mathscr{S}, \mathscr{A})$ whereas the latter derives its existence from the rationality of the agent.

2.1 von Neumann-Morgenstern's Axiomatic Foundation

von Neumann and Morgenstern assume that $(\mathscr{S}, \mathscr{A}, P)$ is a probabilistic space. In other words, P is a probability distribution over $(\mathscr{S}, \mathscr{A})$. As we have seen before, for each decision d, this distribution induces another probability distribution P_d over the space of consequences $(\mathscr{X}, \mathscr{C})$, where $\mathscr{C} = 2^{\mathscr{X}}$. When the support of P_d is finite, i.e., when the number of possible consequences (those with nonzero probabilities) resulting from making decision d is finite, distribution P_d is called a *lottery*. A lottery can therefore be represented as a tuple $\langle x_1, p_1; \ldots; x_n, p_n\rangle$, where the x_i's are some consequences and the p_i's correspond to their probability of occurrence. Note that a lottery is a representation, a summary, of what a decision really is: indeed it only represents synthetically what can result from making the decision. If this summary is faithful, then we can conclude that there exists an equivalence between the preferences of the agent over the decisions and those over their associated lotteries. Let $\mathscr{L}$ be the set of all the possible lotteries and let $\succsim$ be the preference relation of the agent over $\mathscr{L}$. Then we can conclude that $d_1 \succsim_{\mathscr{D}} d_2$ if and only if $P_{d_1} \succsim P_{d_2}$, where P_d represents the lottery associated with decision d. von Neumann and Morgenstern (1944) show that the preferences over $\mathscr{L}$ (and thus over $\mathscr{D}$) of any *rational* agent necessarily follow the expected utility criterion:

$$\text{For all } P, Q \in \mathscr{L}, \quad P \succsim Q \Longleftrightarrow \sum_{i=1}^{n} p_i u(x_i) \geq \sum_{j=1}^{r} q_j u(y_j), \tag{1}$$

where $P = \langle x_1, p_1; \ldots; x_n, p_n\rangle$, $Q = \langle y_1, q_1; \ldots; y_r, q_r\rangle$, and $u(x)$ is a utility function over the space of consequences $\mathscr{X}$ (this function is called the *von Neumann-Morgenstern utility function*). The first axiom exploited by von Neumann and Morgenstern to prove this result is the following:

Axiom 1 (Complete weak order) *$\succsim$ is a complete weak order on $\mathscr{L}$. In other words, $\succsim$ is reflexive (for all $P \in \mathscr{L}$, $P \succsim P$), transitive (for all $P, Q, R \in \mathscr{L}$, $(P \succsim Q) \wedge (Q \succsim R) \Longrightarrow P \succsim R$) and complete (for all $P, Q \in \mathscr{L}$, $(P \succsim Q) \vee (Q \succsim P)$).*
In addition, $\succsim$ is non-trivial, i.e., there exist $P, Q \in \mathscr{L}$ such that $P \succ Q$.

This axiom simply expresses the idea that, given any pair of lotteries, the agent is always capable of determining which one she prefers (completeness) and that if she prefers P to Q and Q to R, then, logically, she will also prefer P to R. This last

property conveys some kind of rationality, although it is possible to find examples in which rational decision makers have intransitive preferences (Anand 1993). Finally, non-triviality just guarantees that we study only situations in which all the decisions are not judged as equivalent by the agent, the decision maker (if this were not the case, a decision making model would be useless).

For the next two axioms, we need to define *mixtures of lotteries*: let P and Q be two lotteries and let $\lambda \in [0, 1]$ be a real number, then $R = \lambda P + (1 - \lambda)Q$, the mixture of P and Q w.r.t. λ, represents the lottery such that, for any consequence $x \in \mathscr{X}$, the probability of occurrence of x is $R(x) = \lambda P(x) + (1 - \lambda)Q(x)$. Intuitively, a mixture essentially amounts to create R in two steps: first, a coin with a probability λ to land on head (and therefore $1 - \lambda$ to land on tail) is flipped; second, if the coin landed on head, then we get lottery P, else we get lottery Q. The probability of occurrence of each consequence x is consequently $\lambda P(x) + (1 - \lambda)Q(x)$.

Axiom 2 (continuity) *For all* $P, Q, R \in \mathscr{L}$ *such that* $P \succ Q \succ R$, *there exist* $\alpha, \beta \in]0, 1[$ *such that:*

$$\alpha P + (1 - \alpha)R \succ Q \succ \beta P + (1 - \beta)R.$$

This axiom conveys the idea that if the agent strictly prefers P to Q, then a lottery resulting from a very small perturbation of P should still be preferred to Q. For instance, if $P = \langle 100, 1\rangle$, i.e., P is the lottery which yields 100 € with certainty, and if $Q = \langle 10, 1\rangle$ and $R = \langle 5, 1\rangle$, then an agent who likes money should have the following preference relation: $P \succ Q \succ R$. If $\alpha = 1 - 10^{-20}$, lottery $\alpha P + (1 - \alpha)R$ is equal to $\langle 100, 1 - 10^{-20}; 5, 10^{-20}\rangle$. The chance of receiving 5 € is so small that the agent is almost assured to win 100 €, which is preferable to Q. Therefore, it is very likely that the agent prefers $\alpha P + (1 - \alpha)R$ to Q. A similar argument can be used with β very close to 0. Here again, Axiom 2 seems quite reasonable. The last axiom used by von Neumann and Morgenstern is the following:

Axiom 3 (independence) *For every* $P, Q, R \in \mathscr{L}$ *and every* $\alpha \in]0, 1]$*:*

$$P \succsim Q \iff \alpha P + (1 - \alpha)R \succsim \alpha Q + (1 - \alpha)R.$$

The interpretation of this axiom follows that of mixtures. We have seen that $\alpha P + (1 - \alpha)R$ corresponds to a lottery created in two steps: first a coin is flipped, with probabilities α and $1 - \alpha$ to land on head and tail respectively and, then, depending on the side on which the coin landed, the agent receives lottery P or R. Following this principle, Axiom 3 can be interpreted as follows: if the coin lands on tail, from both lotteries $\alpha P + (1 - \alpha)R$ and $\alpha Q + (1 - \alpha)R$, the agent receives the same induced lottery R so, logically, in this case, she should be indifferent between $\alpha P + (1 - \alpha)R$ and $\alpha Q + (1 - \alpha)R$. If, on the other hand, the coin lands on head, then, from $\alpha P + (1 - \alpha)R$ and $\alpha Q + (1 - \alpha)R$, she receives lotteries P and Q respectively. As she (weakly) prefers P to Q or is indifferent between these two lotteries, she should also prefer $\alpha P + (1 - \alpha)R$ to $\alpha Q + (1 - \alpha)R$ or be indifferent between them, hence the axiom.

The three above axioms therefore express properties that can be expected from a rational agent. As the next theorem shows, they imply that there exists a unique decision criterion representing the preferences of the agent, and this one is precisely the expected utility criterion:

Theorem 1 (von Neumann-Morgenstern) *The following two assertions are equivalent:*

1. *The preference relation* $\succsim$ *over* $\mathscr{L}$ *satisfies Axioms 1, 2 and 3.*
2. $\succsim$ *is representable by a utility function* $U : \mathscr{L} \mapsto \mathbb{R}$ *such that* $U(P) = \sum_{i=1}^{n} p_i u(x_i)$, *where* $u(x_i) = U(\langle x_i, 1\rangle)$.

Function $u : \mathscr{X} \mapsto \mathbb{R}$ *is called the von Neumann-Morgenstern utility function of the agent and is unique up to scale and location (i.e., up to strictly positive affine transforms).*

This strong relationship between rationality and the EU criterion explains why EU is so popular in the artificial intelligence community but also among decision theorists and operations research scientists. Note also that the above theorem can be generalized, notably by using more general probability measures (Fishburn 1970, 1982). There also exist other axiomatics, like, e.g., the one provided in Jensen (1967), Herstein and Milnor (1953) or in Fishburn and Roberts (1978).

The von Neumann-Morgenstern axiomatics raises one issue: it assumes the existence of an "*objective*" probability distribution over the space of the states of nature —in decision theory, this situation is called "*decision under risk*" (Knight 1921)— and one may wonder whether such a hypothesis is so reasonable in practical decision theory problems and, more generally, in artificial intelligence. As we will see, the answer to this question seems positive because the existence of a probability distribution over the states of nature necessarily follows from the rationality of the agent. This idea has been initially introduced in Ramsey (1931) but went largely unnoticed until the seminal book by Savage (1954) got published. Sixty years later, the idea that probabilities are the only "rational" representation of uncertainties is so deeply anchored into people's minds that, up to recently, it was very difficult in artificial intelligence to imagine a rational decision making process outside the EU framework.[1] When the probability distribution over the states of nature results from the rationality of the agent, this distribution is said to be "*subjective*" and the decisional context is called "*decision under uncertainty*" instead of "*decision under risk*", which is dedicated to the case of objective probabilities.

Let us now study Savage's axiomatics (Savage 1954), which has led to the decision model called "*Subjective Expected Utility*" (SEU). Of course, since the probability distribution over the space of the states of nature $(\mathscr{S}, \mathscr{A})$ results from the rationality

[1] Outside the EU framework, the behavior of an agent cannot be rational (w.r.t. Savage's meaning) and, therefore, it is thought in artificial intelligence that such a behavior must be proscribed. In the 70's and 80's, decision theorists, notably Kahneman, Tversky and Quiggin, suggested that Savage's rationality was not the only possible form of rationality and they proposed to depart from the Savagian framework and developed their own kinds of "rationality". This paved the way to new decision models like, e.g., RDU, that recently attracted the attention of AI researchers.

of the decision maker, this is no more a primitive of the decisional language. The primitive, here, is called an "*act*". Quite similarly to a lottery, this corresponds to the representation/summary of a decision but its description is more precise than a lottery. An *act* f is a function whose domain and codomain are $\mathscr{S}$ and $\mathscr{X}$ respectively. In other words, to each state of nature e, function f assigns the consequence that would result from the occurrence of e if the decision represented by f were made. For pedagogical purposes, let us first consider "*simple acts*", i.e., finite-valued and $\mathscr{A}$-measurable functions. For any simple act f, there exists a finite partition $\{E_i, i \in I\}$ of $\mathscr{S}$ such that, for all $i \in I$, $\{f(e) : e \in E_i\} = \{c_i\}$, where $c_i \in \mathscr{X}$ is a consequence. To put it differently, a simple act yields a finite set of possible consequences and those depend on the realization of some states of nature belonging to E_i. To draw a parallel with lotteries, a simple act corresponds to a description of a lottery with finite support, although this description is more precise than just the lottery. Table 1 highlights their differences: this table shows two acts, f_1 and f_2. In the former, when state e_1 obtains, the resulting consequence is c_1. For both acts, the probabilities of getting c_1 and c_2 are 0.3 and 0.7 respectively. Consequently, both acts correspond to the same lottery $\langle c_1, 0.3; c_2, 0.7\rangle$. However, as can be seen in the table, act f_1 is different from f_2. An act is therefore a description of a decision which is more precise than a lottery. In the sequel, we will explain the interpretations of new concepts and axioms using simple acts but those apply on general acts, not only on simple ones. Finally, let δ_c denote the "*constant*" act yielding consequence c, i.e., the act such that $\delta_c(\mathscr{S}) = \{c\}$. To illustrate visually what acts represent and how they will be combined, we will use figures in which the X and Y axes represent sets $\mathscr{S}$ and $\mathscr{X}$ respectively. In this setting, a simple act is just a stepwise function, as shown in Fig. 1.

In the axiomatic theory of von Neumann-Morgenstern, one of the key ideas was the possibility to combine by "mixture" different lotteries to produce new ones. This was the core of their proof. Here, there exists an equivalent operation on acts, that we will call a "*splicing*" to distinguish it from mixtures.[2] Let f and g be two acts and let $E \subseteq \mathscr{S}$ be an event. The splicing of f and g w.r.t. E, denoted by fEg is the act $h = fEg$ such that $h(s) = f(s)$ for all $s \in E$ and $h(s) = g(s)$ for all $s \in E^C = \mathscr{S} \setminus E$. Figure 1 illustrates this operation. All the primitives and operations necessary to describe the axiomatic theory of SEU are defined, we can now study the axioms provided by Savage. The first one corresponds in essence to Axiom 1 of von Neumann-Morgenstern.

Table 1 Comparison between acts and lotteries

Event	Probability	Act f_1	Act f_2
e_1	0.3	Cons. c_1	Cons. c_2
e_2	0.3	Cons. c_2	Cons. c_1
e_3	0.4	Cons. c_2	Cons. c_2

[2] In his book, Savage did not name this operation. The term "splicing" was introduced in Gilboa (2009).

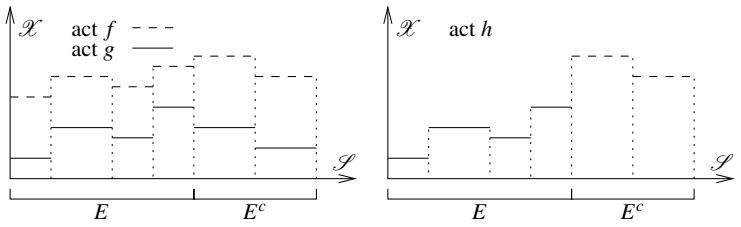


Fig. 1 The concept of splicing

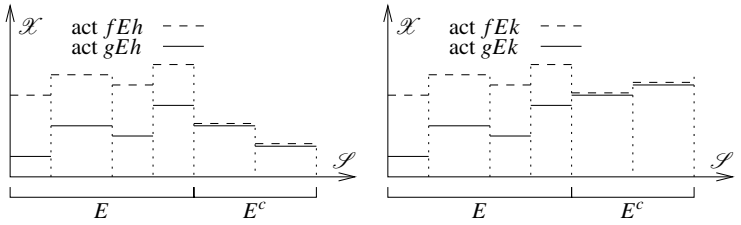


Fig. 2 Illustration of the sure thing principle

Axiom 4 (P1: Weak order on acts) *The set of all the acts is closed under splicing and there exists a complete weak order* $\succsim$ *over the set of acts.*

But Savage's key axiom is the "*Sure Thing Principle*".[3]

Axiom 5 (P2: Sure Thing Principle) *For all acts* $f, g, h, k \in \mathscr{X}^{\mathscr{S}}$ *and all* $E \subseteq \mathscr{S}$*:*

$$fEh \succsim gEh \iff fEk \succsim gEk.$$

This axiom corresponds in spirit to the independence axiom of von Neumann-Morgenstern. Figure 2 provides an illustration: let fEh and gEh be two acts. They yield the same consequences over E^C. Consequently, if the state of nature that obtains belongs to E^C, the agent should be indifferent between both acts. So, if globally, she prefers fEh to gEh, this means that, over E, she prefers the consequences yielded by f to those by g. Now, substitute the common part of both acts h on E^C by another act k. Then, the resulting acts are fEk and gEk. These new acts yield precisely the same consequences over E^C, so the agent should still be indifferent between them if the state of nature that obtains belongs to E^C. And if the state that obtains belongs to E, then both acts yield the same consequences as fEh and gEh, so, globally, if the agent preferred fEh to gEh, she should also prefer fEk to gEk. In other words, the sure thing principle states that, when comparing two acts, the agent only compares the acts on the events on which they differ. This axiom looks quite reasonable.

[3]Most authors name P2 as the "sure thing principle" but it was pointed out by Peter Wakker that, in Savage's book, the sure thing principle refers to axioms P2, P3 and P7.

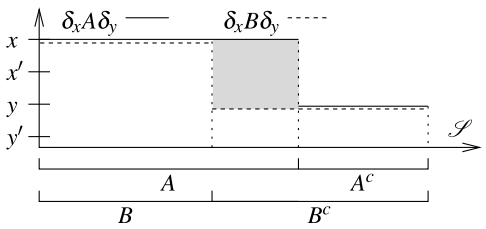


Fig. 3 Interpretation of Axiom P4

Note that Axiom P2 implies the existence of a weak order $\succsim_E$ for every event E defined as $f \succsim_E g$ if and only if $fEh \succsim gEh$ for all h. The next axiom exploits this new preference relation to guarantee, using constant acts, that the agent has a well-defined preference relation $\succsim_{\mathscr{X}}$ over the space of consequences. This axiom relies on *non-null* events, i.e., on events E such that there exist at least two acts f and g such that $f \succ_E g$.

Axiom 6 (P3: Preferences among consequences) *For all consequences $x, y \in \mathscr{X}$ and all non-null events $E \subseteq \mathscr{S}$, $\delta_x \succsim_E \delta_y$ if and only if $x \succsim_{\mathscr{X}} y$, where δ_x and δ_y are constant acts.*

In the SEU framework, the existence of an "objective" probability distribution over the states of nature is never assumed. Rather, the existence of a "subjective" distribution results from the beliefs of the agent herself. The agent must therefore have beliefs that an event A is more or less likely to occur than another event B. This is exactly what the next axiom induces:

Axiom 7 (P4: Preferences over events) *For all consequences $x, x', y, y' \in \mathscr{X}$ such that $x \succ_{\mathscr{X}} y$ and $x' \succ_{\mathscr{X}} y'$, and for all $A, B \subseteq \mathscr{S}$,*

$$\delta_x A \delta_y \succsim \delta_x B \delta_y \iff \delta_{x'} A \delta_{y'} \succsim \delta_{x'} B \delta_{y'}.$$

Figure 3 illustrates this axiom: acts $\delta_x A \delta_y$ and $\delta_x B \delta_y$ differ only on the gray area. On this one, $\delta_x A \delta_y$ yields consequence x and $\delta_x B \delta_y$ yields y, which is not preferred to x. This explains why $\delta_x A \delta_y \succsim \delta_x B \delta_y$. In this figure, the existence of the gray area results from the fact that A contains B and, consequently, it is more "probable" to happen than B. In general, it can be shown that, whenever the agent believes that A is more likely to happen than B, then the preferences of the agent satisfy Axiom P4. Axiom P5 below expresses the fact that all the consequences are not judged as equivalent by the agent (otherwise, it would be impossible to discriminate between acts and SEU would be useless to help the agent in her decision making process):

Axiom 8 (P5: Non-triviality of preferences over the consequences) *There exist two outcomes $x, y \in \mathscr{X}$ such that $\delta_x \succ \delta_y$.*

The five above axioms seem rather reasonable and do not seem too restrictive in the sense that they tend to win unanimous support from people. Yet, as Savage

showed, their combination necessarily induces that the agent models uncertainties using a qualitative probability.[4] To establish the existence of a subjective probability, an additional axiom is needed, that is closely related to the continuity axiom of von Neumann-Morgenstern: assuming that E_i is a highly unlikely event, if f and g are two acts such that $f \succ g$ and if x is an arbitrary consequence, then $\delta_x E_i f$ should be very close to f and, therefore, as $f \succ g$, the agent should also prefer $\delta_x E_i f$ to g. For the same reason, she should also prefer f to $\delta_x E_i g$:

Axiom 9 (P6: Continuity) *For all acts $f, g \in \mathscr{X}^{\mathscr{S}}$ such that $f \succ g$ and for all $x \in \mathscr{X}$, there exists a finite partition $\{E_1, \ldots, E_n\}$ of $\mathscr{S}$ such that $\delta_x E_i f \succ g$ and $f \succ \delta_x E_i g$ for every $i \in \{1, \ldots, n\}$.*

Adding Axiom P6 to the five other axioms necessarily induces the existence of a subjective probability distribution. In addition, all these axioms induce that the agent is an expected utility maximizer, as shown in the following theorem:

Theorem 2 (Savage, 1954[5]) *If the preferences of an agent satisfy axioms P1 to P6, then preference relation $\succsim$ over the set of acts with finite support is representable by a utility function $U(f) = \sum_{s \in \mathscr{S}} p(s)u(f(s))$, where $p(s)$ is the subjective probability of the agent over the state of nature s. In addition, u, the utility function over the set of consequences, is unique up to scale and location.*

Savage has also extended this theorem, notably to the case in which acts are only constrained to be bounded (Savage 1954). Note that there also exist other axiomatics of the EU criterion under uncertainty, notably that of Anscombe and Aumann (1963). All these axiomatics have however in common to rely on axioms that are easily justifiable and that, to some extent, reflect a logical reasoning. In this sense, they constitute the foundation of a rational behavior. From all these axiomatics, it could be easily inferred that only probabilities can "rationally" model uncertainties. This assertion has also been supported for a long time by what decision theorists call "*Dutch books*", which are situations in which using a model of uncertainties different from probabilities inevitably leads the agent to loose some money. As an example, let us consider a bookmaker proposing bets on the three horses of a race. He offers the odds shown in Table 2. Note that the sum of the induced "probabilities" estimated by the bookmaker is equal to 0.95, not to 1. This deviation from a valid probability distribution implies that there is a possibility for gamblers to always win money from the bookmaker. Indeed, gamblers betting the amounts of money shown in the fourth column of the table are guaranteed to win 200 € even though they bet only 190 €. This type of money pump argument has also significantly contributed to establish probabilities as the only reasonable representation of uncertainties in a decision making context.

[4] Note that qualitative probabilities are slightly different from probabilities, see Kraft et al. (1959) for a proof of this assertion.

[5] Savage's theorem is somewhat more general than the theorem mentioned here: acts need not have a finite support, it is sufficient that the set of consequences $\mathscr{X}$ is finite. In this case, the summation needs be substituted by an integral w.r.t. the subjective probability measure.

Table 2 Example of a Dutch book

Horse	Odds	Induced proba.	Bet price	Reimbursement
1	1 against 1	$\frac{1}{1+1} = 0.5$	100 €	100 € of bet + 100 € = 200 €
2	3 against 1	$\frac{1}{3+1} = 0.25$	50 €	50 € of bet + 150 € = 200 €
3	4 against 1	$\frac{1}{4+1} = 0.2$	40 €	40 € of bet + 160 € = 200 €

In the two axiomatics above, that of von Neumann-Morgenstern and that of Savage, the von Neumann-Morgenstern utility function, i.e., the utility representing the agent's preferences over the consequences, is unique up to scale and location. But in decision under certainty, i.e., when the consequences of each action are known with certainty, utility functions (over outcomes) are unique only up to strictly positive increasing transforms. Consequently, we can deduce that von Neumann-Morgenstern utilities must implicitly include some factor related to uncertainties. We will see now that, in reality, this factor represents the attitude of the agent w.r.t. risk.

2.2 Risk Measures

Before defining formally the agent's attitude w.r.t. risk, we need to define the concept of *risk*, and especially how the *quantity* of risk involved in a decision can be measured. A decision can be summarized by an act or a lottery $\langle x_1, p_1; \ldots, x_n, p_n \rangle$. In a sense, the latter correspond to a random variable x whose domain is $x_1, \ldots, x_n$ and the usual risk measure of a real-valued random variable is its variance. So it is tempting to exploit variance as the measure of risk involved in a decision. This idea is supported by the celebrated Arrow-Pratt formula for approximating utility functions, which contains a component related to variance (Pratt 1964; Arrow 1965). But as shows the following example in Ingersoll (1987), this measure is not very well suited: let $L_1 = \langle 0,\ 0.5\ ;\ 4,\ 0.5 \rangle$ and $L_2 = \langle 1,\ 7/8\ ;\ 9,\ 1/8 \rangle$ be two lotteries. Intuitively, observing L_1 and L_2, lottery L_1 seems more risky than L_2 since its consequences are equiprobable whereas, in L_2, it is very likely that the decision yields consequence 1. Unfortunately, the variances of both lotteries are equal.

In decision theory, the most commonly used risk measure is due to Rotschild and Stiglitz (1970, 1971). It is much more robust than variance. It relies on the concept of "*mean-preserving risk increase*" or, as stated usually, "*Mean Preserving Spread*" (MPS). Let us consider the three lotteries P, Q, R of Table 3. Observe the only difference between P and Q: Lottery P yields consequence 4 with probability 0.3 whereas, Q yields consequences 3 and 5 with probability 0.15 (hence, globally, a probability of 0.3 to get consequence "3 or 5"). As a result, Q can be judged as more risky than P since, with a probability of 0.3, the consequence yielded by P is known (i.e., 4) whereas, in Q, with the same probability, we only know that 3 or 5 will be yielded, and there still exists a lottery $\langle 3, 0.5\ ;\ 5, 0.5 \rangle$ to determine which

Table 3 Mean preserving spread: $Y = \text{MPS}(X)$, $Z = \text{MPS}(Y)$ and $Z = \text{MPS}(X)$

X	$P(X)$	Y	$Q(Y)$	Z	$R(Z)$
–2	0.09	–2	0.09	–2	0.09
4	0.30	3	0.15	3	0.15
		5	0.15	5	0.15
10	0.40	10	0.40	10	0.40
16	0.21	16	0.21	12	0.07
				18	0.14

consequence will be yielded. Remark that the expectations of random variables X and Y of Table 3 are equal. This explains why Y is said to be a mean-preserving (same expectation as X) risk increase (w.r.t. X) or, for short, a MPS of X. Similarly, Z is a MPS of Y because their expectations are equal and Y yields consequence 16 with probability 0.21 whereas Z induces lottery $\langle 12, 0.07\ ;\ 18, 0.14\rangle$ instead.

In the rest of this subsection, we will consider that $\mathscr{X}$ is equal to $\mathbb{R}$ and, more generally, that it is a monetary space (this will make the interpretations of the results easier to understand).

Definition 1 (*Mean preserving spread*) Let X and Y be two real-valued random variables. Y is said to be a *Mean Preserving Spread* of X if and only if there exists a white noise Θ, i.e., a random variable whose expectation is equal to 0, such that $Y = X + \Theta$.

Let us call F_X and F_Y the cumulative distribution functions (CDF) of random variables X and Y respectively. In other words, if P_X is the probability distribution of X, then $F_X(x) = P_X(z : z \leq x)$ for every $x \in X$. Figure 4 displays the CDFs of variables X and Z of Table 3. When $X, Z < 3$, the two CDFs are identical. Then, when $x \in [3, 4[$, we have that $F_Z(x) > F_X(x)$. Therefore, we also have that $\int_{x<4} F_Z(x)dx > \int_{x<4} F_X(x)dx$. When $x \in [4, 5[$, the difference $F_X(x) - F_Z(x)$ is positive, so the gap between the two integrals decreases but the two gray regions on the left of Fig. 4 have the same area so, overall, the integral of F_Z is always greater than or equal to that of F_X. This property is general and provides an alternative characterization of MPS:

Definition 2 (*Mean preserving spread*) Let X and Y be two real-valued random variables. Y is said to be a *Mean Preserving Spread* of X if and only if (i) X and Y have the same expectations; and (ii) X and Y satisfy the following equation:

$$\int_{-\infty}^{T} F_Y(x)dx \geq \int_{-\infty}^{T} F_X(x)dx \quad \text{for every } T \in \mathbb{R}. \tag{2}$$

Definition 3 (*2nd order stochastic dominance*) Let X and Y be two real-valued random variables. X dominates stochastically Y at the second order if and only if Eq. (2) is satisfied.

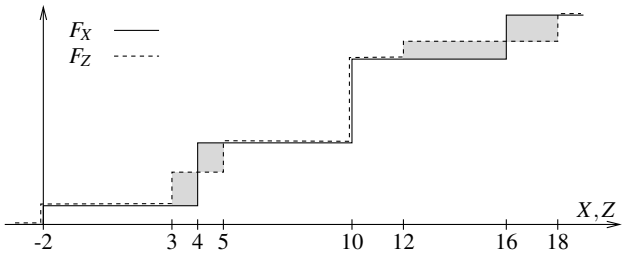


Fig. 4 Interpretation of MPS in terms of cumulative distributions

Rotschild and Stiglitz proved that Definitions 1 and 2 are equivalent. They also provided a characterization in terms of risk aversion, as we will define it in the next subsection: Assertion 3 of the theorem below expresses the fact that Y is a MPS of X if and only if any weakly risk averse agent prefers X to Y.

Theorem 3 (Rotschild and Stiglitz 1970) *Let X and Y be two real-valued random variables with the same expectation. The following three assertions are equivalent:*

1. *$Y = MPS(X)$ (in the sense of Definition 2);*
2. *Y has the same distribution as $X + \Theta$, where Θ is a white noise;*
3. *for any increasing and concave function $u : \mathbb{R} \mapsto \mathbb{R}$, we have that $\int u(x)\, dF_X(x) \geq \int u(x) dF_Y(x)$.*

We can now characterize the behavior of agents w.r.t. lotteries with different amounts of risk. Of special interest, we can now determine if the agent would prefer "taking risks" or not.

2.3 Attitude of Agents with Respect to Risk

The simplest way to estimate whether an agent is risk seeking or risk averse consists of asking her which lottery she would prefer among one lottery X without any risk (it can yield only one consequence, known for sure) and another lottery Y with the same expectation but containing some risk (the lottery can yield several consequences). Note that, as both lotteries have the same expectation, $Y = \text{MPS}(X)$. Assume now that the agent's von Neumann-Morgenstern utility is linear ($u(x) = x$ for simplicity). Then the expected utility of the lottery corresponding to Y is equal to the expectation of Y which, by definition, is equal to that of X and, also, to the expected utility of the lottery associated to X. An agent who is expected utility maximizer shall therefore be indifferent between X and Y. For instance, for the agent, $\left\langle \frac{x_1+x_2}{2}, 1 \right\rangle \sim \left\langle x_1, \frac{1}{2}; x_2; \frac{1}{2} \right\rangle$. These two lotteries have the same expectation (this is the reason why the agent is indifferent between them), but the first one is not risky while the second one is. So we can conclude that the preferences of the agent do not take into account the amount of

risk involved in the lotteries. The agent is thus said to be "*risk neutral*". Of course, if the agent had strictly preferred X to Y, we would say that she has some aversion w.r.t. risk and, therefore, she would be "*risk averse*". Finally, if the agent had strictly preferred Y to X, she would be said to be "*risk seeking*". Arrow and Pratt propose the following definition (Pratt 1964; Arrow 1965):

Definition 4 (*Weak risk attitudes*) An agent is weakly risk averse if, for every real-valued random variable X, she prefers $E(X)$ to random variable X itself: $\langle E(X), 1\rangle \succ X$. An agent is weakly risk neutral (resp. seeking) if $\langle E(X), 1\rangle \sim X$ (resp. $X \succ \langle E(X), 1\rangle$).

We have seen above that a linear von Neumann-Morgenstern utility implies that the agent is risk neutral. Arrow and Pratt have shown that, more generally, the agent's risk attitude is characterized by the concavity or convexity of the von Neumann-Morgenstern utility function:

Theorem 4 *An agent is (weakly) risk averse if and only if her von Neumann-Morgenstern utility function u is concave. She is (weakly) risk neutral if and only if u is linear. Finally, she is (weakly) risk seeking if and only if u is convex.*

Up to now, the risk attitude of the agent was characterized by comparing one risky lottery with a lottery involving no risk. It could be objected that such a comparison is extreme and could introduce some biases. So it might be more appropriate to compare only lotteries involving some risk, some being more risky than others. The concept of mean preserving spread allows to specify such lotteries: it is sufficient to compare lotteries X and Y such that one of them is an MPS of the other. In this case, an agent is risk averse if and only if she prefers lottery X to any MPS(X):

Definition 5 (*Strong risk attitudes*) An agent is strongly risk averse if, for every real-valued random variable X, she prefers lottery X to any lottery Y such that $Y = \text{MPS}(X)$. An agent is strongly risk neutral (resp. seeking) if $X \sim Y$ (resp. $Y \succ X$).

Of course, by definition, strong risk aversion implies weak risk aversion. But in the EU model, the converse is also true:

Theorem 5 (Rotschild and Stiglitz 1970) *In the EU model, the following three assertions are equivalent:*

1. *the agent is weakly risk averse;*
2. *the agent is strongly risk averse;*
3. *the agent's von Neumann-Morgenstern utility is concave.*

As the concavity of the von Neumann-Morgenstern utility function u characterizes the agent's aversion w.r.t. risk, it seems natural to define the intensity of this aversion in terms of properties of u. Arrow and Pratt have proposed to characterize it in terms of a *coefficient of absolute risk aversion*: assume that u is strictly increasing and twice continuously differentiable, with a strictly positive derivative. The coefficient

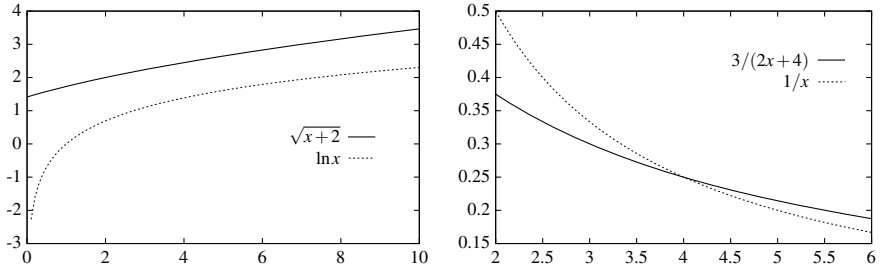


Fig. 5 Coefficients of absolute risk aversion

of absolute risk aversion is defined as function $R_A : \mathbb{R} \mapsto \mathbb{R}$ such that $R_A(x) = -u''(x)/u'(x)$.

This definition can be easily interpreted by considering a risk averse agent. Assume that the set of consequences $\mathscr{X}$ is a monetary space. A *common* agent prefers in general to win more money than less, so her utility $u(x)$ strictly increases with x and, consequently, $u'(x) > 0$. In addition, being risk averse, $u(x)$ is concave, hence $u''(x) < 0$. From these properties, it can be deduced that $R_A(x) > 0$. Consider now utility function $u_1(x) = \ln x$, which implies coefficient $R_A^1(x) = 1/x$. In Fig. 5, it can be observed that the concavity rate of u_1 decreases with x. This translates in terms of coefficient of absolute risk aversion into a decreasing coefficient R_A^1. The level of aversion w.r.t. risk therefore varies with x and, in practice, it is generally strictly decreasing. As a matter of fact, a poor agent is not often prone to take the risk of loosing some money in order to gain more money whereas a wealthy agent is inclined to take such a risk because the same loss of money seems to her relatively much less important than to the poor agent.

Note that R_A can also be exploited to compare the aversions among several agents. Indeed, consider now two utility functions $u_1(x) = \ln x$ and $u_2(x) = \sqrt{x+2}$. These functions induce two coefficients $R_A^1(x) = 1/x$ and $R_A^2(x) = 3/(2x+4)$. Figure 5 displays functions u_1, u_2 as well as their respective coefficients of aversion. From this figure, it can be remarked that the second agent (u_2) is more risk averse for small amounts of money whereas this trend is inverted for larger amounts. Note that such a comparison is meaningful because von Neumann-Morgenstern utilities being unique up to scale and location, R_A remains invariant w.r.t. affine transforms of u.

Clearly, the EU model presents very nice properties. As we have seen, it is justifiable from the viewpoint of the agent's rationality. In addition, its linearity allows for very efficient algorithms, notably in the context of sequential decision making and in that of preference elicitation (Keeney and Raiffa 1993; Chajewska et al. 2000; Boutilier 2002; Wang and Boutilier 2003). However, during the last decades, several criticisms were raised against this model, which led to alternative decision models. The next section shows some of the most important criticisms.

2.4 Some Descriptive Limits of the EU Model

Among the first detractors of the EU model, Allais proposed a celebrated example known as the "*Allais paradox*" (Allais 1953), about which experimental studies have shown that the majority of the surveyed agents have preferences that violate the independence axiom (Axiom 3) and are, therefore, not representable in the EU model. Actually, consider the following two lotteries:

- $L_1 = \langle \text{win } 1\text{M} €, 1\rangle$;
- $L_2 = \langle \text{win } 1\text{M} €, 0.89 \ ; \ 5\text{M} €, 0.1 \ ; \ 0 €, 0.01\rangle$.

Most of the surveyed agents prefer L_1 to L_2 because the uncertainty contained in L_2 is not counterbalanced by the potential gain of 5M €. When faced to the following alternatives:

- $L'_1 = \langle \text{win } 1\text{M} €, 0.11 \ ; \ 0 €, 0.89\rangle$,
- $L'_2 = \langle \text{win } 5\text{M} €, 0.10 \ ; \ 0 €, 0.90\rangle$,

the same agents usually prefer L'_2 to L'_1 because the difference in probability between 0.11 and 0.10 is judged as relatively low and the agents therefore base essentially their preferences on the potential gains of the lotteries. But, if we set: $P = \langle 1\text{M} €, 1\rangle$, $Q = \langle 5\text{M} €, 10/11 \ ; \ 0 €, 1/11\rangle$, $R = \langle 1\text{M} €, 1\rangle$ and $S = \langle 0 €, 1\rangle$, then:

$$\begin{array}{ll} L_1 = 0.11P + 0.89R & L_2 = 0.11Q + 0.89R \\ L'_1 = 0.11P + 0,89S & L'_2 = 0.11Q + 0,89S. \end{array}$$

Therefore, according to the independence axiom, if $L_1 \succ L_2$, the agent shall also have the following preference: $L'_1 \succ L'_2$. Obviously, this is not observed experimentally. This example is quite unsettling because this preference reversal can be explained easily and does not seem to result from some irrational behavior. As we will see in the next section, this example has led researchers to develop new decision models based on different rationality criteria. These models have a higher descriptive power than the EU model and are notably capable of explaining why people tend to prefer L_1 to L_2 and L'_2 to L'_1. Other experimental studies, in particular (Kahneman and Tversky 1972, 1979), highlight other biases w.r.t. the predictions made by the EU model. This is the case, for instance, of the certainty effects.

The second criticism addressed against the EU model concerns the interpretation of the concavity of the von Neumann-Morgenstern utility function u. Indeed, we have seen that in this model a concave utility represents an aversion w.r.t. risk. But u represents the agent's preferences over the space of the consequences and, in general, agents have decreasing marginal preferences over money, i.e., the amount of increase of the agent's satisfaction (as measured by the utility function) tends to decrease when the amounts of money tend to rise. Thus, the satisfaction to increase the agent's wealth from 10 to 20 € is higher than that to increase it from 10010 to 10020 €. In terms of preferences, this decrease necessarily induces the concavity of u. This double interpretation of u's concavity implies that the EU model is unable

to describe the behavior of agents that are at the same time risk averse and that have decreasing marginal preferences.

The third main criticism against the EU model lies in its lack of flexibility to model different types of risk aversions. Indeed, in EU, it is impossible to model an agent who is weakly but not strongly risk averse. But this kind of agent can exist and, more generally, there exist several notions of risk aversion that are not necessarily all equivalent (Chateauneuf et al. 2004). We will see in the next section some "new" decision models that can cope with this lack of flexibility.

The set of criticisms presented here cannot be exhaustive due to lack of space. However, we shall mention two important additional criticisms. First, the formula of the expected utility model combines through multiplications the probabilities of occurrence of the consequences with the utilities. As a consequence, EU necessarily requires the commensurability of preferences and uncertainties: one can "trade" uncertainty for preference satisfaction. For instance, if $\langle x_1, 0.5 \ ; \ x_2, 0.5\rangle \sim \langle x_3, 1\rangle$, the agent is willing to trade/discard some uncertainty (0.5) for a change in consequences (winning x_3 instead of x_1 or x_2, hence a modification in her satisfaction). In addition, even though commensurability may be a reasonable assumption in some practical applications, is it always sensible to model uncertainties by probabilities? According to Savage, this is the only rational representation. However, when considering the example of the Ellsberg's urn (Ellsberg 1961), this justification seems far from being convincing: consider an urn containing red, yellow and black balls. The only information available about these balls is that one third are red and the two remaining third are either yellow or black (but we do not know their respective proportions). With so few information available, it seems difficult for a "rational" agent to estimate the underlying probability distribution over the colors of the balls, and experimental studies highlight this fact. When agents are invited to determine the alternative they prefer among the following ones, whose outcome depends on the color of a ball drawn randomly from the urn:

- Alternative A: win 1M € if the ball is red, else 0 €,
- Alternative B: win 1M € if the ball is black, else 0 €,

most of the agents prefer A to B because, potentially, the urn contains no black ball whereas the urn is guaranteed to contains 1/3 of red balls. On the other hand, when facing the following alternatives:

- Alternative C: win 1M € if the ball is red or yellow, else 0 €,
- Alternative D: win 1M € if the ball is black or yellow, else 0 €,

the agents prefer in general alternative D to C. But this kind of behavior is incompatible with the EU model because it violates the Sure Thing Principle. Indeed, if E represents the event "the drawn ball is red or black", if a_1 and a_2 represent the acts yielding "1M € if red ball, else 0 €" and "1M € if black ball, else 0 €", and if δ_h and δ_k represent the "constant" acts yielding with certainty 0 € and 1M € respectively, then alternatives A and B can be represented by acts $a_1 E \delta_h$ and $a_2 E \delta_h$ respectively, whereas alternatives C and D correspond to acts $a_1 E \delta_k$ and $a_2 E \delta_k$ respectively.

According to the Sure Thing Principle, one of the fundamental principles underlying EU, $A \succ B$ should imply $C \succ D$, which is not the case observed experimentally.

All these descriptive limits have led researchers to propose new models, also relying on rationality criteria, but with a higher expressive power. We will now briefly describe some of them.

3 Non-linear Models for Decision Under Risk

The descriptive limits mentioned above first led decision making researchers to propose models quite similar to EU but, still, weakening one or several axioms of von Neumann-Morgenstern (or of Savage). Let us cite for instance the model proposed in Machina (1982) which discards the independence axiom but is still locally coherent with EU. There also exist models based on security levels like, e.g., that of Jaffray (1988) in which the independence axiom is defined only on pairs of probability distributions that share the same worst consequence.

However, these models have been replaced by what decision theorists call "new" models, which are generalizations of EU. Among the first new models proposed, "*Prospect Theory*" consists of deforming probabilities using an increasing transform (Kahneman and Tversky 1979) in order not to take into account the true probabilities but rather the way agents perceive these probabilities. Although seminal, this model is not used anymore, essentially because it could sometimes propose to the agent to make dominated decisions, i.e., to choose an alternative D_1 such that there existed another alternative D_2 such that, whatever the state of nature that could occur, the consequence yielded by D_2 was judged at least as good as that yielded by D_1 (and it was judged strictly better for at least one state of nature). This feature being very difficult to justify from a rationality point of view, the model is not used anymore. However, it paved the way for the new models, notably for "*Rank Dependent Utility*" (RDU), that we will now describe Quiggin (1982, 1993).

Let x_1, x_2, x_3 be three consequences. Without loss of generality, let us assume that $u(x_2) < u(x_1) < u(x_3)$. According to the EU model, lottery $L = \langle x_1, p_1; x_2, p_2; x_3, p_3\rangle$ is evaluated as $EU(L) = p_1u(x_1) + p_2u(x_2) + p_3u(x_3)$. It is easy to show that this expression is equivalent to:

$$EU(L) = (p_1 + p_2 + p_3)u(x_2) + (p_1 + p_3)[u(x_1) - u(x_2)] + p_3[u(x_3) - u(x_1)]. \tag{3}$$

This new expression can be interpreted as follows: at worst, the agent is guaranteed with probability $p_1 + p_2 + p_3 = 1$ to win consequence x_2. Then, the probability that she gets a consequence strictly better than x_2, i.e., at least as good as consequence x_1 is $p_1 + p_3$. Finally, the probability to win something better than x_1, i.e., x_3, is p_3. The key idea of RDU is to combine this expression with the probability transformation principle of the Prospect Theory. Thus, in its decision making process, RDU does not take into account the true probabilities but only their perceptions by the agent. The score assigned to L by RDU is therefore:

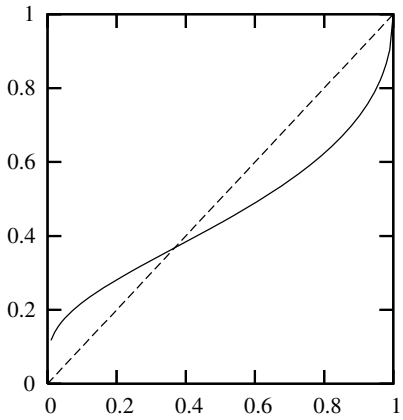


Fig. 6 Probability transformation function

$$RDU(L) = \varphi(p_1 + p_2 + p_3)u(x_2) + \varphi(p_1 + p_3)[u(x_1) - u(x_2)] + \varphi(p_3)[u(x_3) - u(x_1)], \tag{4}$$

where φ is an increasing function from $[0, 1]$ to $[0, 1]$. Experimental studies by Kahneman and Tversky have shown that this function is, in general, similar to that of Fig. 6, whose equation is $\varphi(x) = e^{-\sqrt{-\ln(x)}}$.

Definition 6 (*Rank Dependent Utility (RDU)*) An agent behaves according to the RDU model if her preference relation over the set of lotteries $\mathscr{L}$ is representable by two functions u and φ, where u is the von Neumann-Morgenstern utility over the set of consequences and $\varphi : [0, 1] \mapsto [0, 1]$ is an increasing function such that $\varphi(0) = 0$ and $\varphi(1) = 1$. The agent assigns to every lottery $L = \langle x_1, p_1 \ ; \ \ldots, x_n, p_n \rangle$ such that $u(x_1) \leq u(x_2) \leq \cdots \leq u(x_n)$ utility:

$$RDU(L) = u(x_1) + \sum_{i=2}^{n} \left[\varphi \left(\sum_{k=i}^{n} p(x_k) \right) [u(x_i) - u(x_{i-1})] \right]. \tag{5}$$

As an example, if $u(x) = x/2$ and $\varphi(x) = x^2$, then, to compute the RDU value of lottery $L = \langle 3, 0.2 \ ; \ 10, 0.4 \ ; \ 5, 0.1 \ ; \ 9, 0.3 \rangle$, consequences must first be sorted in increasing utility order: $L = \langle 3, 0.2 \ ; \ 5, 0.1 \ ; \ 9, 0.3 \ ; \ 10, 0.4 \rangle$. Then, the application of Eq. (5) yields:

$$RDU(L) = \varphi(1) \times \frac{3}{2} + \varphi(0.8) \times \left[\frac{5}{2} - \frac{3}{2} \right] + \varphi(0.7) \times \left[\frac{9}{2} - \frac{5}{2} \right] + \varphi(0.4) \times \left[\frac{10}{2} - \frac{9}{2} \right].$$

There exist alternative definitions of RDU. Let us show one of them that will prove useful for highlighting the connection between RDU and another more general model: Choquet expected utility.

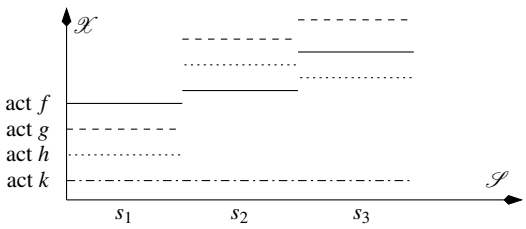


Fig. 7 Illustration of comonotonicity

Definition 7 (*Rank dependent utility (RDU)*) Let u and φ be the functions defined in Definition 6. Let X be a random variable whose probability distribution is P. Then:

$$RDU(X) = \int_{-\infty}^{0} [\varphi(P(u(X) > t)) - 1]dt + \int_{0}^{\infty} \varphi(P(u(X) > t))dt.$$

Note that the Allais paradox can be explained by RDU. This is notably the case when utility u is linear and the probability transform φ is like the one suggested by Kahneman and Tversky: $\varphi(x) = e^{-\sqrt{-\ln(x)}}$. The expressive power of RDU is therefore higher than that of EU. It generalizes the latter since, when $\varphi(x) = x$, RDU boils down to EU. Note also that, when $\varphi(p) \leq p$ for every p, the agent always underestimate the probabilities of the utility increases $u(x_i) - u(x_{i-1})$ (see Eqs. (3) and (4)). This can be interpreted as a kind of pessimism under risk (since the agent takes more into account the worst consequences than the best ones).

The axiomatic foundations of RDU are quite complicated (Quiggin 1982; Wakker 1994; Chateauneuf 1999), so in this chapter, we will not detail them. However, to let the reader understand the key feature of RDU, we will now focus on RDU's main properties: the comonotonic independence axiom in von Neumann-Morgenstern's framework and the comonotonic sure thing principle in Savage's framework (Chew and Wakker 1996). Here, we chose to present only the latter because it is somewhat simpler to understand than the former. For this purpose, we need to define "*comonotonic acts*": two acts f and g are said to be comonotonic if there exists no pair of states of nature $s, s' \in \mathscr{S}$ such that $f(s) \succ_{\mathscr{X}} f(s')$ and $g(s) \prec_{\mathscr{X}} g(s')$. Intuitively, two acts are comonotonic if their variations (in terms of preferences over the consequences) do not evolve in the opposite directions when moving from one state of nature to another. For instance, in Fig. 7, in which preferences over the consequences increase along the vertical axis, f and g are comonotonic, as well as g and k, and h and k. But g and h are not comonotonic because $g(s_3) \succ_{\mathscr{X}} g(s_2)$ and $h(s_2) \succ_{\mathscr{X}} h(s_3)$. Note that comonotonicity is not a transitive property since g and k are comonotonic, as well as k and h, but g and h are not comonotonic. The key idea of RDU consists of imposing the "Sure Thing Principle" only over comonotonic acts:

Axiom 10 (comonotonic sure thing principle) *Let* $\{A_1, \ldots, A_n\}$ *be a partition of* $\mathscr{S}$ *and let* $f : A_i \mapsto x_i$ *and* $g : A_i \mapsto y_i$ *be two acts such that* $x_1 \leq x_2 \leq \cdots \leq x_n$ *and*

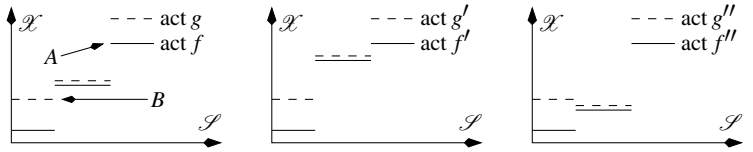


Fig. 8 The comonotonic sure thing principle

Table 4 The Allais paradox and the comonotonic acts

act	A_1 ($P(A_1) = 0.01$)	A_2 ($P(A_2) = 0.89$)	A_3 ($P(A_3) = 0.10$)
L_1	1M €	1M €	1M €
L_2	0 €	1M €	5M €
L'_1	1M €	0 €	1M €
L'_2	0 €	0 €	5M €

$y_1 \leq y_2 \leq \cdots \leq y_n$. *Assume that there exists* $i_0 \in \{1, \ldots, n\}$ *such that* $x_{i_0} = y_{i_0}$. *Let* $f' : A_i \mapsto x'_i$ *and* $g' : A_i \mapsto y'_i$ *be two other acts such that:*

$$\begin{cases} x'_{i_0} = y'_{i_0}; \quad \textit{and} \quad x'_i = x_i \textit{ and } y'_i = y_i \textit{ for every } i \neq i_0, \\ x'_1 \leq \cdots \leq x'_n \textit{ and } y'_1 \leq \cdots \leq y'_n. \end{cases}$$

Then $f \succsim g \Longrightarrow f' \succsim g'$.

This principle is illustrated in Fig. 8: The common part of acts f and g can vary only between points A and B. Thus, acts f' and g' satisfy the constraints of the above definition, which is not the case for acts f'' and g''. Table 4 shows the acts corresponding to the Allais paradox mentioned in the preceding section. In this table, the A_i's are sorted in such a way that acts L_1 and L_2 correspond to f and g of Axiom 10. It can be seen that quadruple (L_1, L_2, L'_1, L'_2) does not satisfy the premises of Axiom 10 (see the difference between L_1 and L'_1). As a consequence, the Allais paradox does not violate the comonotonic sure thing principle. This is the reason why RDU can explain why agents prefer L_1 to L_2 and L'_2 to L'_1.

The RDU model is in fact a particular case of a more general model: Choquet expected utility (CEU), that we will briefly describe after introducing the concept of capacity:

Definition 8 (*Capacity*) A capacity $\mu : 2^{\mathscr{S}} \mapsto [0, 1]$, where $\mathscr{S}$ is the set of states of nature, is a function satisfying the following two properties:

1. $\mu(\emptyset) = 0$ and $\mu(\mathscr{S}) = 1$;
2. For every pair $A, B \subseteq \mathscr{S}$, we have that $A \subseteq B \Longrightarrow \mu(A) \leq \mu(B)$.

Here, a capacity must be understood as a generalization of the concept of probability distribution.[6] Indeed, any probability distribution satisfies properties (1) and (2) above. This is also the case for all the probability transforms of the RDU model. Therefore, capacities allow to define a more general decision model:

Definition 9 (*Choquet expected utility (CEU)*) An agent behaves according to the CEU model if her preference relation over the set of acts $\mathscr{X}^{\mathscr{S}}$ is representable using two functions u and μ, where u is the utility function over the consequences and $\mu : 2^{\mathscr{S}} \mapsto [0, 1]$ is a capacity. The agent assigns to each act f utility:

$$CEU(f) = \int_{Ch} u(f)d\mu = \int_{-\infty}^{0} [\mu(u(f) > t) - 1]dt + \int_{0}^{\infty} \mu(u(f) > t)dt. \tag{6}$$

It has been proved in Wakker (1990) that CEU reduces to RDU when Axiom 11 below is added to the axiomatics of CEU (Schmeidler 1986; Gilboa 1987; Wakker 1990). It is generally believed that this axiom is attractive for a "rational" decision model since it expresses the fact that if, for every consequence x, the probability of winning at least x is higher with act f than with act g, the agent should prefer f to g.

Definition 10 (*First order stochastic dominance*) For every act h, let $F_h(x) = P(\{s \in \mathscr{S} : h(s) \preceq x\})$ denote the cumulative distribution of h. Let f and g be two acts and let F_f and F_g be their respective cumulative distributions. Then f stochastically dominates g at the first order if, for every $x \in \mathbb{R}$, we have that $F_f(x) \leq F_g(x)$.

Axiom 11 (First order stochastic dominance) *Let f and g be two acts. If f stochastically dominates g at he first order, then $f \succsim g$.*

We will see again the CEU model and its usefulness for decision making under uncertainty in the next section. To complete our overview of RDU, we must mention some results about risk aversion.We have seen earlier that, in the EU model, strong risk aversion is equivalent to weak risk aversion, which also corresponds to the concavity of the von Neumann-Morgenstern utility u. Is this also the case in RDU? A first answer to this question can be found in Chew et al. (1987), where it is proved that a RDU agent is strongly risk averse if and only if her utility u is concave and her probability transform φ is convex. Similarly, the agent is strongly risk seeking if and only if u is convex and φ is concave. To our knowledge, there does not exist yet any complete characterization of weak risk aversion in the RDU model. Only sufficient conditions have been proposed and those do not require the concavity of u (Chateauneuf and Cohen 1994). In terms of risk aversion, the expressive power of RDU is therefore higher than that of EU. Finally, note that other concepts of risk aversion designed specifically for RDU have been proposed. Those are different from both strong and weak risk aversions. For instance, Quiggin suggested to substitute

[6]For an interpretation in terms of weights of agents' coalitions or of criteria, see chapter "Multicriteria Decision Making" of this volume.

strong risk aversion by monotonic risk aversion (Quiggin 1992): let X and Y be two random variables. Y is said to be a monotonic mean preserving spread (MMPS) of X if $Y = X + Z$, where Z is a white noise, and X and Z are comonotonic. An agent is monotonic risk averse if she does not like monotonic risk increase, i.e., if $Y = \text{MMPS}(X)$, then $X \succsim Y$.

Up to now, we have only studied decision models relying on the existence of probability distributions to model uncertainties. But, what can we do if there does not exist sufficient information to construct one, like in the Ellsberg's urn example? The goal of the next section is to provide some keys to answer to this question.

4 Decision Models Outside the Probabilistic Framework

Let us recall the Ellsberg's urn problem: this is an urn containing 99 balls, which can be either red, yellow or black. The only information available to the agent is that one third of the balls is red and the remaining two third are either yellow or black (but their respective proportions are unknown). Agents bet on the color of a ball to be drawn from the urn. Thus, an agent is asked which alternative she prefers between alternatives A and B below, and which one she prefers between C and D:

- Alternative A: win 1M € if the drawn ball is red, else win 0 €,
- Alternative B: win 1M € if the drawn ball is black, else win 0 €,
- Alternative C: win 1M € if the drawn ball is red or yellow, else win 0 €,
- Alternative D: win 1M € if the drawn ball is black or yellow, else win 0 €.

Most of the human agents prefer A to B and D to C. As we have seen before, EU cannot account for such preferences (violation of the sure thing principle). RDU can neither model these preferences. Indeed, if it could then, assuming that the agent prefers winning more money than less, and denoting by P_r, P_y, P_b the probabilities that the drawn ball is red, yellow and black respectively, we have that $A \succ B \iff \text{RDU}(A) > \text{RDU}(B) \iff \varphi(P_r) > \varphi(P_b)$ and $D \succ C \iff \varphi(P_b + P_y) > \varphi(P_r + P_y)$. But this is impossible to have both inequalities satisfied because φ is an increasing function. Here, the problem is that there does not exist a unique probability distribution compatible with the information available to the agent. Therefore, in this case, we should not try to use a decision model that relies on a unique probability distribution but rather on a model that relies on the set of all the distributions compatible with the available information. Here, it is easy to see that this set is convex: if P and Q are two compatible probability distributions, for every $\alpha \in [0, 1]$, we have that $\alpha P + (1 - \alpha)Q$ is also compatible with the available information. As a consequence, to represent the uncertainties in the Ellsberg's urn, it is sufficient to know the boundary of the convex hull of all the compatible distributions. But since the probability of any event and that of its complementary event sum always to 1, the lower bounds on the probabilities are sufficient to characterize all the convex hull. Those correspond to a function $\mu : 2^{\mathscr{S}} \mapsto [0, 1]$ such that, for every $A \subseteq \mathscr{S}$, $\mu(A) = \min_{\{P\text{compatibles}\}} P(A)$. For the Ellsberg's urn, this function

Table 5 The belief function of the Ellsberg's urn and its Möbius inverse

Evt	$\emptyset$	$\{R\}$	$\{Y\}$	$\{B\}$	$\{R, Y\}$	$\{R, B\}$	$\{Y, B\}$	$\mathscr{S}$
f	0	1/3	0	0	1/3	1/3	2/3	1
ϕ	0	1/3	0	0	0	0	2/3	0

μ, also called a "belief function", is described in Table 5. Indeed, the min of $P(Y)$ is equal to 0 because it is possible that the urn contains no yellow ball. On the other hand, min $P(Y, B) = 2/3$ because, for all the probability distributions P compatible with the Ellsberg's urn, we have $P(Y, B) = 2/3$. More formally, belief functions (Dempster 1967; Shafer 1976) are defined as follows (see chapter "Representations of Uncertainty in AI: Beyond Probability and Possibility" of this volume):

Definition 11 (*Belief function*) A belief function μ is a capacity (in the sense of Choquet) which is ∞-monotone, i.e., it is such that for all $n \geq 2$, and for all $A_1, \ldots, A_n \in 2^{\mathscr{S}}$:

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{\emptyset \subset I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} \mu\left(\bigcap_{i \in I} A_i\right).$$

To any capacity (and *a fortiori* to any belief function) is associated its *Möbius inverse* ϕ defined by: $\phi(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \mu(B)$ for every $A \subseteq \mathscr{S}$. Intuitively, ϕ represents the information/the belief about the realization of event A that is not captured in its subevents. For instance, in Table 5, $\phi(\{R, Y\}) = 0$ because the agent has no more information about the chances of realization of R or Y than she has of R alone because there is no information available on the proportion of yellow balls in the urn. Above, we have characterized ϕ in terms of μ but it is also possible to characterize μ in terms of ϕ. Indeed, it is not difficult to show that $\mu(A) = \sum_{B \subseteq A} \phi(B)$ for all events A. This formula simply states that the agent's belief about event A corresponds to the sum of all her "elementary" beliefs on the realizations of A's subevents. Thus, Belief $\mu(\{R, Y\})$ about the realization of event $\{R, Y\}$ corresponds to the belief generated by the sum of the information available about R alone, Y alone and the compound (R or Y) that could not be captured in singletons $\{R\}$ and $\{Y\}$. From a mathematical point of view, this translates as $\mu(\{R, Y\}) = \phi(\{R\}) + \phi(\{Y\}) + \phi(\{R, Y\})$.

In (Jaffray 1989), Jaffray observed that the set of all the belief functions is a *mixture set*, i.e., it is closed under mixture operations or, in other words, any convex combination of belief functions is another belief function. In addition, he remarked that this is the key property used by von Neumann and Morgenstern to develop their axiomatic foundation of EU. As a consequence, it is possible to substitute in each axiom probabilities by belief functions. Expected utility thus boils down to a Choquet integral with respect to capacity μ. More precisely, in von Neumann-Morgenstern's framework, the probability distribution over the space of the states of nature generates, for each decision, a probability distribution over the outcomes

of the decision, which is translated as a lottery. Here, Jaffray showed that the belief function over the space of the states of nature generates, for each decision, a belief function over the space of consequences. Let us call $\mathscr{G}$ the space of these functions.

Theorem 6 (Jaffray 1989) *The two assertions below are equivalent:*

1. *Preference relation $\succsim$ over $\mathscr{G}$ satisfies Axioms 1, 2, 3, where lotteries over $\mathscr{L}$ are substituted by $\mathscr{G}$, the set of belief functions over the space of consequences.*
2. *$\succsim$ is representable by a utility function*

$$U : \mathscr{G} \mapsto \mathbb{R}$$

such that $U(\mu) = \int u d\mu$.

Function $u : \mathscr{X} \mapsto \mathbb{R}$ is called the von Neumann-Morgenstern utility function of the agent and is unique up to scale and location.

Therefore, the Choquet integral provides an attractive decision framework for situations in which probabilities are inadequate to model uncertainties. Thanks to the following definitions, it can be appropriately redefined in terms of Möbius inverses rather than belief functions: a belief function e_B is said to be elementary and concentrated on B if $e_B(A) = 1$ when $A \supseteq B$ and $e_B(A) = 0$ otherwise. In other words, its Möbius inverse ϕ_B is such that $\phi_B(B) = 1$ and $\phi_B(A) = 0$ for every $A \neq B$. Let μ be a belief function whose Möbius inverse is ϕ. The focal set $\mathscr{C}_\mu$ of μ is defined as $\mathscr{C}_\mu = \{B : \phi(B) > 0\}$. From these two definitions, it can be inferred that, for every belief function μ, and for every consequence set A, $\mu(A) = \sum_{B \subseteq A} \phi(B) = \sum_{B \in \mathscr{C}_\mu} \phi(B) e_B(A)$. But Theorem 6 trivially implies that, for every convex combination $\{\lambda_i,\ i = 1, \ldots, n : \lambda_i \geq 0 \text{ and } \sum_{i=1}^n \lambda_i = 1\}, U\left(\sum_{i=1}^n \lambda_i \mu_i\right) = \sum_{i=1}^n \lambda_i U(\mu_i)$. As a consequence, if $\mu = \sum_{B \in \mathscr{C}_\mu} \phi(B) e_B$, $U(\mu) = \sum_{B \in \mathscr{C}_\mu} \phi(B) U(e_B)$. Let us denote by $u(B) = U(e_B)$ the utility of set of consequences B. Then, we get a linear utility model called *Belief expected utility (BEU)*:

Theorem 7 (Belief expected utility (BEU) – Jaffray 1989) *The following two assertions are equivalent:*

1. *Preference relation $\succsim$ over $\mathscr{G}$ satisfies Axioms 1, 2, 3, where lotteries over $\mathscr{L}$ are substituted by belief functions over $\mathscr{G}$.*
2. *$\succsim$ is representable by a utility function $U : \mathscr{G} \mapsto \mathbb{R}$ such that $U(\mu) = \sum_{B \in \mathscr{C}_\mu} \phi(B) u(B)$, where $u(B)$ is the utility of set of consequences B and ϕ is the Möbius inverse of μ.*

Table 6 illustrates the computation of U on the four alternatives A, B, C, D of the Ellsberg's urn. Assume that $u(\{0\}) = 0, u(\{1\text{M}\}) = 1$ and $u(\{0, 1\text{M}\}) = \alpha$. Then $A \succ B$ and $D \succ C$ is equivalent to $\alpha < 1/2$. Therefore BEU is capable of representing "common" agent's preferences on the Ellsberg's urn.

However, the BEU formula clearly highlights its limits w.r.t. EU: in the EU model, the agent's utility function u needs be elicited only over the space of consequences $\mathscr{X}$ whereas with BEU or CEU, it must be elicited on $2^{\mathscr{X}}$. Unfortunately, elicitation, i.e., the learning of the agent's preferences, is a complex and time consuming process.

Table 6 BEU utilities for the Ellsberg's urn problem

Evts	0	1M€	{0,1M€}
balls	$\{B, Y\}$	$\{R\}$	$\mathscr{S}$
μ	2/3	1/3	1
ϕ	2/3	1/3	0

$BEU(A) = 2/3u(\{0\}) + 1/3u(\{1\text{M}\}) = 1/3$

Evts	0	1M€	{0,1M€}
balls	$\{R, Y\}$	$\{B\}$	$\mathscr{S}$
μ	1/3	0	1
ϕ	1/3	0	2/3

$BEU(B) = 1/3u(\{0\}) + 2/3u(\{0,1\text{M}\}) = 2/3\alpha$

Evts	0	1M€	{0,1M€}
balls	$\{B\}$	$\{R, Y\}$	$\mathscr{S}$
μ	0	1/3	1
ϕ	0	1/3	2/3

$BEU(C) = 1/3u(\{1\text{M}\}) + 2/3u(\{0,1\text{M}\})$

Evts	0	1M€	{0,1M€}
balls	$\{R\}$	$\{B, Y\}$	$\mathscr{S}$
μ	1/3	2/3	1
ϕ	1/3	2/3	0

$BEU(D) = 1/3u(\{0\}) + 2/3u(\{1\text{M}\}) = 2/3$

Therefore, to fix this problem, Jaffray proposed to add a new axiom called a "*dominance*" axiom to BEU. This axiom expresses the fact that, without any knowledge, within a set of consequences $\{x_1, \ldots, x_k\}$, the agent has no reason to believe that a consequence is more likely to be yielded than any other. So the agent can summarize the information about the set of consequences by defining her preferences taking into account only the worst and the best consequences of the set. Consequently, utility $u(B)$ of a set of consequences B boils down to utility $u(m_B, M_B)$, where m_B and M_B denote the worst and the best consequences of B respectively.

Axiom 12 (Dominance) *For every set of consequences $B \subseteq \mathscr{X}$, let m_B and M_B denote the worst and the best consequences of B respectively. Let e_B be the elementary belief function concentrated on B. Then, for every $B, B' \subseteq \mathscr{X}$, if $m_B \succsim_{\mathscr{X}} m_{B'}$ and $M_B \succsim_{\mathscr{X}} M_{B'}$ then $e_B \succsim e_{B'}$.*

Theorem 8 (Jaffray's model 1989) *The following two assertions are equivalent:*

1. *Preference relation $\succsim$ over $\mathscr{G}$ satisfies Axioms 1, 2, 3 and 12, where lotteries over $\mathscr{L}$ are substituted by belief functions over $\mathscr{G}$.*
2. *$\succsim$ is representable by a utility function $U : \mathscr{G} \mapsto \mathbb{R}$ such that*

$$U(\mu) = \sum_{B \in \mathscr{C}_\mu} \phi(B) u(m_B, M_B).$$

Functions U and u are unique up to scale and location. In addition, u is a nondecreasing function of m and M and the von Neumann-Morgenstern utility $u(x)$ of consequence x is equal to $u(x, x)$.

As a consequence of this theorem, utility $u(m, M)$ takes into account two factors: (i) the attitude of the agent w.r.t. risk (concavity of $u(x, x)$), but also (ii) the attitude w.r.t. ambiguity when $M \neq m$. The model can be further refined using the Hurwicz criterion (Hurwicz 1951):

Definition 12 (Hurwicz criterion) for every (m, M), let us call the "local optimism/pessimism criterion" the value $\alpha(m, M)$ for which the agent is indifferent between the following two alternatives:

1. winning m with probability $\alpha(m, M)$ and M with probability $1 - \alpha(m, M)$,
2. winning at least m and at most M, without any further information.

Thanks to this criterion, utility $u(m, M)$ can be redefined as $\alpha(m, M)u(m) + [1 - \alpha(m, M)]u(M)$, with $u(x)$ the von Neumann-Morgenstern utility function of the agent. In this context, coefficient α expresses the attitude of the agent w.r.t. ambiguity and the concavity of u expresses the agent's attitude w.r.t. risk. Now, the task of eliciting the agent's preferences (the learning of function u) has a complexity similar to that in the EU model.

4.1 Qualitative Decision Models Under Uncertainty

In parallel to the research works made in the field of mathematical economics, decision under uncertainty has received attention in artificial intelligence. In particular, researchers investigated qualitative models, which describe preferences only through ordinal information (Tan and Pearl 1994; Boutilier 1994; Dubois and Prade 1995; Brafman and Tennenholtz 1996; Lehmann 1996; Dubois et al. 1997). Thus, within the framework of possibilistic lotteries (Dubois and Prade 1995), Dubois and Prade proposed a counterpart to the von Neumann-Morgenstern axiomatic foundation: they axiomatized "*qualitative utilities*", which generalize Wald criterion (Wald 1950) for comparing possibility distributions. A possibility distribution is characterized by a function π which assigns to each consequence x its possibility $\pi(x) \in L$, L being an ordered set. The pessimistic qualitative utility model is based on an L-valued utility function u defined over the set of consequences $\mathscr{X}$, with L an ordered set. This function assigns to every possibilistic lottery π the following value:

$$U^{-}(\pi) = \min_{x \in \mathscr{X}} \max\{n(\pi(x)), u(x)\}$$

where n is a decreasing function which inverses the scale of L. Typically, when $L = [0, 1]$, n is chosen as $n(x) = 1 - x$. Value U^{-} indicates to which extent, by choosing π, the agent is sure to get a consequence having a "good" utility value. In the same possibilistic framework, there exists a more optimistic version which evaluates to which extent it is possible that the agent gets a consequence with a "good" utility value. This version consists of assigning to every possibilistic lottery the following quantity:

$$U^{+}(\pi) = \max_{x \in \mathscr{X}} \min\{\pi(x), u(x)\}$$

The axiomatic foundation of Savage has also been revisited in order to propose qualitative counterparts to the EU model. Thus, Dubois, Prade and Sabbadin (Dubois et al. 1998) proposed axiomatic justifications for the optimistic and pessimistic qualitative utility criteria when comparing acts in the sense of Savage. This led to the following models:

$$U^-(f) = \min_{s\in\mathscr{S}} \max\{n(\pi(s)), u(f(s))\}$$
$$U^+(f) = \max_{s\in\mathscr{S}} \min\{\pi(s), u(f(s))\}$$

For every act f in $\mathscr{X}^{\mathscr{S}}$. $U^+(f)$ evaluates to which extent there exists a consequence of f which is at the same time very good and very plausible. On the other hand, $U^-(f)$ evaluates to which extent all the consequences in act f are plausible and good. These formula are therefore the numerical translations of logic principles. For more details, see Dubois et al. (1999). Dubois, Prade and Sabbadin have also proposed an axiomatic foundation of the Sugeno integral for comparing acts (Dubois et al. 1998), which led to the following model:

$$S_v(f) = \max_{x\in\mathscr{X}} \min\{v(F_x), u(x)\}$$

where $F_x = \{s \in \mathscr{S} : f(s) \geq x\}$ and v is a capacity defined on $2^{\mathscr{S}}$.

These models depart from EU notably by their weakening of the "sure thing principle", which becomes the "weak sure thing principle":

Axiom 13 (Weak Sure Thing Principle) *For every* $f, g, h, h' \in \mathscr{X}^{\mathscr{S}}$ *and for every* $A \in 2^{\mathscr{S}}$*, we have that* $fAh \succ gAh \Rightarrow fAh' \succsim gAh'$.

This axiom is important because, although it is weaker than the sure thing principle, it is sufficient to enable the computation of optimal policies in dynamic decision problems by backward induction. For more details on this point, see Sabbadin (1998).

Finally, pure ordinal aggregation rules (derived from majority rules used in voting) have been proposed under the name of "*lifting rules*" (Dubois et al. 2002, 2003). To compare acts, they only rely on relative events likelihoods and on a preference relation over the consequences. They are defined as:

$$f \succsim g \text{ if and only if } \{s \in \mathscr{S} : f(s) \succsim_{\mathscr{X}} g(s)\} \trianglerighteq \{s \in S : g(s) \succsim_{\mathscr{X}} f(s)\}$$

where $\succsim_{\mathscr{X}}$ is the projection on the consequence scale of preference relation $\succsim$ restricted to the constant acts, and $\trianglerighteq$ is a relative likelihood relation over the events. Their axiomatic justification is based on the introduction, in Savage's framework, of an axiom compelling the purely ordinal nature of the rule (Dubois et al. 2002, 2003):

Axiom 14 (Ordinal invariance) [*for every* $s \in \mathscr{S}$, $(f(s) \succsim_{\mathscr{X}} g(s)$ *if and only if* $f'(s) \succsim_{\mathscr{X}} g'(s))$ *and* $(g(s) \succsim_{\mathscr{X}} f(s)$ *if and only if* $g'(s) \succsim_{\mathscr{X}} f'(s))$]
$\Longrightarrow$ ($f \succsim g$ *if and only if* $f' \succsim g'$).

This axiom states that preference $f \succsim g$ among two acts f and g, characterized by consequence vectors $(f(s_1), \ldots, f(s_n))$ and $(g(s_1), \ldots, g(s_n))$ respectively, does not depend on the relative positions of these consequences in the agent's preference scale, i.e., it only depends on preferences $f(s) \succsim_{\mathscr{X}} g(s)$ and $g(s) \succsim_{\mathscr{X}} f(s)$ for all the states of nature $s \in \mathscr{S}$. This model reminds of the relative concordance rules introduced in chapter "Multicriteria Decision Making" of this volume in multicriteria decision making. Such rules do not necessarily induce transitive preferences, except when the beliefs over the events are highly hierarchical systems (see Dubois et al. 2002, 2003 for more details). Here again, in order to obtain transitive preferences without constraining arbitrarily the beliefs over the events, it can be advantageous to introduce reference points in the model and to propose rules like:

$$f \succsim_r g \text{ if and only if } \{s \in \mathscr{S} : f(s) \succsim_{\mathscr{X}} r\} \trianglerighteq \{s \in \mathscr{S} : g(s) \succsim_{\mathscr{X}} r\}$$

in which r represents a reference consequence on scale $\mathscr{X}$. For more details on this type of models, see Perny and Rolland (2006).

5 Sequential Decision Models

In practical situations, a decision is seldom made independently of the other decisions. Therefore, agents often have to choose among sets of decisions that must be made consecutively, each one having some impact on the next ones. In this section, we will study such problems and some decision models that were designed for that purpose.

Graphical models are well-suited for this task. "*Decision trees*" are certainly one of the most popular models. Their graphs contain two types of nodes: "*decision nodes*", drawn as rectangles, which represent the alternatives among which the agent has to choose; and "*chance nodes*", draw as circles, which represent the uncertainties about the events. All these nodes are put into the graph in such a way that time always increases from the left to the right of the graph. Finally, to the leaves of the tree are assigned the utilities of the consequences resulting from the sequence of decisions and the set of events made from the root of the tree up to the leaves. Figure 9 represents a simple decision tree corresponding to the following problem (Raiffa 1968): An oil wildcatter must decide either to drill or not. He is uncertain whether the hole is dry, wet or soaking. If he decides to drill, then, his gain will depend on the quantity of oil in the hole: if the hole is dry (no oil), he will loose 1M €; if the hole is wet, he will win 2M €; finally, if the hole is soaking, he will win 10M €. At a cost of 10K €, the wildcatter can make seismic soundings which help determine the geological structure of the site. The soundings will disclose whether the terrain below has no structure (*NoS*), in which case there is not much chance that the hole contains some oil, or open structure (*OpS*), in which case the presence of oil is somewhat more probable, or closed structure (*ClS*), in which case there are high chances that the hole contains a lot of oil. This problem can be modeled by a decision tree in the following way: The first decision to be made consists of making or not seismic soundings. This

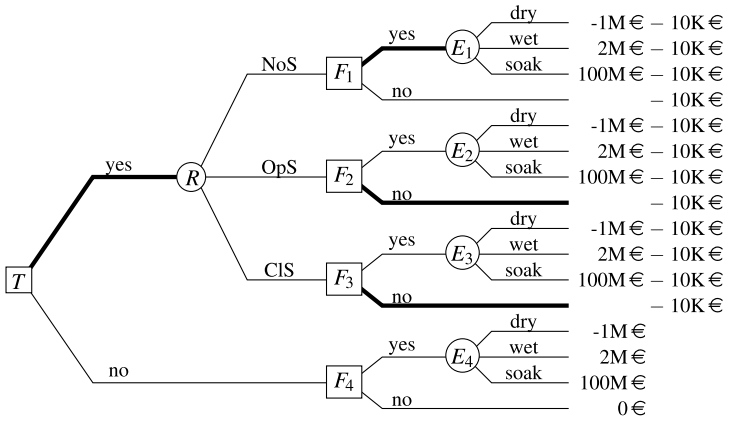


Fig. 9 Decision tree for the oil wildcatter problem

decision is represented by node T in Fig. 9. If the oil wildcatter decides to make the soundings, we pass through the upper branch, else in the lower branch. Once the test is made, the wildcatter gets back the result R of the test. Of course, this result is only known after making the seismic soundings and, therefore, after making the decision to make the seismic soundings. This is the reason why node R must be located on the right of node T (time increases from left to right). Whatever the result of the test, upon knowing this result, the oil wildcatter must decide whether he will drill or not (nodes F_1). If he decides not to drill, then he will have lost the price of the seismic soundings, i.e., 10K €. This information can be found on the leaves of the tree. If the oil wildcatter decides to drill, then he will win the amount of money depending on the quantity of oil in the hole minus the price of the seismic soundings. This quantity (the E_i's) is unknown when the agent makes the decision to drill or not, hence the E_i's must be located on the right of F_1 in the decision tree. In the end, we get Fig. 9. In general, on the branches outgoing chance nodes, are indicated the beliefs the agent has that the events will occur. Those are often the conditional probability that the event will occur given the values taken by all the preceding nodes, i.e., all the nodes to the left, up to the root. For instance, on the upper branch on the right of E_3 should be stored $P(E_3 = \text{dry}|F_3 = \text{yes}, R = \text{ClS}, T = \text{yes})$. Some variables can be independent from others, so this expression can often be simplified. Here, for instance, it is obvious that the state of the hole does not depend on the decisions of the agent, so the above conditional probability is equivalent to $P(E_3 = \text{dry}|R = \text{ClS})$. Probabilities on the branches outgoing the E_i's therefore differ from one E_i to the other.

In addition to their capacity to model sequential decision making problems, decision trees can also be exploited to help agents making the best decisions. For this purpose, whatever the decision criterion chosen (EU, RDU, *etc.*), the idea is to look for an *optimal strategy*, i.e., in *every* decision node accessible given the set of all the decisions made previously, the choice of an alternative/decision among those

possible at that node. Thus, a strategy considers all the states of nature possible. For instance, in Fig. 9, the set of bold edges represents a strategy: when T="yes" is selected, as it is not possible to know which value R will take, we need to consider all the possible values for R and an alternative needs be selected for each node F_i. Note that, when the uncertainty within the chance nodes is modeled by probabilities, a strategy precisely corresponds to a lottery. Indeed, consider the strategy in bold edges in Fig. 9. This one represent the fact that the agent will loose 10K € if $R =$ "Ops" or $R =$ "ClS" and that, if $R =$ "NoS", he will win 100M–10K € if $E_1 =$ "soak", 2M–10K € if $E_1 =$ "wet" and –1M–10K € if $E_1 =$ "dry". Therefore, this corresponds to lottery:

$$\langle \text{–10K €}, P(R = \text{Ops or Cls}) \ ; \ \text{100M–10K €}, P(R = \text{NoS}, E_1 = \text{soak}) \ ; \\ \text{2M–10K €}, P(R = \text{NoS}, E_1 = \text{wet}) \ ; \ \text{–1M–10K €}, P(R = \text{NoS}, E_1 = \text{dry}) \rangle.$$

Therefore, finding the EU optimal strategy in a decision amounts to find the strategy whose corresponding lottery is optimal, i.e., it is maximal w.r.t. the EU criterion. Luckily, to determine it, it is not necessary to compute all the lotteries and to extract the best one. Actually, the above strategy can be rewritten as follows:

$$\begin{array}{rl} \langle & \text{–10K €} \ , \ P(R = \text{Ops}) \ ; \\ & \text{–10K €} \ , \ P(R = \text{Cls}) \ ; \\ & \text{100M–10K €} \ , \ P(R = \text{NoS}) \times P(E_1 = \text{soak} | R = \text{NoS}) \ ; \\ & \text{2M–10K €} \ , \ P(R = \text{NoS}) \times P(E_1 = \text{wet} | R = \text{NoS}) \ ; \\ & \text{–1M–10K €} \ , \ P(R = \text{NoS}) \times P(E_1 = \text{dry} | R = \text{NoS}) \ \rangle. \end{array} \tag{7}$$

Remark that the last three lines correspond to $P(R = \text{NoS})$ times the following lottery:

$$\begin{array}{rl} \langle & \text{100M–10K €} \ , \ P(E_1 = \text{soak} | R = \text{NoS}) \ ; \\ & \text{2M–10K €} \ , \ P(E_1 = \text{wet} | R = \text{NoS}) \ ; \\ & \text{–1M–10K €} \ , \ P(E_1 = \text{dry} | R = \text{NoS}) \ \rangle. \end{array} \tag{8}$$

which is nothing else than the lottery resulting from the bold strategy in the subtree whose root is F_1. If, in the bold strategy of Fig. 9, Decision $F_1 =$ "yes" is substituted by $F_1 =$ "no", it is easy to see that the resulting lottery will differ from that of Eq. (7) only by the last three lines of the latter that are substituted by $P(R = \text{NoS})$ times lottery $\langle \text{–10K €}, 1 \rangle$, which is nothing else than the lottery corresponding to the strategy of the subtree rooted at the lower branch of F_1. Consequently, to compare according to the EU criterion two lotteries L_1, L_2 that differ only in a subtree of the decision tree, it is sufficient to compute their respective lotteries in this subtree and to select the one with the highest EU score. As a matter of fact, the expectations of the sub-lotteries in the other subtrees are identical for both L_1 and L_2 so, due to the linearity of EU, they are irrelevant to compare L_1 and L_2. This justifies that the following dynamic programming-based algorithm by backward induction can determine the EU-optimal strategy in all the decision tree: first, select the decisions that maximize EU in all the subtrees rooted at the decision nodes that are the closest

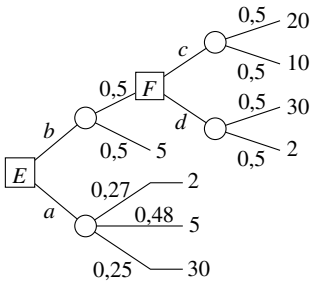


Fig. 10 Decision tree and the RDU criterion

to the leaves of the decision tree (in Fig. 9, this corresponds to the subtrees rooted at $F_i, i = 1, \ldots, 4$, respectively); then substitute these subtrees by leaves whose utility values are the expectations of these decisions, and iterate this process until reaching the root of the decision tree. The decision selected at each step of this algorithm constitute the EU-optimal strategy.

The goal of this chapter is not to develop computational decisional algorithmics, so we will not detail further this backward induction mechanism. However, it was useful to mention it when considering features of the "new" decision models like RDU. Actually, for these nonlinear models, backward induction produces incorrect results, as we will show in the next example. Suppose that the probability transformation function of the agent is $\varphi(x) = e^{-\sqrt{-ln(x)}}$, as suggested by Kahneman and Tversky, and that her utility function is $u(x) = x$. Now, consider the decision tree of Fig. 10. On the arcs outgoing from the chance nodes are indicated the probabilities of occurrence of their respective events and, on the right of the leaves are displayed the utilities of the consequences of the decisions. Calculating the RDU values of the strategies in this decision tree, we have that:

$$\begin{aligned}
RDU(a) &= 2 + (5-2)\varphi(0,73) + (30-5)\varphi(0,25) = 11,41\\
RDU(bc) &= 5 + (10-5)\varphi(0,5) + (20-10)\varphi(0,25) = 10,26\\
RDU(bd) &= 2 + (5-2)\varphi(0,75) + (30-5)\varphi(0,25) = 11,46\\
RDU(c) &= 10 + (20-10)\varphi(0,5) = 14,35\\
RDU(d) &= 2 + (30-2)\varphi(0,5) = 14,18.
\end{aligned}$$

In other words, in the subtree rotted at F, Strategy c is preferable to d, but in the subtree rooted at E, the optimal strategy is bd rather than bc.

This phenomenon is not restricted to the RDU criterion: it is general and appears as soon as the criterion departs from EU. In fact, to produce correct results, backward induction requires two properties: consequentialism and dynamic consistency. The first one states that, in each subtree, the optimal strategy depends only on this subtree and not on the rest of the decision tree. The second property states that an optimal strategy in a subtree is an extension of optimal strategies in its own subtrees. As an example, if, in Fig. 10, bd is an optimal strategy for the subtree rooted at E, then d must also be an optimal strategy in the subtree rooted at F. Unfortunately,

consequentialism + dynamic consistency implies the "sure thing principle" (or at least a slightly weakened version) which leads to the EU criterion.

To complete our brief overview of sequential decision making, note that there exist compact representations of decision trees like, for instance, influence diagrams (Howard and Matheson 1984; Shachter 1986; Jensen et al. 1994). The first key idea consists of considering decision trees as representations of "big" multivariate functions. The case of the decision trees with a symmetric structure simplifies the illustration of this idea: consider the trees of Fig. 11. Instead of considering the utility values independently from one leaf to the other, consider the *set* of all these utility values as the result of a function depending on the values of D and O that led to the corresponding leaves. Similarly for the probabilities indicated on the branches of the decision tree, do not consider the values separately but as a whole as the probability distribution $P(O|D)$ depending on the values of D and O. The second key idea consists of exploiting the structural independences inherent to the decision problem. There is often a large number of such independences and those usually greatly simplify the "big" functions mentioned earlier. As an example, observe the 4 decision trees of Fig. 11. At first sight, they seem quite similar. However, upon examining carefully the probabilities and the consequences/utilities displayed beside the branches of the tree, fundamental differences can be observed among these trees. In the first one, probabilities and utilities differ from one another on all the branches and, therefore, none of the functions $P(O|D)$ and $u(D, O)$ can be simplified. This is precisely what is represented by influence diagram 1 in Fig. 12: circles represent chance nodes, to which are associated the conditional probabilities of these nodes given their parents in the graph (like in a Bayesian network (Pearl 1988)); lozenges represent the utility multivariate functions and the variables they depend on correspond to those at the tails of their ingoing arcs. In tree 2 of Fig. 11, it can be noticed that utility values depend on the branch outgoing from O where they are located but they do not depend on D. In other words, utility $u(D, O)$ can be summarized as $u(O)$ and this is precisely what influence diagram 2 of Fig. 12 represents. In Tree 3, utilities

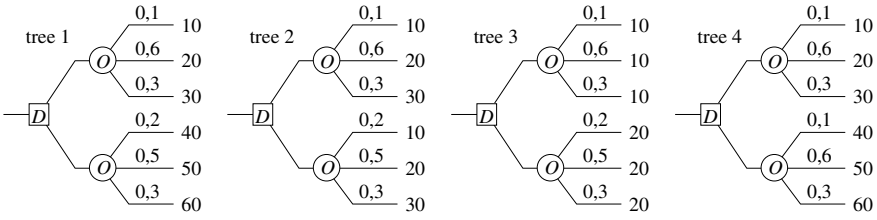


Fig. 11 Structural dependences in decision trees

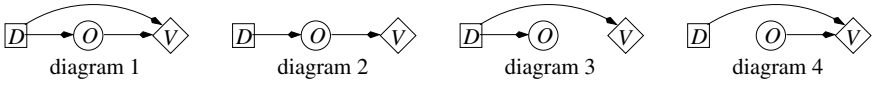


Fig. 12 Influence diagrams

depend on D but not on O, hence influence diagram 3. Finally, in Tree 4, probabilities $P(O|D)$ do not depend on the value of D, which corresponds to influence diagram 4. To complete our description of influence diagrams, note that, although no function is associated with decision nodes, the latter can also have ingoing arcs. In this case, these arcs indicate the nodes (decisions and/or chances) whose values are known to the agent when she makes her decision.

To conclude this section, note that models for representing sequential decision making problems are not restricted to decision trees and their compact representations (e.g., influence diagrams). Other formalisms do exist, which can be better suited for particular tasks. For instance, we can cite Markov decision processes (MDP) (Bellman 1957; Howard 1960; Puterman 1994) or partially observed MDPs (Sondik 1971; Monahan 1982), which are especially useful in planning. Although these models have been based initially on probabilities, their possibilistic counterparts have been proposed in the literature (Fargier et al. 1998; Sabbadin 2001). In this chapter, we will not develop further these models since chapter "Planning in Artificial Intelligence" of Volume 2 is devoted to them.

6 Conclusion

This chapter has provided a brief and non exhaustive overview of the theory of decision making under uncertainty. As we have seen, justifying mathematically the proposed decision making models, relying on simple axioms reflecting commonsense features that are expected to be satisfied by any "*rational*" agent, has been one of the major concerns in the decision theoretic community. These axioms enable to justify to users these models and, more importantly, the recommendations they provide. This is a key point to make human agents/decision makers accept these models. Currently, the main research topics of the field are threefold and are focused on: (i) the elicitation of preferences; (ii) the models of uncertainty and their learning; and (iii) the recommendation algorithms based on these models. Researches on preference elicitation focus on the minimization of the number of questions to ask to the agent to capture her preferences, but also on how to focus questions in order to elicit only the parts of the utility function that are needed to make "good" recommendations (Wang and Boutilier 2003; Gonzales and Perny 2004; Boutilier et al. 2010; Lu and Boutilier 2011). As for the uncertainties, new compact graphical models have been introduced recently (*Probabilistic Relational Models*, *Markov Logic Networks*, *Multi-Entity Bayesian networks*, etc), which notably enable learning from relational databases probability distributions defined over high-dimensional spaces, taking into account generic domain knowledge (Getoor and Taskar 2007; Kok and Domingos 2009; Khosravi et al. 2010). Finally, recommendation algorithms have to address problems over combinatorial spaces of ever increasing sizes (de Salvo Braz et al. 2005; Regan and Boutilier 2011).

For many years, in artificial intelligence, expected utility (EU) has been considered as the only reasonable model for decision under uncertainty. However, these

last years, new decision theoretic models like RDU or Choquet have been introduced in the major AI conferences and their place shall increase in the next years. Indeed, they are not only capable to model faithfully the behaviors of agents facing uncertainty and ambiguity, but they also proved to be very useful for modeling fair and robust decision making problems. Finally, their expressive power should make them the models of choice for preference elicitation for high stakes strategic decision problems. However, exploiting such models requires a high level of information about the preferences of the agents as well as about the likelihoods of the events that may occur. Unfortunately, in some AI decision problems (like planning in partially known environments, preference elicitation and recommendation), the information available does not usually allow to precisely quantify the utility of an action or the probability of an event. In such situations, by relying on an ordinal representation of preferences and uncertainties, the qualitative models presented in the preceding sections prove to be better suited. To a large extent, these models are still unknown outside the academic world but, in the near future, their exploitation in industrial applications should increase significantly.

References

Allais M (1953) Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. Econometrica 21:503–546

Anand P (1993) The philosophy of intransitive preference. Econ J 103(417):337–346

Anscombe F, Aumann R (1963) A definition of subjective probability. Ann Math Stat 34:199–205

Argall BD, Chernova S, Veloso M, Browning B (2009) A survey of robot learning from demonstration. Robot Auton Syst 57:469–483

Arrow KJ (1965) Aspects of the theory of risk bearing, chapter The theory of risk aversion. Yrjo Jahnsson Fondation, pp 90–120

Bellman R (1957) Dynamic programming. Princeton University Press, Princeton

Bernoulli D (1738) Specimen theoriae novae de mensura sortis. Commentarii academiae scientiarum imperialis Petropolitanae 5:175–192

Bleichrodt H (1996) Applications of utility theory in the economic evaluation of health care. PhD thesis, Erasmus University, Rotterdam, the Netherlands

Boutilier C (1994) Towards a logic for qualitative decision theory. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'94), pp 75–56

Boutilier C (2002) A POMDP formulation of preference elicitation problems. In: Proceedings of the national conference on artificial intelligence (AAAI'02), pp 239–246

Boutilier C, Regan K, Viappiani P (2010) Simultaneous elicitation of preference features and utility. In: Proceedings of the national conference on artificial intelligence (AAAI'10), pp 1160–1167

Brafman RI, Tennenholtz M (1996) On the foundation of qualitative decision theory. In: Proceedings of the national conference on artificial intelligence (AAAI'96), pp 1291–1296

Chajewska U, Koller D, Parr R (2000) Making rational decisions using adaptive utility elicitation. In: Proceedings of the national conference on artificial intelligence (AAAI'00), pp 363–369

Chateauneuf A (1999) Comonotonicity axioms and RDU theory for arbitrary consequences. J Math Econ 32:21–45

Chateauneuf A, Cohen M (1994) Risk-seeking with diminishing marginal utility in a non-expected utility model. J Risk Uncertain 9:77–91

Chateauneuf A, Cohen M, Meilijson I (2004) Four notions of mean-preserving increase in risk, risk attitudes and applications to the rank-dependent expected utility model. J Math Econ 40(6):547–571

Chew S, Karni E, Safra Z (1987) Risk aversion in the theory of expected utility with rank dependent preferences. J Econ Theory 42:370–381

Chew S, Wakker PP (1996) The comonotonic sure thing principle. J Risk Uncertain 12:5–27

Conati C, Gertner AS, VanLehn K, Drudzel MJ (1997) On-line student modeling for coached problem solving using Bayesian networks. In: Proceedings of the international conference on user modeling (UM'97)

Dasgupta P (2006) Distributed automatic target recognition using multiagent UAV swarms. In: Proceedings of the international conference on autonomous agents and multiagent systems (AAMAS'06), pp 479–481

de Salvo Braz R, Amir E, Roth D (2005) Lifted first-order probabilistic inference. In Proceedings of the international joint conference on artificial intelligence (IJCAI'05), pp 1319–1325

Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–339

Doucet A, Johansen A (2011) The Oxford handbook of nonlinear filtering, chapter a tutorial on particle filtering and smoothing: fifteen years later. Oxford University Press, Oxford, pp 656–704

Dubois D, Fargier H, Perny P (2003) Qualitative decision theory with preference relations and comparative uncertainty: an axiomatic approach. Artif Intell J 148(1):219–260

Dubois D, Fargier H, Perny P, Prade H (2002) Qualitative decision theory: from Savage's axioms to nonmonotonic reasoning. Int J Assoc Comput Mach 49(4):455–495

Dubois D, Fargier H, Prade H (1997) Decision-making under ordinal preferences and uncertainty. In: Proceedings of the conference on Uncertainty in artificial intelligence (UAI'97), pp 157–164

Dubois D, Le Berre D, Prade H, Sabbadin R (1999) Using possibilistic logic for modeling qualitative decision: ATMS-based algorithms. Fund Inform 37(1–2):1–30

Dubois D, Prade H (1995) Possibility theory as a basis of qualitative decision theory. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'95), pp 1924–1930

Dubois D, Prade H, Sabbadin R (1998) Qualitative decision theory with Sugeno integrals. In: Proceedings of the conference on uncertainty in artificial intelligence (UAI'98), pp 121–128

Ellsberg D (1961) Risk, ambiguity and the Savage axioms. Q J Econ 75:643–669

Fargier H, Lang J, Sabbadin R (1998) Towards qualitative approaches to multistage decision making. Int J Approx Reason 19:441–471

Fishburn PC (1970) Utility theory for decision making. Wiley, New York

Fishburn PC (1982) The foundations of expected utility. Kluwer

Fishburn PC, Roberts FS (1978) Mixture axioms in linear and multilinear utility theories. Theory Decis 9:161–171

Franklin R, Spiegelhalter D, Macartney F, Bull K (1991) Evaluation of an algorithm for neonates. Br Med J 302:935–939

Getoor L, Taskar B (2007) Introduction to statistical relational learning. MIT Press, Cambridge

Gilboa I (1987) Expected utility with purely subjective non-additive probabilities. J Math Econ 16:65–88

Gilboa I (2009) Theory of decision under uncertainty. Econometric society monographs. Cambridge University Press, Cambridge

Gonzales C, Perny P (2004) GAI networks for utility elicitation. In: Proceedings of the international conference on principles of knowledge representation and reasoning (KR'04), pp 224–234

Herstein IN, Milnor J (1953) An axiomatic approach to measurable utility. Econometrica 21:291–297

Horvitz E, Barry M (1995) Display of information for time-critical decision making. In: Proceedings of the conference on uncertainty in artificial intelligence (UAI'95), pp 296–305

Horvitz E, Breese J, Heckerman D, Hovel D, Rommelse K (1998) The Lumière project: Bayesian user modeling for inferring the goals and needs of software users. In: Proceedings of the conference on uncertainty in artificial intelligence (UAI'98), pp 256–265

Howard RA (1960) Dynamic programming and Markov processes. MIT Press, Cambridge
Howard RA, Matheson JE (1984) Influence diagrams. In: Howard R, Matheson J (eds) Readings on the principles and applications of decision analysis, vol 2. Strategic Decision Group, Menlo Park, pp 719–762
Hurwicz L (1951) Optimality criteria for decision making under ignorance, vol 370. Cowles Commission discussion paper, Statistics
Ingersoll J (1987) Theory of financial decision making. Rowman and Littlefeld
Jaffray J-Y (1988) Choice under risk and the security factor: an axiomatic model. Theory Decis 24(2):169–200
Jaffray J-Y (1989) Linear utility theory for belief functions. Oper Res Lett 8:107–112
Jensen F, Jensen FV, Dittmer SL (1994) From influence diagrams to junction trees. In: Proceedings of the conference on uncertainty in artificial intelligence (UAI'94)
Jensen FV, Kjærulff U, Kristiansen B, Langseth H, Skaanning C, Vomlel J, Vomlelova M (2001) The SACSO methodology for troubleshooting complex systems
Jensen NE (1967) An introduction to Bernoullian utility theory. I: utility functions. Swed J Econ 69:163–183
Kahneman D, Tversky A (1972) Subjective probability: a judgment of representativeness. Cogn Psychol 3:430–454
Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–291
Keeney RL, Raiffa H (1993)Decisions with multiple objectives - preferences and value tradeoffs. Cambridge University Press, Cambridge. (Version originale en 1976 chez Wiley)
Khosravi H, Schulte O, Man T, Xu X, Bina B (2010) Structure learning for Markov logic networks with many descriptive attributes. In: Proceedings of the national conference on artificial intelligence (AAAI'10)
Knight F (1921) Risk, uncertainty and profit. Houghton Miffin
Kok S, Domingos P (2009) Learning Markov logic network structure via hypergraph lifting. In: Proceedings of the international conference on machine learning (ICML'09)
Kraft CH, Pratt JW, Seidenberg A (1959) Intuitive probability on finite sets. Ann Math Stat 30:408–419
Lehmann D (1996) Generalized qualitative probability: Savage revisited. In: Proceedings of the conference on uncertainty in artificial intelligence (UAI'96), pp 381–388
Lu T, Boutilier C (2011) Robust approximation and incremental elicitation in voting protocols. In Proceedings of the international joint conference on artificial intelligence (IJCAI'11), pp 287–293
Machina M (1982) Expected utility analysis without the independence axiom. Econometrica 50:277–323
Monahan GE (1982) A survey of partially observable Markov decision processes: theory, models and algorithms. Manag Sci 28:1–16
Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufman Publishers Inc
Perny P, Rolland A (2006) Reference-dependent qualitative models for decision making under uncertainty. In: Proceedings of the European conference on artificial intelligence (ECAI'06), pp 422–426
Pratt J (1964) Risk aversion in the small and in the large. Econometrica 32:122–136
Puterman ML (1994) Markov decision processes: discrete stochastic dynamic programming. Wiley, New York
Quiggin J (1982) A theory of anticipated utility. J Econ Behav Organ 3:323–343
Quiggin J (1992) Increasing risk: another definition. In: Chikan A (ed) Progress in decision, utility and risk theory. Kluwer, Dordrecht
Quiggin J (1993) Generalized expected utility theory: the rank-dependent model. Springer, Berlin
Raiffa H (1968) Decision analysis: introductory lectures on choices under uncertainty. Addison-Wesley, Reading

Ramsey FP (1931) Truth and probability. In: Ramsey F (ed) The foundations of mathematics and other logical essays. Harcourt, Brace and Co, California
Regan K, Boutilier C (2011) Robust online optimization of reward-uncertain MDPs. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'11), pp 2165–2171
Rotschild M, Stiglitz J (1970) Increasing risk I: a definition. J Econ Theory 2:225–243
Rotschild M, Stiglitz J (1971) Increasing risk II: its economic consequences. J Econ Theory 3:66–84
Sabbadin R (1998) Une Approche Ordinale de la Décision dans l'Incertain: Axiomatisation, Représentation Logique et Application à la Décision Séquentielle. Thèse de doctorat, Université Paul Sabatier, Toulouse, France
Sabbadin R (2001) Possibilistic Markov decision processes. Eng Appl Artif Intell 14:287–300
Savage LJ (1954) The foundations of statistics. Dover
Schmeidler D (1986) Integral representation without additivity. In: Proceedings of the American mathematical society (AMS), vol 97, pp 255–261
Shachter R (1986) Evaluating influence diagrams. Oper Res 34:871–882
Shafer G (1976) Mathematical theory of evidence. Princeton University Press, Princeton
Sondik E (1971) The optimal control of partially observable Markov processes. PhD thesis. Stanford University
Sordoni A, Briot J-P, Alvarez I, Vasconcelos E, Irving M, Melo G (2010) Design of a participatory decision making agent architecture based on argumentation and influence function: application to a serious game about biodiversity conservation. RAIRO Oper Res 44(4):269–284
Tan S, Pearl J (1994) Qualitative decision theory. In: Proceedings of the national conference on artificial intelligence (AAAI'94), pp 928–933
von Neumann J, Morgenstern O (1944) Theory of games and economic behaviour. Princetown University Press, Princetown
Wakker PP (1990) Under stochastic dominance Choquet expected utility and anticipated utility are identical. Theory Decis 29:119–132
Wakker PP (1994) Separating marginal utility and risk aversion. Theory Decis 36:1–44
Wald A (1950) Statistical decision functions. Wiley, New York
Wang T, Boutilier C (2003) Incremental utility elicitation with the minimax regret decision criterion. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'03), pp 309–316

Collective Decision Making

Sylvain Bouveret, Jérôme Lang and Michel Lemaître

Abstract This chapter introduces two prominent models of collective (multiagent) decision making, namely the basic ordinal model, and the utilitarian (numerical/quantitative) model. These models are then illustrated on three major collective decision making problems: voting, fair division and auctions. For each of these three problems we give a formal definition and we discuss the main links with computer science and artificial intelligence.

1 Introduction

1.1 Collective Decision Making Problems

This chapter focuses on collective decision making (CDM) problems, in which a group of people (agents) has to make a collective decision cooperatively. The chosen decision, to be selected among a set of eligible decisions, will engage each agent. Most procedures presented in this chapter are centralized procedures.

Typical CDM problems examples are: political elections; private everyday votes (for example, friends choosing a restaurant); fair allocation (for example, dividing goods in a divorce, allocating courses to students in a university...); a jury seeking for a consensus in a court.

The study of CDM problems dates back to Antiquity. The name "social choice" refers nowadays to the formal study of such collective problems. Nicolas de

S. Bouveret (✉)
Université Grenoble-Alpes, CNRS, LIG, 38000 Grenoble, France
e-mail: sylvain.bouveret@imag.fr

J. Lang
LAMSADE-CNRS, PSL Research University, Université Paris Dauphine, Paris, France
e-mail: lang@lamsade.dauphine.fr

M. Lemaître
Formerly ONERA, Toulouse, France
e-mail: michel.lemaitre.31@gmail.com

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_18

Condorcet (1743–1794) was one of the first to formalize some CDM problems. His contribution to the field of voting systems (Condorcet 1735) is widely recognized. Other prominent contributors are Kenneth Arrow (1921–2017), celebrated for his famous impossibility theorem (see p. 4), and Amartya Sen (born in 1933), well known for his work on social inequalities (Sen 1970).

Classical social choice theory has never been very much concerned about algorithmic issues. This is where computer science and more specifically artificial intelligence and operations research come into play: a recent research field has emerged, named *computational social choice*, bringing computer science and social choice together. Two research directions have appeared: the first one (from social choice theory to computer science) aims at exploiting social choice theory concepts and procedures in order to solve problems arising in computer science applications (for example, aggregation procedures for web page ranking and information retrieval, using voting for pattern recognition and classification, or computational resource allocation). The other direction (from computer science to social choice theory) aims at using notions and methods coming from computer science (representation languages, complexity, algorithmics, interaction protocols...) in order to solve complex group decision making problems. This last direction is by far the most important.

Formally, a CDM problem consists of three elements: a set of *agents* $\mathcal{N} = \{1, \ldots, n\}$; a set of *eligible decisions* or *alternatives* $\mathcal{X}$; an expression of individual preferences (or sometimes *beliefs* — we will go back to this later) of each agent over the alternatives. The expected result is, as the case may be, the choice of a "socially optimal" alternative, the choice of a set of alternatives, or a ranking of the alternatives.

Three of the most important sub-fields of social choice are:

- *voting*: agents (or *voters*) express their preferences over alternatives (in this case *candidates*) and must agree on the choice of a candidate (or a subset of candidates).
- *fair allocation*: a common resource has to be divided amongst agents expressing their preferences about the possible shares they can possibly receive.
- *judgement aggregation*: agents express their beliefs over the real word and must come up to a common conclusion.

The first two examples above concern *preference* aggregation (the most frequent case in social choice), whereas in the last case, *belief* aggregation is at stake. Belief aggregation is addressed specifically in chapter "Belief Revision, Belief Merging and Information Fusion" of this volume and will not be discussed in this one. In the following we are only concerned with preference aggregation, focusing successively on voting, fair allocation, and finally on combinatorial auctions, which are a special case of resource allocation.

There are many models dedicated to the problem of aggregating individual preferences into a collective one. We will now present the two prominent models on which most works on CDM are based.

1.2 The Basic Model: Ordinal Preferences

Let $\mathscr{P}$ be the set of total preorders[1] on $\mathscr{X}$. In the ordinal model, the preferences of an agent i are represented by the *individual preorder* $\succeq_i \in \mathscr{P}$.

Let $G : \mathscr{P}^n \to \mathscr{P}$ be a *collective preorder aggregation function*. The *collective preorder* $\succeq_{col} = G(\langle \succeq_1, \succeq_2, \ldots, \succeq_n \rangle)$ (or *social welfare ordering*) represents the collective preference which results from the aggregation by G of the individual preference profile $\langle \succeq_1, \succeq_2, \ldots, \succeq_n \rangle$. A collectively preferred alternative is an alternative maximizing the collective preorder $\succeq_{col}$.

Let us give a simple example of a preorder aggregation procedure: let $N(a)$ be the number of times for which an alternative a is (one of) the most preferred in the individual preorders. Now define $a \succeq_{col} b \equiv N(a) \geq N(b)$. This is obviously a preorder. On the other hand, the aggregation which prefers alternative a to alternative b when a majority of agents prefer a to b is not a preorder, because it may generate a cyclic collective preference, in which case the preference relation is not transitive — this is the celebrated Condorcet paradox, see Sect. 2.

In this context, the centralized CDM problem consists in defining an aggregation function G having "good" properties. What are these "good" properties for CDM? We now introduce the main ones.

1.2.1 The Pareto-Efficiency Property and the Unanimity Principle

Informally, an efficient alternative is an alternative which satisfies all agents "as well as possible". The simplest and mostly used expression of efficiency is the *Pareto-efficiency* property, based on *Pareto-dominance*. Let $\langle \succeq_1, \succeq_2, \ldots, \succeq_n \rangle$ be a preference profile. We say that alternative a Pareto-dominates alternative b when $a \succeq_i b$ for all agents, with $a \succ_i b$ for at least one agent ($\succ_i$ designates the strict part of $\succeq_i$, that is $a \succ_i b \equiv [a \succeq_i b$ and not $b \succeq_i a]$). A Pareto-efficient (or Pareto-optimal) alternative is a non-dominated one. It is such that we cannot switch to another alternative increasing strictly the satisfaction of an agent, without strictly decreasing the satisfaction of another agent. We say that an aggregation function G satisfies the Pareto-efficiency property if the alternatives collectively preferred are Pareto-efficient.

The *unanimity principle* simply requires that the aggregation function G satisfies the Pareto-efficiency property.

1.2.2 The Independence of Irrelevant Alternatives (IIA) Property

This natural property asks that for each pair of alternatives a and b, the strict collective preference between a and b ($a \succ_{col} b$ or $b \succ_{col} a$) only depends on the way each agent strictly compares a and b ($a \succ_i b$ or $b \succ_i a$) — the other alternatives are irrelevant.

[1] A preorder $\succeq$ is a binary relation that is reflexive and transitive. In a total preorder (or weak order), no pair of alternatives is incomparable: $x \succeq y$ or $y \succeq x, \forall x, y \in \mathscr{X}$.

1.2.3 Arrow's Theorem

Most results in classical social choice theory consist of *impossibility* or *possibility theorems* of the following form: *there is no collective decision procedure satisfying a set of natural and desirable conditions* $R_1, \ldots, R_p$, or *the set of collective decision procedures satisfying the set of natural and desirable conditions* $R_1, \ldots, R_p$ *is exactly the set of procedures of form* F. A celebrated example is Arrow's theorem (1951). Consider strict preference profiles $\langle \succ_1, \ldots, \succ_i, \ldots, \succ_n \rangle$ on $\mathscr{X}$ (total strict orders). Let $\mathscr{S}$ be the set of all possible strict profiles. Arrow's theorem states that if there are at least 3 alternatives, then any aggregation function G defined on $\mathscr{S}^n$ satisfying the unanimity principle and the IIA property is *dictatorial*, meaning that there is an agent i such that for any profile P, $G(P) = \succ_i$.[2]

1.3 The Utilitarian Model, or the Model of Quantitative Preferences

The utilitarian model (or numerical or quantitative preference model) represents the preferences of agent i by an *individual utility function* $u_i : \mathscr{X} \to \mathbb{R}$. To each alternative a corresponds a vector $\langle u_1(a), u_2(a), \ldots, u_n(a) \rangle$ (utilities of a for each agent) called the *utility profile* of a.

In order to compare the (quantitative) satisfaction of two agents for a given alternative, utilities must be defined on a common scale. But this is not always possible: *interpersonal comparison of utility* is a critical question in CDM. Actually, agents may use their own non commensurable utility scales. However in the following we will assume (unless explicitly stated) that agents' utilities are expressed on a common utility scale.[3]

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a *collective utility aggregation function*, and let $u : \mathscr{X} \to \mathbb{R}$ be the function defined this way: $u(a) = g(\langle u_1(a), u_2(a), \ldots, u_n(a) \rangle)$, for any alternative a. The function u, called *collective utility function* or *social utility function*, represents the collective preference obtained through aggregation by g of individual utility functions u_i. A collectively preferred alternative is an alternative maximizing this function u.

Given individual and collective utility functions, we can easily recover the ordinal model by defining individual and collective preorders on $\mathscr{X}$ as follows: for any agent i, $a \succeq_i b \equiv u_i(a) \geq u_i(b)$, and $a \succeq_{col} b \equiv u(a) \geq u(b)$. The individual utility function u_i (respectively the collective utility function u) is said to *represent* the individual preorder $\succeq_i$ (respectively the collective preorder $\succeq_{col}$). In this way, each

[2] Arrow's theorem also holds (in a weaker form) when the individual preferences are preorders (that is, with possible indifference between alternatives).

[3] A straightforward way to obtain a common utility scale, often used in fair allocation problems, is to normalize the individual utility of each agent relatively to the utility she would get if she was given all the resource (Kalai-Smorodinsky normalization).

purely ordinal property (such as the Pareto-efficiency property) can be expressed in the utilitarian model.

The main two collective utility aggregation functions are sum $u(a) = \sum_{i \in \mathcal{N}} u_i(a)$, and minimum: $u(a) = \min_{i \in \mathcal{N}} u_i(a)$. These two functions respectively correspond to the main two agendas of the utilitarian model, namely *classical utilitarianism* and *egalitarianism*. Classical utilitarianism (sum) seeks to produce collective utility, irrespective of the agent from which this utility comes from (hence ignoring any equity concern). On the other hand, egalitarianism (min) seeks to maximize and equalize at the same time individual utilities: it selects an alternative which maximizes the satisfaction of the least satisfied agent, hence conveying a very strong equity flavor.

Classical utilitarianism and egalitarianism are two opposite and extreme attitudes towards CDM.[4] In classical utilitarianism, an agent is a "collective utility producer". The marginal collective utility produced by an agent does not depend on her present degree of individual utility. Hence, the collective preference maximization could indeed lead to lower the satisfaction of the least satisfied agents, if more satisfied agents "produce" more utility. Agents must show a high degree of solidarity: some of them could be sacrificed on the altar of the collective utility maximization. Conversely, in egalitarianism, even a large utility increment of an agent already satisfied does not compensate for a tiny loss of utility of the least satisfied agent.

These two variations of utilitarianism are linked with two different approaches in philosophy and economics: Rawls (1971) and Sen (1970) for egalitarianism, and Harsanyi (1955) for classical utilitarianism are often advocated.

The utilitarian model often refers to an efficiency definition which is a refinement of Pareto-efficiency, namely *sum-efficiency*. A sum-efficient alternative is an alternative maximizing the sum of individual utilities, that is, alternatives maximizing the sum of individual utilities. A sum-efficient alternative is Pareto-efficient, but the converse is not true in general.

In the utilitarian model, maximizing the collective utility function yields Pareto-efficient (or Pareto-optimal) decisions if and only if the aggregation function g is strictly increasing. This is the case for sum, but not for min. The leximin total preorder is a refinement of the total preorder induced on $\mathcal{X}$ by min, of which the maximization always results in a Pareto-efficient alternative.

1.4 Centralized Versus Distributed CDM

Another essential dichotomy exists — orthogonal to the ordinal versus quantitative/numerical preferences one — namely the way agents interact in the decision making. In a *centralized* CDM resolution, a central authority (arbitrator, chairper-

[4]For example, egalitarianism prefers the utility profile $\langle 10, 10, 10 \rangle$ to the profile $\langle 9, 100, 100 \rangle$. Classical utilitarianism prefers $\langle 1, 100, 100 \rangle$ to $\langle 66, 67, 67 \rangle$, and even to $\langle 2, 99, 99 \rangle$. See Sect. 3 for two families of trade-offs between utilitarianism and egalitarianism.

son, Home Office, auctioneer, ...) gathers in a first phase the preferences of the agents (or at least a part of them, being informative enough for the decision to be made), and then decides on the optimal alternative and communicates it to the agents. The interaction phase between the central authority and the agents intended for collecting preferences is generally called the *elicitation* phase (the interested reader can see chapter "Compact Representation of Preferences" of this volume for more details). In fully *distributed* CDM resolution, there is no central authority, agents interact freely, and negotiate to reach a common consensus. There are also intermediate CDM frameworks, that lie between centralized resolution and distributed negotiation.

This chapter concerns mostly centralized CDM resolution, because of its importance, and because distributed decision (in particular negotiation) is considered in detail in chapter "Negotiation and Persuasion among Agents" of this volume.

1.5 Discussion

In practice, the nature of the CDM problem at stake dictates the choice of a particular model of preferences (ordinal or numerical/quantitative) and type of resolution (centralized or distributed). By way of illustration, voting theory generally assumes ordinal preferences and centralized resolution; fair allocation with money (as with auctions), numerical preferences and centralized resolution; some fair allocation problems — such as *cake-cutting*, see Sect. 3.5 — ordinal preferences and distributed resolution.

Some difficulties may impair the use of a centralized CDM model:

- agents might be unable to reveal their preferences or simply refuse to do so, hence complicating the elicitation phase;
- in the real world, the agents often have intricate preferences that are difficult to translate into preorders or utilities: agents are often sensitive to several criteria (see Chapter "Compact Representation of Preferences" of this volume), and are often not indifferent to other agents preferences, as well as to some social norms.

However, to be accepted by the agents at stake, the CDM model and resolution should be based on clear concepts, easy to explain and use.

Finally, we should be aware of the limitations of the standard CDM models presented above. Their interest can be mostly found for CDM problems having a technical aspect — typically, the routine allocation of numerous physical resources — for which direct negotiations are hardly possible, and for which a kind of automatic processing is required. These models can serve as well to build some technically relevant solutions, initiating a negotiation process.

The following sections are devoted to three specific CDM problems. The first concerns voting, making use of the ordinal model of preferences. The next one is dedicated to fair allocation problems, for which equity is a strong concern. We present in the last section the auction problem, a specific allocation problem in which agents

interact in a limited way, and which is solved by the mean of maximizing the profit of a particular agent (the auctioneer).

2 Voting

The use of voting procedures for collective decision making is not only of tremendous importance in large-scale political contexts, but it is also more and more applied in low-stake contexts such as social networks, workplaces or other local communities (and perhaps also societies of autonomous agents), which explains why it has acquired so much importance in the last fifteen years in the artificial intelligence literature. Since there are now hundreds of papers on voting in the mainstream AI conferences (such as AAAI or IJCAI) and journals (such as AIJ or JAIR) as well as in more specialised conferences (such as AAMAS), we cannot report on every research stream and we will only give a brief overview of the main topics (measured in number of papers). For a finer overview we advise to consult Chaps. 2–10 of the Handbook of Computational Social Choice (Brandt et al. 2016b).

2.1 Introduction to Voting Theory

A common assumption in voting theory is that the agents (which will be called 'voters' in this section) have ordinal preferences, and furthermore that these preferences are *linear orders* (or *rankings*) over the set of alternatives. (There are exceptions to this, such as in approval voting, which will be discussed further.)

Let $N = \{1, \ldots, n\}$ be a set of voters, and $\mathscr{X} = \{x_1, \ldots, x_m\}$ be a set of *alternatives*, or *candidates*. A *profile* is a collection of n votes, where each vote is a linear order over $\mathscr{X}$:

$$P = \langle V_1, \ldots, V_n \rangle = \langle \succ_1, \ldots, \succ_n \rangle$$

where V_i (also denoted $\succ_i$) is the vote expressed by voter i.

A *resolute voting rule* F is a function that maps each profile P to a candidate $F(P)$ in $\mathscr{X}$, who is the socially preferred candidate.

An *irresolute voting rule* F is a function that maps each profile P to a nonempty subset of $\mathscr{X}$: $F(P)$ is the set of socially preferred candidates, called *cowinners*; the candidate that will be chosen in the end will be one of the candidates of $F(P)$, obtained by means of a *tie-breaking mechanism* whose specification is outside the definition of F.[5]

[5]The reason why we sometimes need irresolute rules is the possibility of a tie: suppose for instance that we have two candidates a and b, $n = 2q$ voters, and a profile P containing q votes $a \succ b$ and q votes $b \succ a$. For an irresolute voting rule, we simply let $F(P) = \{a, b\}$, and the final winner will be chosen by the tie-breaking mechanism. For a resolute rule, however, we have to specify the tie-breaking mechanism as part of the rule. For this, a choice must be made: either we give up

When there are only two candidates a and b, the arguably most reasonable irresolute rule is the *majority rule*:

$$\text{maj}(V_1, \ldots, V_n) = \begin{cases} \{a\} & \text{if a strict majority of voters prefers } a \text{ to } b \\ \{b\} & \text{if a strict majority of voters prefers } b \text{ to } a \\ \{a, b\} & \text{otherwise} \end{cases}$$

May's theorem (1952) gives an axiomatic characterisation of the majority rule.

Things become more complicated when the number of candidates is at least 3. We now give an incomplete (but representative) list of voting rules. Unless stated otherwise, we define only their irresolute version; again, a resolute version can be obtained by composition with a tie-breaking mechanism.

A *positional scoring rule* is defined by a vector $\mathbf{s} = \langle s_1, \ldots, s_m \rangle$ of m integers, with $s_1 \geq \cdots \geq s_m$ and $s_1 > s_m$: each time voter i ranks candidate x in position j, x gets a score $score_i(x) = s_j$; the cowinners for the scoring rule $F_{\mathbf{s}}$ are the candidates maximizing $s(x) = \sum_{i=1}^{n} score_i(x)$. Here are three important examples of positional scoring rules:

- *plurality*: $s_1 = 1, s_2 = \cdots = s_m = 0$ (the cowinners are the candidates ranked first most often);
- *veto* (or *antiplurality*): $s_1 = s_2 = \cdots = s_{m-1} = 1$, $s_m = 0$ (the cowinners are the candidates ranked last least often);
- *Borda*: $s_1 = m - 1, s_2 = m - 2, \ldots s_m = 0$.

Consider the profile P composed of one vote $c \succ a \succ b \succ d$, two votes $a \succ b \succ d \succ c$ and two votes $d \succ b \succ c \succ a$: the cowinners for plurality are a and d; for Borda and veto, it is b.

Another important family of voting rules is that of the rules *based on the majority graph*. Given two candidates x and y, and a profile P, let $N_P(x, y)$ be the number of voters who prefer x to y in P. The *majority graph* M_P associated with P is the directed graph whose vertices are the candidates, and which contains an edge from x to y if and only if $N_P(x, y) > \frac{n}{2}$. A voting rule is *based on the majority graph* if the cowinners can be computed from M_P.

A candidate x is *Condorcet winner* for P if for any $y \neq x$, we have $N_P(x, y) > \frac{n}{2}$, that is, if it beats every other candidate in a pairwise duel by a majority of votes. Clearly, x is Condorcet winner for P if M_P contains an edge from x to every other candidate. Of course, when there exists a Condorcet winner, it is unique. However, for some profiles, there is no Condorcet winner (see the example below). A voting rule is *Condorcet-consistent* if it elects the Condorcet winner when there exists one.

The majority graph M_P associated with the previous profile P is given in Fig. 1. Each candidate being dominated by another candidate, there is no Condorcet winner for P. If the first voter, instead of voting $c \succ a \succ b \succ d$, had voted $c \succ b \succ a \succ d$, then the edge $a \to b$ would have been replaced by an edge $b \to a$ and b would have been a Condorcet winner.

---

neutrality and use a predefined priority relation over candidates; or we give up *anonymity*, and use a predefined priority relation over voters or sets of voters.

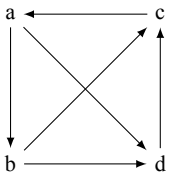


Fig. 1 Majority graph M_P

Here are three examples of rules based on the majority graph:

- *Copeland*: the *Copeland score* of a candidate x with respect to a profile P is the number of candidates that x beats in the majority graph M_P, plus half the number of candidates for which there is a pairwise tie with x (there is a pairwise tie between x and y if M_P contains neither an edge from x to y nor one from y to x). The Copeland (co)winners are the candidates with largest Copeland score. For example, for the profile P above, the Copeland winners are a and b.
- *Slater*: a *Slater order* for P is a linear order on $\mathscr{X}$ minimizing the number of edges disagreeing with M_P. A Slater winner is a candidate ranked first in some Slater order. For the profile P above, the unique Slater order is $a \succ b \succ d \succ c$, with only one disagreement with M_P (about (a, c)), and the Slater winner is a.
- *Banks*: for $S \subseteq \mathscr{X}$, let $M_P^{\downarrow S}$ be the restriction of M_P to S. $M_P^{\downarrow S}$ is a *maximal acyclic subtournament* of M_P if $M_P^{\downarrow S}$ is acyclic and for each S' such that $S \subset S' \subseteq \mathscr{X}$, $M_P^{\downarrow S'}$ is not acyclic. Then x is a Banks winner if x is non-dominated in a maximal acyclic subtournament of M_P. For the profile P above, the maximal acyclic subtournaments of M_P are obtained for $\{a, b, d\}$, $\{a, c\}$ and $\{b, c, d\}$, and the Banks winners are a, b and c.

Clearly, these three rules are Condorcet-consistent. For a survey of rules based on the majority graph, with a focus on computation, see Brandt et al. (2016a).

The *weighted majority graph*, or *pairwise comparison matrix* W_P is defined by: for each $x, y \in \mathscr{X}$, $x \neq y$, $W_P(x, y)$ is the number of voters who prefer x to y minus the number of voters who prefer y to x. A voting rule is *based on the weighted majority graph* if the (co)winners can be computed from the weighted majority graph (or pairwise comparison matrix) W_P.

Consider the profile Q:

$$\begin{aligned} &4 \text{ voters} : a \succ b \succ c \succ d \\ &2 \text{ voters} : b \succ c \succ d \succ a \\ &3 \text{ voters} : c \succ d \succ a \succ b \end{aligned}$$

The weighted majority graph for Q is

	a	b	c	d
a	–	5	−1	−1
b	−5	–	3	3
c	1	−3	–	9
d	1	−3	−9	–

Here is a rule based on the weighted majority graph: the *Simpson* (or *maximin*) rule outputs the candidates maximizing $\min_{y \neq x} W_P(x, y)$. The maximin winner for Q is a, with $\min_{y \neq a} W_Q(a, y) = -1$. Clearly, $\min_{y \neq x} W_P(x, y) > 0$ if and only if x is a Condorcet winner for P, and there cannot be two candidates with this property, therefore, the maximin rule is Condorcet-consistent.

Another rule based on the weighted majority graph is the *Kemeny* rule, defined as follows: the *Kemeny score* $K(V, P)$ of a linear order V with respect to profile P is defined by $K(V, P) = \sum_{(x,y) \in \mathscr{X}, x \neq y} W_P(x, y)$. A *Kemeny consensus* for P is a linear order V^* maximizing $K(V^*, P)$, and a *Kemeny winner* is a candidate ranked first in a Kemeny consensus. The Kemeny rule is Condorcet-consistent as well. The Kemeny rule can be used a voting rule, but perhaps even more so as a social welfare function (outputting the set of Kemeny consensuses). Also, it is easily adaptable to truncated votes, which explains why it is used for the aggregation of rankings of web pages given by different search engines (Dwork et al. 2001). On profile Q, the Kemeny consensus is $a \succ b \succ c \succ d$, with Kemeny score 18.

Some Condorcet-consistent rules are not based on the weighted (and a fortiori, unweighted) majority graph. Here is an example: the Dodgson rule[6] is defined as follows: for each $x \in \mathscr{X}$, $D(x)$ is the smallest number of elementary changes needed for making x a Condorcet winner, where an elementary change consists in swapping two adjacent candidates in a vote.

In order c to become a Condorcet winner for Q, it has to move one position up in two out of the first 6 votes; as for a, it needs to move two positions up in one of the last 5 votes; b and d need respectively 3 and 7 elementary changes in order to become Condorcet winners: therefore, a and c are the Dodgson cowinners for Q.

Here are now two rules that proceed by *successive rounds*. First, *single transferable vote* (STV) proceeds in $n - 1$ rounds, as follows:

1. let y be the candidate ranked first by the smallest number of voters (using a tie-breaking mechanism if necessary);[7]
2. eliminate y; the votes where y was ranked first are 'transferred' to the voter's preferred candidate among those who remain;
3. iterate the process until there remains only one candidate.

Consider the profile R containing 3 votes $a \succ d \succ b \succ c$, 4 votes $b \succ d \succ a \succ c$, 3 votes $c \succ d \succ a \succ b$ and 2 votes $d \succ c \succ b \succ a$. At the first round, d is eliminated;

[6] Charles Dodgson was better known under the name of Lewis Carroll.

[7] Another way of handling ties consists in considering all tie-breaking possibilities and gather the corresponding winning candidates; the resulting rule is called the *parallel universe* version of STV (Conitzer et al. 2009).

the votes of the two voters who preferred d are transferred to their second choice, that is, c. At the second round, we have the reduced following profile: 3 votes $a \succ b \succ c$, 4 votes $b \succ a \succ c$, 3 votes $c \succ a \succ b$ and 2 votes $c \succ b \succ a$: a is eliminated. At the last round, only b and c remain; 7 voters out of 12 prefer b to c, and the winner is b.

When there are only three candidates, STV coincides with *plurality with runoff*, which is defined more generally as follows: the first round selects the two candidates with the largest plurality scores (again, using tie-breaking if necessary), and the winner of the second round is selected according to majority.[8]

Social choice theorists have studied some desirable properties of voting rules. *Condorcet-consistency* is one of them; note that no positional scoring rule is Condorcet-consistent (Moulin 1988), and that STV and plurality with runoff are not Condorcet-consistent either. We give three other important properties, which for the sake of brevity we define for resolute rules only:

- *monotonicity*: when x is the winner for profile P, it remains the winner for a profile obtained from P by moving x up in some vote, the rest being unchanged;
- *participation*: when x is the winner for P, the winner for a profile obtained from P by adding one more vote is either x, or a candidate which the new voter prefers to x;
- *reinforcement*: when x is elected separately by two profiles, it is also elected by their union.
- *clone-proofness*: if a candidate x is cloned into a set of clones $\{x^1, \ldots, x^p\}$, and assuming that these clones of x will be ranked contiguously (in an arbitrary order) in each vote, and that the rest of the vote is equal to the vote before x was cloned, then the winner after cloning x will be (a) the same winner as before cloning x, if this winner was not x, and (b) one of the clones of x, if the winner was x.

For instance, positional scoring rules satisfy monotonicity, participation and reinforcement, but not clone-proofness; Copeland and maximin satisfy monotonicity, but not participation, reinforcement, nor clone-proofness; more generally, as soon as there are at least 4 candidates, Condorcet-consistency is incompatible with participation and with reinforcement. STV fails monotonicity, participation, reinforcement and Condorcet-consistency, but satisfies clone-proofness. Plurality with runoff fails to satisfy all these properties! For a survey on voting rules and their properties, see Brams and Fishburn (2004) and Zwicker (2016).

In *approval voting* (see Laslier and Sanver 2010 for an extensive survey), the input is different: each voter specifies an *approval ballot*, which is subset of candidates she approves; the cowinners are the candidates approved by the largest number of voters. After adapting the properties we listed above to approval ballots, we obtain that monotonicity, participation, reinforcement and clone-proofness are satisfied by approval voting, and that Condorcet-consistency is not.

[8] Plurality with runoff is used for political elections in many countries, such as France. STV – arguably better than plurality with runoff – is used for political elections in some countries such as Ireland and Australia.

2.2 Computing Voting Rules

Many voting rules are computable in polynomial time. This is the case for positional scoring rules and plurality with runoff, that are computable in time $O(nm)$, and for Copeland, Simpson, and STV,[9] computable in time $O(nm^2)$.

But for some other rules, winner determination is hard. The first article that shows that a voting rule is computationally hard is Bartholdi et al. (1989b), which shows that Dodgson and Kemeny rules are NP-hard. The exact complexity of the problem was determined by Hemaspaandra, Hemaspaandra and Rothe (1997): deciding whether x is a Dodgson winner is Θ_2^{P}-complete, that is, needs a logarithmic number of calls to NP-oracles.

The Kemeny, Slater and Banks rules are also hard to compute. Deciding if x is a Kemeny winner is Θ_2^{P}-complete (Rothe et al. 2003).[10] Deciding if x is a Banks winner is NP-complete (Woeginger 2003); however, it is possible to find an arbitrary Banks winner in polynomial time by a greedy algorithm (Hudry 2004b). Note that when comparing Banks to Kemeny, an important difference is that for Banks, since winner determination is "only" in NP, we can always find a succinct certificate for verifying that x is a winner (such a certificate is the subset S such that $M_P^{\downarrow S}$ is maximal acyclic), while for Kemeny, certificates are exponentially large (unless the polynomial hierarchy collapses). Finally, winner determination for the Slater rule is NP-hard, even under the restriction that ties between candidates do not occur (Ailon et al. 2005; Alon 2006; Conitzer 2006); but it is not known whether the problem is in NP or not.

These hardness results do not mean that we should give up using these rules, especially when they have good properties. Here are three ways of dealing with hardness.

First, *practical computation*: sometimes using a translation into a well-known setting with good solvers, such as integer linear programming; and sometimes using specific heuristics.

Second, *approximation*: interestingly, a polynomial approximation algorithm of a voting rule defines a new voting rule — sometimes already known under an other name, sometimes not. For example, let us consider the *Tideman rule*, defined as follows: if x, y are two candidates, let $Deficit(x, y) = \max(0, 1 + \lfloor \frac{N(y,x)-N(x,y)}{2} \rfloor)$ ($Deficit(x, y)$ is the number of votes which x needs in order to win against y, if that is possible) and the *Tideman score* is defined by $T(x) = \sum_{y \neq x} Deficit(x, y)$. The Tideman winner is the candidate minimizing the Tideman score. This rule is computable in $O(nm^2)$, and is a good approximation of the Dodgson rule, in the following sense: under the assumption that the profiles are uniformly distributed (also called *impartial culture assumption*), the probability that a Tideman winner is a Dodgson winner converges asymptotically towards 1 when the number of voters tends to infinity (McCabe-Dansted et al. 2008). Also, sometimes it is possible to design an approximation of a voting rule that not only is easier to compute than

[9] For its version where ties are broken as soon as they appear.

[10] For the more general problem of the computation of median orders, see Hudry (2004a).

the original rule, but also satisfies more desirable properties! For instance, while the Dodgson rule does not satisfy monotonicity, monotonic polynomial approximation of it have been designed by Caragiannis et al. (2009).

Third, *fixed-parameter tractable algorithms*: sometimes a rule is hard but becomes polynomial-time computable when the number of candidates is fixed; this is the case for instance for the Kemeny rule, for which winner determination can be made by inspecting each of the $m!$ orders and computing their scores in polynomial time.

These following three handbook chapters review the complexity, the approximation, and the practical computation of these rules that are hard to compute: Brandt et al. (2016a) for rules based on the majority graph, such as Banks or Slater; Fischer et al. (2016) for rules based on the weighted majority graph, such as Kemeny; and Caragiannis et al. (2016a) for other rules, such as Dodgson.

2.3 Voting on Combinatorial Domains

In many contexts, a decision has to be taken over several variables that may be intercorrelated. Two typical examples:

- *multiple referenda*: variables correspond to binary issues. For example, the inhabitants of a town may have to decide whether the town should build a swimming pool or not, and whether it should build a tennis court or not.
- *committee elections*: for example, a president, a vice-president and a secretary have to be elected, and some candidates (not necessarily the same ones) run for these positions. Sometimes there are no specific positions and the aim is just to elect k people.

In these situations, the space of candidates is a *combinatorial domain*: it consists in a Cartesian product $\mathscr{X} = D_1 \times \cdots \times D_m$, where D_i is a finite domain of values for variable X_i.

When the preferences of a voter on the values of a variable do not depend on the values of other variables, there are said to be *separable*. When all the voters have separable preferences, the vote can be decomposed into several independent voting processes, each bearing on a variable: for instance, there will be a vote about the swimming pool, and independently, a vote about the tennis court. Problems arise when some voters have nonseparable preferences. Consider the following example: there are two binary variables P (build a swimming pool), T (build a tennis court), and five voters whose preferences are

$$\begin{array}{ll} \text{voters 1 and 2}: & P\overline{T} \succ \overline{P}T \succ \overline{P\,\overline{T}} \succ P\overline{T} \\ \text{voters 3 and 4}: & \overline{P}T \succ P\overline{T} \succ \overline{P\,\overline{T}} \succ P\overline{T} \\ \text{voter 5}: & P\overline{T} \succ P\overline{T} \succ \overline{P}T \succ \overline{P\,\overline{T}} \end{array}$$

A first problem is concerned with the way the voters can express their preferences on $\{S, \overline{S}\}$ and on $\{T, \overline{T}\}$. This is not a problem for voter 5, whose preferences are

separable. On the other hand, for voters 1–4, this is problematic. Take for example voter 2. If she votes for the swimming pool, she can favour, according to the votes of other voters, $S\overline{T}$ (her best candidate) or ST (her worst candidate); and if she votes against the swimming pool, she can favour one of her two intermediate candidates. In both cases, she can feel regret once the final outcome of the vote is known. A second problem is that the outcome of the vote can be extremely bad. If the voters majoritarily vote 'optimistically', the outcome will be ST, which is the worst alternative for all voters but one. Such paradoxes have been studied under the name *multiple election paradoxes* (Brams et al. 1998; Lacy and Niou 2000).

When there is no guarantee that voters have separable preferences, the decomposition into independent voting processes is thus a bad idea, and other solutions must be found. There is no perfect solution; some possibilities:

1. ask voters to specify their preference relation *explicitly* on the set of all alternatives.
2. restrict the possible combinations of values for which one can vote.
3. ask voters to report a small part of their preference relation (e.g., their top alternative), and apply a voting rule that needs only this information (e.g., plurality).
4. ask voters to report their preferred alternative(s) and complete their preferences using a *distance* between alternatives.
5. use a *compact preference representation language* in which the voters' preferences will be represented succinctly.
6. *sequential voting*: vote about the variables one ofter the other, and communicate the outcome for a variable to the agents before they vote on the next variable.

One has to keep in mind that there are $\Pi_{1 \leq i \leq m} |D_i|$ alternatives. Therefore, as soon as there are more than three or four variables, Solution 1 is unrealistic.

Solution 2 is somewhat arbitrary: who decides which combinations are allowed? Moreover, in order this method to be realistic, the number of possible combinations has to be limited to a small number: voters thus express their preferences on a very small part of the set of alternatives.

Solution 3 is likely to give completely insignificant results as soon as the number of variables is significantly larger than the number of voters ($2^m \gg n$). For example, consider 5 voters and 6 binary variables, that is, 2^6 candidates, and choose plurality as the voting rule; one can expect the votes to be completely scattered, for example 001010: 1 vote; 010111: 1 vote; 011000: 1 vote; 101001: 1 vote; 111000: 1 vote; all other candidates: 0 vote. This solution is then completely pointless.

Solution 4 presupposes the existence of a natural and objective (voter-independent) distance between alternatives. It is used, among others, for defining the *minimax* committee election rule (Brams et al. 2007), and other rules, as well as in belief merging (see chapter "Main Issues in Belief Revision, Belief Merging and Information Fusion" of this volume). This solution is communicationwise cheap; it is however more costly in terms of computation, and it requires a significant domain restriction.

Solution 5 comes down to aggregate preferences specified in a compact representation language (see chapter "Compact Representation of Preferences" of this volume), such as CP-nets. It is potentially highly costly in terms of computation.

Finally, Solution 6 is an interesting trade-off: it is relatively cheap in communication and computation, and it is applicable to nonseparable preferences. However, in order it to work well, the following domain restriction has to be made (Lang and Xia 2009): there must exist a linear order on the variables $X_1 > \cdots > X_p$, common to all voters, such that the preferences of each voter on X_i are independent of the values of $X_{i+1}, \ldots, X_p$: for example, for the choice of a collective menu, `MainDish` > `FirstCourse` > `Wine` looks reasonable enough.

More details on voting over combinatorial domains can be found in Lang and Xia (2016).

2.4 Computational Barriers to Strategic Behaviour

A key problem in voting theory is that in some circumstances, some voters have an incentive to report insincere preferences in order to give more chances of winning to a candidate they prefer to the one who would be elected normally. Such a behaviour is called *manipulation*.

Consider for example plurality with runoff applied to the following profile: 8 votes $a \succ b \succ c$, 4 votes $c \succ b \succ a$ and 5 votes $b \succ a \succ c$. At the first round, c is eliminated, and at the second round, b is elected. Suppose now that 2 of the 8 first voters (those whose preference is $a \succ b \succ c$) decide to vote $c \succ b \succ a$ (all other votes being unchanged). The new profile is then composed of 2 votes $c \succ a \succ b$, 6 votes $a \succ b \succ c$, 4 votes $c \succ b \succ a$ and 5 votes $b \succ a \succ c$. At the first round, b is eliminated, and at the second round, a is elected. Since the actual preferences of these two voters are $a \succ b \succ c$, they are better off, since a is now the winner.

This example is not an isolated case. Indeed, the Gibbard–Satterthwaite theorem (Gibbard 1973; Satterthwaite 1975) shows that when there are at least three candidates, any voting rule which is nondictatorial and surjective (that is, for each candidate x, there is a profile for which x wins) is manipulable: for some profiles, some voters will have an incentive to report insincere preferences.

Although it is not possible to find a reasonable rule which is not manipulable, a way of limiting the impact of manipulation consists in making sure that a manipulation, whenever there is one, *is hard to compute*; this has lead computer scientists to study the *computational resistance to manipulation*. In practice, one considers that for a given voting rule, if finding a manipulation is NP-hard, then one can assume that voters – whose rationality is limited – will give up the idea of looking for one. Let us state the problem more formally by defining the following problem called COALITIONAL CONSTRUCTIVE MANIPULATION: for a voting rule F, given a distinguished candidate $x \in \mathscr{X}$, and the votes $\succ_1, \ldots, \succ_k$ of voters $1, \ldots, k$, is there a vote $\succ_i$ for each of the voters $i = k+1, \ldots, n$ such that x is elected by application of F on the profile $\langle \succ_1, \ldots, \succ_k, \succ_{k+1}, \ldots, \succ_n \rangle$?

The first articles on this topic have been written by Bartholdi et al. (1989a) and Bartholdi and Orlin (1991). Then this question came back in the early 2000s, with Conitzer and Sandholm (2002a). Since then, more than thirty papers on the problem

of complexity of (several variants of) manipulation have been written. They are surveyed by Conitzer and Walsh (2016). (See next section for another interpretation.)

Let us start by an example illustrating the *constructive manipulation of the Borda rule by a single voter*. Consider the following profile: $P = \langle a \succ b \succ d \succ c \succ e,\ b \succ a \succ e \succ d \succ c,\ c \succ e \succ a \succ b \succ d,\ d \succ c \succ b \succ a \succ e \rangle$. The current Borda scores (from these 4 votes) are a: 10, b: 10, c: 8, d: 7 and e: 5. Obviously, the last voter can make a or b win. Can she make c win? Yes, by ranking c first, then ranking in second position the least threatening candidate (e), then the least threatening after e (d), then a, then b (or vice versa). The final scores are then a: 11; b: 10; c: 12; d: 9; e: 8. Can she make d win? The same algorithm leads to rank d first, then e, then c, then, without loss of generaltiy, a, then b. The final scores are then a: 11; b: 10; c: 10; d: 11; e: 8: the existence of a constructive manipulation for d here depends on the tie-breaking priority order (there exists a constructive manipulation for d if and only if d has a higher priority than a or than b). On the other hand, there exists no constructive manipulation for e. The greedy algorithm we have applied (rank first the candidate that we want to be the winner, then the others by increasing order of their current Borda score, possibly taking tie-breaking priority into account) gives a successful manipulation if and only if there is one: the manipulation of the Borda rule by a single voter is therefore polynomial.

What about the same problem for two voters or more? Consider a profile for which the current Borda scores are a: 12; b: 10; c: 9; d: 9; e: 4; f: 1, with tie-breaking priority $a > b > c > d > e > f$. Generalizing the previous greedy algorithm does not work: suppose that the last two voters want e to win; after they rank it first, e has 14 points, and after they both rank f second, f has 9 points. They can continue with ranking d third for one of them and fifth for the other one (d then has 13 points). Then there are two ways of going further, depending on whether c is ranked once third and once fifth, or twice fourth; one can check that in the first case, it will not be possible to make e win, but that in the second case it will. This example suggests that computing a manipulation of the Borda rule by two voters or more is hard; its NP-hardness was long conjectured, and was proven independently by Betzler et al. (2011) and Davies et al. (2011).

Such complexity studies were done for numerous voting rules, in several contexts (constructive or destructive manipulation, by a single voter or a coalition of voters, by weighted or unweighted voters, with a restriction to single-peakedness profiles or not, etc.). We refer to Conitzer and Walsh (2016) for detailed results.

Some other works have also considered the issue of the *average* complexity of manipulation, starting from the constatation that an NP-hardness result talks about the worst case and does not guarantee that computing a manipulation will be *usually* hard. The results in this direction tend to show that there does not exist any rule that is *often enough* hard to manipulate (Procaccia and Rosenschein 2007).

Beyond manipulation by coalitions of voters, there exist other types of strategic behaviour, such as "procedural control": some voting rules can be strategically controlled by the central authority ('chair') in charge of the election. The first article on this topic (Bartholdi et al. 1992) defines several types of control: by addition, suppression or partitioning of candidates or voters. For example, for control by addition

of candidates, the chair can add a certain number of candidates in the hope of diluting the support to the candidates that can beat her favorite candidate. For each type of control and each voting rule F, three possibilities exist:

- F is *insensitive to control*: the chair can never act so as to change the winner.
- F is *resistent to control*: F is not insensitive to control but the control problem is computationally hard.
- F is *vulnerable to control*: F is not insensible to control and the control problem is computationally easy.

For example, the plurality rule is computationally resistant to control by addition or suppression of candidates, but computationally vulnerable to control by suppression of voters (Bartholdi et al. 1992).

Other types of strategic behaviour, related to control, have been considered more recently, such as *bribery*, *control of sequential voting on a combinatorial domain* or *manipulation by cloning candidates*.

For a synthesis on computational barriers to strategic behaviour, see the work by Conitzer and Walsh (2016) for manipulation and by Faliszewski and Rothe (2016) for control and bribery.

2.5 Incomplete Knowledge and Communication

We consider here questions such as: given an *incomplete* description of the votes, is the outcome already determined? If not, what are the candidates who can still win and what are the relevant pieces of information to ask to the voters? How can we do that in order to minimize the amount of communication exchanged between the voters and the central authority?

For example, let us consider the following *partial profile*, with 4 candidates and 9 votes, out of which only 8 have been expressed:

$$\begin{array}{l} 4 \text{ voters} : c \succ d \succ a \succ b \\ 2 \text{ voters} : a \succ b \succ d \succ c \\ 2 \text{ voters} : b \succ a \succ c \succ d \\ 1 \text{ voter} : \ ? \succ ? \succ ? \succ ? \end{array}$$

If the voting rule is plurality, then it is not difficult to see that the outcome is already determined independently of the last vote (the winner is c), while for Borda, the partial scores (computed from the 8 votes expressed) are a: 14; b: 10 ; c: 14; d: 10; thus, only a and c can win, and in order to determine the winner one needs only to know whether the last voter prefers a to c or vice versa. This problem is known under the name *vote elicitation* (Conitzer and Sandholm 2002b; Walsh 2008).

More generally, in order to model situations where the central authority has an *incomplete knowledge* of the voters' preferences, one considers that each voter has expressed a *partial order* over candidates, and a *partial profile* is a n-uple

of partial orders $P = \langle P_1, \ldots, P_n \rangle$. A *completion* of P is a (complete) profile $T = \langle T_1, \ldots, T_n \rangle$, where each T_i is a linear order extending P_i. Then one defines the *possible and necessary winners* (Konczak and Lang 2005) with respect to a voting rule and a partial profile:

- c is a *possible winner* if c is a winner for some completion of P.
- c is a *necessary winner* if c is a winner for each completion of P.

Thus, in the above example, c is a necessary winner for plurality; for Borda, the possible winners are a and c, and there is no necessary winner.

The computation of possible and necessary winners has received a significant attention, starting by Xia and Conitzer (2008). There also exists a probabilistic version, where one counts the extensions where a candidate wins (Bachrach et al. 2010).

Several classes of situations deserve a special attention:

1. *Possible and necessary winners for the addition of voters*: some voters have expressed their votes entirely, whereas the others have not expressed anything: the partial profile is $P = \langle P_1, \ldots, P_{n-k} \rangle$, where P_i is a linear order on $\mathscr{X}$. This class of situations corresponds, with a different interpretation, to a coalitional manipulation problem: more precisely, let us consider the coalition A composed of the last k voters. Then x is a possible winner if the coalition A can make x win (or equivalently, A has a constructive manipulation for x), whereas x is a possible winner if A cannot prevent x from winning (or equivalently, A has no destructive manipulation against x).
2. *Possible and necessary winners for the addition of candidates*: the voters have expressed their preferences on a fixed subset of candidates, and nothing on the other candidates: the partial profile is $P = \langle P_1, \ldots, P_n \rangle$, where P_i is a linear order on $\{x_1, \ldots, x_{m-k}\} \subseteq \mathscr{X}$. This class of situations occurs when new candidates declare in the curse of the process: one can for instance think of a Doodle poll for finding the date of a meeting, where voters have expressed their preferences on a first set of time slots, and when new time slots that were previously not considered can become possible after this first vote; or else, consider a hiring committee where a preliminary vote occurs between the candidates already interviewed and a new candidate is declared admissible (Chevaleyre et al. 2010). As an example, consider 12 voters, an initial set of candidates $X = \{a, b, c\}$ and a new candidate y. If the voting rule is plurality with tie-breaking priority order $a > b > c > y$, the partial profile is such that the plurality scores before y is taken into account are $a : 5$, $b : 4$, $c : 3$. On can check that a and b are possible winners, but not c. For instance, for b, it is enough that 2 of the voters who ranked a first now rank y first: the new plurality scores are $a : 3$, $b : 4$, $c : 3$, $y : 2$, and the winner is b.
3. *Truncated ballots*: every voter has expressed a partial ranking consisting of her top k candidates.
4. *Incomplete lists*: every voter i has expressed a ranking of an arbitrary subset $S_i \subseteq \mathscr{X}$ of alternatives (the candidates in $X \setminus X_i$ being incomparable with those in S_i, and incomparable with each other). This class of situations occurs when voters have an informed opinion on a subset of alternatives only: for instance, on

a web application for evaluating restaurants, a voter can evaluate only those she has tried.

A problem closely related to the search of possible winners for the addition of candidates is that of manipulation by candidate cloning (Elkind et al. 2010); the difference is that for candidate cloning, although we don't know how the clones of a candidate will be ranked by a voter in the profile after cloning, still, we know that they will be ranked contiguously in each vote.

Another topic is the design of *communication protocols* for voting rules. The definition of a voting rule does not say anything on the way the winner will be determined by the central authority. On the other hand, a *communication protocol* for a voting rule specifies the pieces of information that each voter will communicate in each round, in such a way that at the end of the protocol, the result will be known. (More generally, a protocol can be seen as an algorithm where atomic instructions are replaced by *communication actions* between agents, in such a way that an agent, in a given round, communicates information based on *her knowledge*.) The cost of a protocol is the total number of bits exchanged in the worst case. The *(deterministic) communication complexity* of a voting rule F is the cost of the cheapest protocol for F: it measures the minimal amount of information to be communicated so that the result of the vote can be determined. The communication complexity of voting rules is studied in detail by Conitzer and Sandholm (2005).

A trivial protocol for an arbitrary voting rule F is the following: each voter i sends her vote V_i to the central autority (this requires $n \log m!$ bits), and then the central authority sends back the name of the winner to all the voters (this requires $n \log m$ bits). The communication complexity of a voting rule is thus in $O(n \log m!)$. However, for some voting rules there exist cheaper protocols. This is obvious for instance for plurality, since it suffices for each voter to send the name of their preferred candidate to the central authority: the communication complexity of plurality is therefore at most in the order of $n \log m$ (it is in fact *exactly* of this order; the proof of the lower bound is nontrivial Conitzer and Sandholm 2005); but it is also the case for many other voting rules, such as plurality with runoff (in the order of $n \log m$), STV (in the order of $n(\log m)^2$), etc.

Another related problem is the *compilation of the votes of a subelectorate*. In a context where the votes do not come in a single round (consider for instance a political election where the votes of the citizens living abroad come with a few days delay, or to a Doodle pool where some persons are late in responding). In this case, it makes sense to compile the votes expressed so far, using as little space as possible, so as to "prepare the ground" for the time where the remaining votes are known. The *compilation complexity* of a voting rule is the minimal size for compiling a profile. It is identified, for some rules, by Chevaleyre et al. (2009) and Xia and Conitzer (2010).

For more details about voting with incomplete preferences, communication protocols, vote compilation, as well as the learning of (some classes of) voting rules and the robustness of voting rules, see the chapter by Boutilier and Rosenschein (2016).

2.6 Some Other Issues

For the sake of brevity, there are a number of other research topics at the meeting point of voting and AI which we have not discussed in this section.

One which is especially relevant to this book is *group planning*. It is addressed for the fist time by Ephrati and Rosenschein (1993): each agent has her own goal; at each round, agents vote on the next action to perform without revealing their preferences explicitly. More generally, *collective combinatorial optimization* deals with the design of methods for the collective version of specific combinatorial optimization problems, such as shortest path finding (Klamler and Pferschy 2007), minimum spanning tree (Darmann et al. 2009); egalitarian versions of some other combinatorial optimization problems are studied by Galand and Perny (2006), Escoffier et al. (2013).

A few other topics, such as: *randomized voting*, *iterative voting*, *computer-assisted theorem proving in social choice*, *approximate notions of single-peakedness*, the *computational aspects of apportionment and districting*, *group classification*, *group recommendation*, *social choice and crowdsourcing*, and *dynamic social choice*, are briefly reviewed in the chapter by Brandt et al. (2016c, Sect. 4), and the first four are reviewed in more detail by Endriss (2017). Lastly, new approaches to the *rationalization of voting rules*, partly originating in AI research, are reviewed in the chapter by Elkind and Slinko (2016).

3 Fair Allocation

Quand on partage le gâteau, l'important est : qui tient le couteau ?
(When the cake is divided, the main question is: who holds the knife?)
Bernard Maris, French economist, killed on January 7, 2015 in Paris

3.1 Fair Allocation Problems

Every CDM process is guided, explicitly or not, by the desirable properties that the collective decision must satisfy. We have seen in the introduction the most prominent of these properties: *efficiency*, which is often implemented by Pareto-efficiency. Another property which is very often required is *fairness*. Indeed, the essence of CDM is very often to look for admissible compromises between the agents antagonistic interests and preferences, which is a possible definition of fairness. We will later introduce several formal models of fairness.

The need for fairness is particularly strong in a kind of CDM problems called *fair allocation* or *fair division* problems, which are the subject of this section. Here, the alternatives are just *allocations* of goods (or resources) to agents. Even if the

traditional CDM problem assumes that the agents have preferences over all the alternatives, it is commonly assumed that each agent only cares about what she receives (her own *share*). In other words, an agent will be indifferent between two allocations where she receives the same share.

Fair allocation problems can be divided into classes. The first one concerns *divisible* goods or resources, like money, time, water or land. For a long time this class has been explored by economists, using continuous mathematics in microeconomics. The second one is about *indivisible* objects or resources, like works of art, pieces of furniture, teaching time slots, cars or houses. Mixed problems exist, a classical example being fair division problems with indivisible goods, but monetary compensation (money is a special divisible good, others are indivisible).

The contribution of artificial intelligence to the field of fair allocation mainly concerns fair allocation of indivisible goods without monetary compensation, which are the most difficult from an algorithmic point of view, because of their strong combinatorial nature. Actually, consider the allocation of m objects to n agents (each object must be allocated to one and only one agent), the number of possible allocations is n^m: the size of the solution space increases exponentially with the size of the problem instance. For similar reasons, artificial intelligence is involved also in fair allocation of divisible and heterogeneous resources (*cake-cutting*), see Sect. 3.5.

3.2 Some Real World Fair Allocation Problems

Before going further and in order to emphasize their importance, we enumerate a set of real world fair allocation problems.

1. frequency allocation to radio stations, land division, mining and natural resource sharing (Antarctic, ocean floor, Moon), common exploitation of a scientific facility, such as Earth-observing satellites (Lemaître et al. 1999);
2. fair representation (apportionment) (Balinski and Young 2001), setting up electoral boundaries;
3. fair allocation of kidneys or other organs to transplant;
4. allocation of positions in public entities;
5. sharing operating costs of international organisations, assessment of taxes;
6. allocation of permits to discharge pollutants;
7. sharing water treatment facilities between localities;
8. dividing estates in inheritance or divorce;
9. sharing time slots in schools, hospitals, airports,...;
10. allocating tasks or offices to employees, rooms to students, articles to reviewers, quotas of refugees to countries.

Notice that a lot of these problems concerns fair division of indivisible and non-shareable goods without monetary compensations.

3.3 How to Define Fairness?

Even if fairness is sometimes defined using the prominent Aristotelian's principle "equal treatment of equals, unequal treatment of unequals", there is no definitive and universal definition of fairness, but a number of properties corresponding to different formal definitions. None of these properties is universally considered to be the right notion of fairness, but each one conveys a different aspect of fairness. Some of these properties are defined on the collective preference (such as anonymity, separability, inequality reduction) while others apply to allocations (proportionality, envy-freeness).

Even if there is no universal notion of fairness, two properties are commonly required: *unanimity* and *anonymity*. Unanimity corresponds, in the context of fair division, to the Pareto-efficiency property, already discussed in Sect. 1.

3.3.1 Anonymity

If equal agents should be treated equally, then the *anonymity* property should be the first prerequisite. This properties conveys the fact that the collective preference should not depend on the agents' identities. Formally, for all permutation of agents σ, then the collective preorder aggregation function G, must satisfy

$$G(\succeq_1, \succeq_2, \ldots \succeq_n) = G(\succeq_{\sigma(1)}, \succeq_{\sigma(2)}, \ldots \succeq_{\sigma(n)}).$$

3.3.2 The Tension Between Unanimity and Strict Equality

In this paragraph we adopt the utilitarian model, with a common scale of utilities: for example, it is meaningful to say that a given allocation satisfies agent 1 more than agent 2.

In general, unanimity and strict equality cannot be satisfied at the same time. That is, there is in general no Pareto-efficient allocation giving to each agent the same amount of individual utility, as the following abstract situation involving two agents confronted to four possible allocations illustrates:

allocations	u_1	u_2
a	4	4
b	3	6
c	7	5
d	2	11

Allocation a is perfectly equitable, but is dominated by allocation c (c is better than a for each agent). Hence a is ruled out, in spite of its perfect equity. How to choose then between b, c and d? No one dominates another. Egalitarianism (maximizing the min) selects c whereas classical utilitarianism chooses the sum-efficient but obviously

unfair d. From an equity point of view, we are inclined to select c as the "optimal" allocation. But there are less obvious situations. Imagine having to choose between two allocations associated to the following utility profiles: $\langle 1, 49, 50\rangle$ and $\langle 2, 2, 96\rangle$. Or what about the case $\langle 14, 43, 43\rangle$ and $\langle 15, 15, 70\rangle$?

3.3.3 The Priority Principle

Another principle is sometimes considered in situations where anonymity is not completely relevant: the *priority* principle. Following this principle, an allocation decision should be based on agents' characteristics. For example, in the kidney allocation problem, patients having waited longer than others could have priority, or those having a longer life expectancy after transplant. Birthright is also a typical example.

3.3.4 Independence of Unconcerned Agents (IUA)

This property, also called *separability*, applies to the collective preference. According to IUA, the collective preference should be such that an agent who is indifferent between two allocations can be ignored when choosing between these two allocations (she is not concerned by the choice). In other words, if this property is not satisfied, the collective preference between two allocations for which an agent is indifferent, depends on the precise individual utility of this agent for these allocations, which is hardly acceptable.

Consider the following example (Moulin 1988). There are three agents and the collective utility aggregation function g is the median. Let a and b be two allocations with respective utility profiles $\langle 0, 2, 3\rangle$ and $\langle 0, 1, 4\rangle$. We have $g(\langle 0, 2, 3\rangle) > g(\langle 0, 1, 4\rangle)$ hence $a \succ_{col} b$. Now, consider two other allocations a' and b', with respective profiles $\langle 5, 2, 3\rangle$ and $\langle 5, 1, 4\rangle$ (utilities of agents 2 and 3 are not modified, but utility of agent 1 is raised from 0 to 5. We have now $g(\langle 5, 2, 3\rangle) < g(\langle 5, 1, 4\rangle)$, that is $b' \succ_{col} a'$. The preference is reversed. Agent 1 is not concerned by the choice between a and b, but her individual utility influences the choice between allocations which does not change utilities of others! The collective preorder represented by the median does not satisfies the IUA property.

This property is connected to the following important result. A collective preorder is continuous and satisfies the IUA property if and only if this preorder is represented by an additive collective utility aggregation function $g(\overrightarrow{u}) = \sum_i f(u_i)$ where f is continuous and increasing.

3.3.5 Inequality Reduction

This property supposes the utilitarian model with a common individual utility scale.

We have to define first what is called an *inequality reducing transfer*, or *Pigou-Dalton, transfer*. Let $\overrightarrow{u} = \langle u_1, u_2, \ldots, u_n\rangle$ be a utility profile, with $u_1 < u_2$.

Consider a utility transfer from agent 2 to agent 1 (from the richest to the poorest) such that after this transfer, $\overrightarrow{u}$ becomes $\overrightarrow{v}$ with $u_1 + u_2 = v_1 + v_2$ and $|v_2 - v_1| < |u_2 - u_1|$. Such transfer is said to reduce inequalities.

The inequality reduction property requires that any inequality reduction transfer does not strictly decrease the collective utility. Formally, the preorder $\succeq_{col}$ represented by the aggregation function g reduces inequalities when, for any pair of utility profiles $\overrightarrow{u}$ and $\overrightarrow{v}$ equal except on their first and second components, such that $u_1 < u_2$, $|v_2 - v_1| < |u_2 - u_1|$ and $u_1 + u_2 = v_1 + v_2$, we have $g(\overrightarrow{u}) \leq g(\overrightarrow{v})$.

Here is an example with three agents and $g(\langle u_1, u_2, u_3\rangle) = u_1^2 + u_2^2 + u_3^2$. Let a and b be two allocations respectively associated to the utility profiles $\langle 0, 3, 4\rangle$ and $\langle 1, 2, 4\rangle$. From a to b, inequalities are reduced, however $g(\langle 0, 3, 4\rangle) = 25 > g(\langle 1, 2, 4\rangle) = 21$ which means that a is preferred to b. We conclude that the collective preference does not obey the inequality reduction property in this case.

An interesting fact is connected with this property and the separability (IUA) property: the preorder $\succeq_{col}$ represented by the additive aggregation function $g(\overrightarrow{u}) = \sum_i f(u_i)$ reduces inequalities if and only if f is a concave function. In the previous example, $f(x) = x^2$ is convex.

We now turn to properties which apply to allocations.

3.3.6 Proportionality (or Proportional Fair Share) and Max-Min Fair Share

An allocation satisfies proportionality when each agent gets at least $1/n$ of the total utility she would have received if she receives alone all objects (n is the number of agents). This property was coined by Steinhaus in 1948 in the context of continuous fair division (cake-cutting) problems.

Proportionality has been adapted to indivisible goods without monetary compensation by Budish (2011), which defines the *max-min fair share* property. The original definition is purely ordinal. In utilitarian terms, the max-min fair share of an agent is the maximal utility that she can hope to get from an allocation if all the other agents have the same preferences as her, when she always receive the worst share (it is the best of the worst shares). An allocation satisfies the max-min fair share property if each agent receives a utility no less than the utility of her max-min fair share. Proportionality implies max-min fair share.

3.3.7 Envy-Freeness

This very general and intuitive property does not require interpersonal comparisons of utility (just like proportionality). It both applies to the ordinal and cardinal (utilitarian) models.

An allocation is envy-free if no agent strictly prefers the share of another agent to her own share. It is a kind of stability property. Formally, let $a_{i/j}$ be an allocation

identical to a except that agent i gets the share that was the share of agent j in a. We say that a is envy-free if $a \succeq_i a_{i/j}$ for all agents i, j.

Envy-freeness and Pareto-efficiency are generally not compatible. Furthermore, the problem of determining whether an allocation satisfying envy-freeness and Pareto-efficiency exists is quite complex; even in most reasonable settings (see for instance, Bouveret and Lang 2008).

Under additive preferences, envy-freeness implies proportionality. Other fairness properties and their relations to each other are presented in the paper by Bouveret and Lemaître (2016).

3.4 Main Aggregation Functions

In the utilitarian model, a family of aggregation functions is particularly interesting in the context of fair division, namely the *root-power quasi-arithmetic means* family, defined as follows (assuming strictly positive utilities):

$$g_p(\overrightarrow{u}) = \left(\frac{1}{n}\sum_i u_i^p\right)^{1/p}, p \neq 0 \qquad g_0(\overrightarrow{u}) = \left(\prod_i u_i\right)^{1/n}, \text{for } n = |\overrightarrow{u}|$$

The family is parameterized by the real number p. Functions of this family are additive[11] and hence induced preorders obey the separability (IUA) property (see p. 23). When $p = 1$, g_1 is the standard arithmetic mean, corresponding to classical utilitarianism. The case $p = 0$ corresponds to the Nash function, which is independent of individual scale of agents utilities, a particularly interesting property. The collective preorder induced by g reduces inequalities if and only if $p < 1$. Finally, when p tends to $-\infty$, g tends toward the min function, and the induced preorder tends toward the leximin preorder.[12]

Notice that this family creates a continuum between the extremes classical utilitarianism (sum) and egalitarianism (min).

Another family of interest is the family of *ordered weighted averages* (OWA) (Yager 1988), a variant of the weighted averages in which weights do not hold on the components but on the ranks. A n-OWA is a family of aggregation functions from $\mathbb{R}^n$ into $\mathbb{R}$, taking $\overrightarrow{w} = \langle w_1, \ldots, w_n\rangle \in [0, 1]^n$ as parameter, with $\sum_i w_i = 1$, defined by $O_{\overrightarrow{w}}(\overrightarrow{a}) = \sum_{i=1}^n w_i \cdot a_i^\star$, where $\langle a_1^\star, a_2^\star, \ldots, a_n^\star\rangle$ is $\langle a_1, a_2, \ldots, a_n\rangle$ once sorted weakly increasing. OWAs can express: the mean ($w_i = 1/n$ for all i); the

[11] Strictly speaking, these function are not additive. However the preorders they induce can be represented by additive functions, derived from original ones by increasing transformations. Even g_0 is additive in the broad sense of the term, because the additive function $\sum_i \log(u_i)$ represents the same preorder.

[12] The leximin preorder is a refinement of the preorder induced by the min function which satisfies unanimity (Pareto-efficiency). The leximin preorder is precisely the one which at the same time reduces inequalities and is independent of the common scale of utilities.

min ($w_1 = 1$, and $w_i = 0$ for all $i > 1$); the median ($w_{(n+1)/2} = 1$, and $w_i = 0$ for $i \neq (n+1)/2$); parameterizable compromises between min and mean, for example $w_i = \alpha^i$, $0 < \alpha < 1$ (with a suitable normalization); a function which tends towards a representation of the leximin preorder (the previous one when α tends towards 0).

3.5 *Procedural Allocation of a Divisible and Heterogeneous Resource (Cake-Cutting)*

The previous allocation model — choosing an allocation that maximizes an appropriate collective utility function — is based on two implicit assumptions. First, each agent should completely and honestly report her preferences (under the form of a utility function). Second, the agents rely on a central entity that is in charge of computing the allocation. In some cases, the agents do not wish to publicly report their preferences, and even if they accept to do so, nothing prevents them from acting strategically and misreporting them. Finally, the agents can simply refuse to trust a central authority. Hence the model based on the central optimization of a collective utility function is not adapted to every situation.

There exists a very different kind of allocation *procedures*, that have been studied for years. These procedures are by essence distributed and output a fair and efficient allocation from the preferences or a small fraction of them reported (honestly if possible) by the agents. These procedures — often called *mechanisms* — are particularly used in the context of the allocation of an infinitely divisible and heterogeneous good. The traditional metaphor is the *cake-cutting* situation: a rectangular cake (formally modeled has the [0, 1] real interval) is the common divisible and heterogeneous resource, and has to be shared among the n starving invitees, which all have a particular utility function on this interval.[13]

The allocation procedures that are studied in this context are similar to games in which the agents interact. The most prominent procedure in this context is the well-known "I cut, you choose" game, that can be used to cut a cake between two participants which can be roughly described as follows. One of the two participants is the divider. The other one is the chooser. Provided that the cake can be divided in all possible ways, that it is heterogeneous and that the participants can have different tastes for different parts of the cake, cutting the cake in two equal parts is in general not Pareto-optimal. The safest action for the divider is to cut the cake into two pieces that are equal *from her point of view*. Then the chooser will take the best of the two pieces. Under mild natural assumptions,[14] it is easy to see that the resulting allocation is Pareto-efficient, envy-free and proportional (satisfying the fair share

[13] Or at least an ordinal function on intervals: between two intervals, each agent must be able to determine which of the two is better.

[14] The agents are rational (they decide so as to maximize their satisfaction) and their utility is additive in the ordinal sense: if $A \succ B$, $C \succ D$ and $A \cap C = \emptyset$, then $A \cup C \succ B \cup D$.

property). However, as we shall see later, the generalization of this protocol to three agents or more is not straightforward.

The allocation problem of divisible and heterogeneous goods has a lot of real-world applications, among which we can mention the time-sharing of a common facility, or the land division problem.

A large number of works are dedicated to this problem, essentially in the field of economics. The seminal books by Robertson and Webb (1998), Young (1994, Chaps. 8 and 9) and Brams and Taylor (1996) are good surveys of the topic. The interested reader can also have a look at the more recent paper by Brams et al. (2006). Plenty of procedures are now well-known and well-studied, adapted to different contexts, and characterized by their fairness and efficiency properties. Some impossibility theorems have also been described.

More recently, researchers in artificial intelligence have contributed to the field of cake-cutting. They have mainly focused on the algorithmic complexity of the proposed procedures. The analysis of the complexity bounds requires the introduction of precise models of interaction between the agents. From this point of view, the researchers in AI also contribute: for instance, the classical model of interaction, that has been proposed by Robertson and Webb (1998), has been recently extended by a group of computer scientists (Brânzei et al. 2016). For interested readers, the paper by Procaccia (2009) proposes an interesting survey of the complexity bounds of cake-cutting procedures and the chapter by Procaccia (2016) gives an overview on algorithmic aspects of cake-cutting.

To give an idea of the difficulty of the mathematical problems we have to deal in this area and how computer scientists have contributed to the field, let us go back to the aforementioned problem of finding a protocol to find an envy-free cake cutting for three agents or more. In the early 60's, Selfridge and Conway independently came up with a protocol that returns an envy-free cake cutting in a bounded number of steps for three agents (Brams and Taylor 1996). A few decades later, Brams and Taylor (1995) came up with a general envy-free protocol that works for any number of agents. This protocol is guaranteed to terminate in finite time; however, the number of queries needed can be unbounded, even for four agents. The problem of finding a protocol that returns an envy-free allocation in a bounded number of queries for any number of agents had been opened for decades until it was finally solved by two computer scientists, Aziz and Mackenzie (2016). They came up with an algorithm that finds an envy-free allocation in less than $n^{n^{n^{n^n}}}$ queries, hence closing what has been described by several researchers as one of the main challenges in the field.

To finish this overview, we can mention some extension of the cake-cutting problem (called online cake-cutting) in the case where some participants arrive and depart when the allocation process is ongoing (Walsh 2010).

3.6 Fair Division and Computer Science

As we have seen earlier, resource allocation problems have long been mainly studied by economists, either from the normative or axiomatic point of view (as in the works by Young 1994 and Moulin 2003 for instance) or from the procedural point of view like in the works by Brams and Taylor (1996, 2000) about cake-cutting. However, like in voting, computer scientists and AI and OR researchers have started for a couple of years to investigate the computational aspects of resource allocation problems: compact preference representation, algorithmic or complexity issues... The vitality of the field is well illustrated by the chapters dedicated to fair division in the survey books on computational social choice (Brandt et al. 2016b, Chaps. 11–13) (Rothe and Rothe 2015, Chaps. 7 and 8).

3.6.1 Compact Preference Representation

Even if a lot of work has been done in recent years on the topic of compact preference representation (see chapter "Compact Representation of Preferences" of this volume dedicated to this topic), only a small fraction of this work directly concerns resource allocation problems, with the notable exception of combinatorial auctions (as we will see in Sect. 4). On the one hand the domain of compact preference representation is quite young and resource allocation problems only represent a small fraction of the individual or collective decision making problems involving compact representation issues. On the other hand, a lot of works dedicated to algorithmic or complexity issues of fair division problem simply rule out these compact representation problems by assuming that the preferences are additive: see for instance the paper by Lipton et al. (2004) mainly dedicated to additive preferences, or the works by Bezáková and Dani (2005), Bansal and Sviridenko (2006), Asadpour and Saberi (2007) on the Santa-Claus problem.

The first papers explicitly dedicated to compact preference representation in fair division problems date back to the works by Chevaleyre et al. (2004) on k-additive functions and those by Bouveret et al. (2005) and Bouveret and Lang (2008) mainly concerning logic-based compact representation. We can also mention an adaptation of the language of CP-nets for the compact preference representation in the context of fair division problems, that was proposed by Bouveret et al. (2009). It seems however that not much work has been done since on compact representation in fair division, and that AI researchers have focused on simpler settings like additive preferences. One of the reasons might be that it is not so clear whether the benefit of using an expressive compact representation language is worth the additional complexity cost and elicitation burden or not, and that the use of simple additive preferences might well be enough for most applications.

3.6.2 Complexity and Algorithmic Issues

At the beginning of computational social choice, most works especially dedicated to algorithmic aspects of resource allocation are from the fields of combinatorial auctions or operations research. In the latter domain, fair division problems have been mostly considered from the point of view of "fair" multicriteria optimisation problems, that is, optimisation problems where the criteria should be maximized (or minimized) while being made as equal as possible. For instance, leximin, or (inequality-reducing) OWA optimization problems belong to this kind of problems. Among these works, to cite only two, Ogryczak (1997) applies fair optimization to fair facility location problems, and Luss (1999) to fair division.

Several works have followed the earlier paper about algorithmic aspects of fair division in different fields of research, both in artificial intelligence and operations research. For instance, a very active stream of works has concerned algorithmic and complexity aspects of a particular fair division problem, the *Santa-Claus Problem*. This problem can be formulated as follows: Santa-Claus has to allocate a set of m (indivisible, non shareable) toys to a set of n children. Each child has an additive utility function on the set of toys. The allocation must be made so as to maximize the utility of the least satisfied child. Of course, in spite of this special formulation, this problem is nothing else than a fair division problem with indivisible goods, additive preferences, and under the egalitarian criterion. Several papers (just to name a few, Bezáková and Dani 2005; Bansal and Sviridenko 2006; Asadpour and Saberi 2007) have investigated the complexity and approximation algorithms of this problem. They have also drawn a interesting parallel with scheduling problems that have led to fruitful approximation approaches.

Still in the context of additive preferences, several works have also investigated other fairness criteria. The seminal paper by Lipton et al. (2004) is one of the first works concerning the complexity and approximation of computing an allocation minimizing the envy between agents.[15] de Keijzer et al. (2009) have extended the latter paper by proving that the problem of determining whether an envy-free and Pareto-efficient allocation was NP-complete. Later on, the case of ordinal separable (*i.e.* additive) preferences has been investigated by Bouveret et al. (2010), and further by Aziz et al. (2015) that have introduced interesting notions of envy-freeness and Pareto-efficiency based on stochastic dominance. Concerning utility maximization problems, Bouveret and Lemaître (2009) have focused on the computation of leximin-optimal solutions using constraint programming approaches. Golden and Perny (2010) and Lesca and Perny (2010) have also studied preference aggregation in particular in the context of fair division problems, focusing on fairness criteria like Lorenz optimality, maximization of an OWA or of a Choquet integral (extension of the OWA that can take into account positive or negative interactions between agents).

Finally, several papers go beyond additive preferences and investigate theoretical complexity issues related to the use of compact preference representation languages

[15] This work introduces an interesting extension of the aforementioned envy-freeness criterion, by proposing different *measures* of envy.

in the context of fair division. For utility maximization problems, the paper by Dunne (2005) was one of the first works investigating the complexity of fair division problems with preference representation based on Straight-Line Programs. At the same time, Bouveret et al. (2005) have focused on preference representation based on propositional logic. All these results have been extended to other social welfare functions, other languages like k-additive languages, and approximation issues by Nguyen et al. (2014). Finally, the complexity of finding envy-free allocations with logic-based compact preference languages has been investigated by Bouveret and Lang (2008).

3.6.3 Distributed Allocation and Communication Complexity

Even if distributed allocation and negotiation issues in fair division will mainly be discussed in chapter "Negotiation and Persuasion among Agents" of this volume, an overview of computer science aspects of resource allocation would be incomplete without evoking this domain. In the absence of any central authority, the natural way of computing an "optimal" allocation is to start from an initial allocation and then let the agents changing it using multilateral negotiation. In this framework, the main desirable properties are related to the convergence of the negotiation process, and the complexity is not defined in terms of computation, but in terms of communication costs (number of steps, size of the messages exchanged...).

The first theoretical results in this domain date back to Sandholm (1998). The notion of communication complexity has been imported in fair division in particular by Endriss and Maudet (2005) and Dunne et al. (2005), that mainly focus on the number of swaps needed to reach an optimal allocation. We can also mention, among other works on the subject, the paper by Chevaleyre et al. (2007) that focuses on a relaxation of the envy-freeness criterion, for which the agents only have a limited knowledge of the other agents. Finally, a paper by Chevaleyre et al. (2017) analyzes the fairness properties (like envy-freeness or proportionality) of the allocations obtained after the convergence of the negotiation process with several kinds of deals, and also in the case where the set of possible swaps is constrained by a graph.

3.6.4 Recent Trends

We can observe an interesting recent trend in the community of computational fair division. As already noticed in the section dedicated to compact preference representation, a significant trend is to abandon complex theoretical frameworks and look for simple models that are inspired by practical applications and can be used in real situations. This has led to a stream of works trying to implement fair division in practice (see for instance the wonderful web application Spliddit[16]), or take inspiration

[16] http://www.spliddit.org/.

from practice to find better ways of defining fairness or motivating fairness criteria. With simplicity in mind, most of these works are based on additive preferences.

Among the works proposing some alternative approaches to fairness, Bouveret and Lemaître (2016) introduce a scale of five fairness criteria of increasing strength. This scale can be used as a measure of the level of conflict of the agents' preferences: the higher criterion it is possible to satisfy, the less conflicting the preferences are, and the more likely it will be possible to find a satisfactory allocation. Among the five criteria, the maximin share and the Competitive Equilibrium from Equal Income (CEEI) criteria were already known in economics, but had been ignored so far by computer scientists. This is no longer the case, and it had led to fruitful works on theoretical properties, approximation and complexity, either about CEEI (Brânzei et al. 2015; Aziz 2015), or about the maximin share (Procaccia and Wang 2014; Amanatidis et al. 2015; Kurokawa et al. 2015). Caragiannis et al. (2016b) have also revisited a long-standing criterion, the seminal Nash social welfare function, and shed a new light on its appealing fairness properties.

Finally, we can also mention an interesting work that perfectly characterizes the fruitful collaboration between economics and computer science. Dickerson et al. (2014) analyze, both in practice and analytically, the probability of existence of an envy-free allocation, depending on the ratio between the number of objects and the number of agents: when this ratio is low, an envy-free allocation is very likely not to exist, and the opposite when the ratio is high enough. Furthermore, the simulations show a very interesting phase-transition phenomenon.

4 Combinatorial Auctions

4.1 From Classical to Combinatorial Auctions

Auctions is probably one of the most widely studied collective decision making problem in the economical literature in the last fifty years. In its most general definition, an auction is simply a structured mechanism in which some agents, the bidders, compete for some objects to buy. The mechanism (which is in practice implemented by a central entity, the auctioneer) is in charge of determining which objects each agent will get at which price. A wide variety of mechanisms are studied by economists and used in practice. Just to name a few, an auction can be *open* if the participants publicly announce their bids, or *sealed* if the bidders only reveal this information to the auctioneer and hide it from the other participants. An auction is *ascending* if the bidders iteratively increase their bids until no agent is willing to pay a higher price, and *descending* if the price proposed for an object decreases until at least one bidder declares to be interested. In a *first price* auction, the winner should pay the price corresponding to the highest bid, whereas in a *second price* auction, she should pay the second highest price.

The most common auctions types are the *English auction* (open first-price ascending) which is commonly used in artwork sales. The *Dutch auction* (open first-price descending, in which the auctioneer progressively decreases the price proposed for an object until a bidder accepts and pays this price) is traditionally used for perishable products like tulips in the Netherlands. The *Vickrey auction* (sealed second-price) is also called philatelic auction because it is used in the United States for collection stamps sales. Finally, the sealed first-price auction is classically used for the attribution of government contracts.

Auction theory has been studied for about 50 years mostly in economics — the first theoretical work on auctions is generally attributed to Vickrey (1961) — but computer scientists have recently paid an increasing attention to this field, mainly with the study of *combinatorial auctions*.

The study of combinatorial auctions in computer science dates back to the work of Rassenti et al. (1982). The starting point of this work is that classical auctions mechanisms, sequential by nature (that is, selling objects one by one) can be quite inefficient and barely adapted to the situations where the bidders have non-modular preferences on the objects; in other words, when they have are preferential dependencies.

Let us consider a simple example where three objects are for sale: a vinyl disc player (p) and two (indivisible) sets of vinyl discs: the first set contains records from the Beatles (b) and the second one records from the Rolling Stones (s). Agent 1 is very interested in having one of the two sets of records (no matter which of the two), but does not own any disc player. Moreover, she is not interested in buying the disc player alone, because she does not have any vinyl disc to listen to. She is for instance ready to pay €100 for $\{p, b\}$ or $\{p, s\}$, €110 for $\{p, b, s\}$ but nothing for individual objects. In other words, p and b are complementary, p and s are as well, but s and b are substitutes. Agent 2 is also interested in the sets of records, but she already owns a disc player. Let us say that she is thus ready to pay €30 for $\{b\}$ or for $\{s\}$, €10 for $\{p\}$, €40 for $\{p, b\}$ or $\{p, s\}$ and €70 for $\{p, b, s\}$ (her preferences are additive).

If the objects are sold sequentially, the first agent will probably have difficulties to bid for them. Not knowing Agent 2's preferences, she will probably not take the risk of bidding for one of the two sets of records if she does not know for sure whether she will get the disc player (and the other way around). Agent 2 will not have the same difficulties: her preferences being additive, she can safely bid on each of the three objects separately. Not only sequential allocation can have a negative effect on bidding, but it can also harm the overall auction efficiency. For instance, in the latter auction, if Agent 1 is risk-adverse and chooses not to bid at all, the three objects will go to Agent 2 for a total price of €70 (provided that it is a first-price auction). If all three objects had been allocated to Agent 1, the auctioneer would have earned €110.

An obvious solution to this problem is to sell *bundles* of objects instead of selling them individually.[17] However, it it not obvious how to do so, because the way the bundles are formed should be related to the preferential dependencies of the agents. In some cases it is reasonable to assume that these dependencies are the same (in a shoe

[17] Other methods exist; see for instance simultaneous ascending auctions (Cramton 2006).

sale for instance, we can reasonably assume that the agents will only be interested in pairs of shoes, not individual shoes), but it is not always the case. For instance, in the latter auction, would it be more relevant to sell p and s together, or b and s together? The only solution to this problem is to sell all the objects simultaneously, and provide a way for the bidders to choose themselves the bundles they want to bid for. This is the basic idea behind combinatorial auctions. It has motivated the first works in this domain, concerning the allocation of take-off and landing slots in airports (Rassenti et al. 1982). In this application, the notion of preferential dependency is naturally present (what would an airline do with a take-off slot without the corresponding landing slot?).

It is not a surprise that this extension of classical auctions has been mainly developed and studied in the field of computer science and artificial intelligence. A lot of problems that arise in combinatorial auctions are well-known in computer science. As we shall see, the combinatorial blow-up induced by the representation of the allocation space calls for compact representation bidding languages. Furthermore, the problem of determining the optimal allocation is a lot more complex than in traditional auctions and induces intricate algorithmic issues. Finally, even if we will elude these aspects in the chapter, issues related to the design of truth-telling auctions mechanisms and their resistance to manipulation is a crucial topic. They are not only related to combinatorial auctions, but have a special formulation in this context. All these topics are covered in details in the reference book by Cramton et al. (2006).

In what follows, we will denote by $\mathcal{O}$ the finite set of objects to be allocated to the agents (the set of objects the agents bid for). Given a set of n agents $\mathcal{N}$ and a set of objects $\mathcal{O}$, an *allocation* $\overrightarrow{\pi}$ is a vector $\langle \pi_1, \ldots, \pi_n \rangle$, where for all i, $\pi_i \subset \mathcal{O}$ denotes the *share* received by agent i. In this section, we will only focus on allocations satisfying the preemption constraint, that is, such that $\forall i \neq j : \pi_i \cap \pi_j = \emptyset$ (an object cannot be allocated to two different agents).

4.2 Bidding Languages

As we have seen, the main difference between combinatorial and classical auctions is the bidding set, which is namely the set of objects $\mathcal{O}$ for classical auctions and the set of bundles $2^{\mathcal{O}}$ for combinatorial auctions. From the theoretical point of view, changing the bidding set does not make a huge difference in the formal definition of the problem. However, in practice, the combinatorial dimension of the bidding set induces crucial representation[18] and computation issues.

The most prominent bidding languages used in combinatorial auctions are the ones from the family of XOR / OR / OR* languages (Nisan 2006; Fujishima et al. 1999; Sandholm 2002).

[18] A simplistic representation of an agent's utility function requires $2^m - 1$ values, corresponding to the number of non-empty subsets of $\mathcal{O}$.

Definition 1 (*XOR / OR / OR⋆ languages*) Let $\mathscr{O}$ be a finite set of objects. An *atomic bid* on $\mathscr{O}$ is a pair $\langle \mathscr{S}, w \rangle \in 2^{\mathscr{O}} \times \mathbb{R}^+$. A set $\{\langle \mathscr{S}_1, w_1 \rangle, \ldots, \langle \mathscr{S}_p, w_p \rangle\}$ of atomic bids is said to be *admissible* if $\mathscr{S}_i \cap \mathscr{S}_j = \emptyset$ for all $i \neq j$ in $\{1, \ldots, p\}$.

A bid expressed in the *XOR language* is a finite set of atomic bids

$$\langle \mathscr{S}_1, w_1 \rangle \text{ XOR } \ldots \text{ XOR } \langle \mathscr{S}_p, w_p \rangle.$$

The utility function associated to a bid $\mathscr{M}$ in the XOR language, mapping each set of objects π (in other words each possible share) to the price the agent is ready to pay for it, is defined as follows:

$$\begin{aligned} u : 2^{\mathscr{O}} &\to \mathbb{R}^+ \\ \pi &\mapsto \max_{\substack{\langle \mathscr{S}_i, w_i \rangle \in \mathscr{M} \\ \mathscr{S}_i \subseteq \pi}} w_i \end{aligned}$$

A bid expressed in the *OR language* is a finite set of atomic bids

$$\langle \mathscr{S}_1, w_1 \rangle \text{ OR } \ldots \text{ OR } \langle \mathscr{S}_p, w_p \rangle.$$

The utility function associated to a bid $\mathscr{M}$ in the OR language is:

$$\begin{aligned} u : 2^{\mathscr{O}} &\to \mathbb{R}^+ \\ \pi &\mapsto \max_{\substack{\mathscr{M}' \subseteq \mathscr{M} \\ \mathscr{M}' \text{ admissible}}} \sum_{\substack{\langle \mathscr{S}_i, w_i \rangle \in \mathscr{M}' \\ \mathscr{S}_i \subseteq \pi}} w_i \end{aligned}$$

A bid expressed in the *OR⋆language* is a bid expressed in the OR language in which one or several dummy objects $d \notin \mathscr{O}$ can be present.

In the XOR language, an agent can cast a custom set of atomic bids. Each atomic bid gives the price the agent is ready to pay for the corresponding bundle. Given a set of objects the price an agent is ready to pay for it is the price of the best bundle it contains.

The OR language works the same way, except that in this language, the prices are interpreted additively.

By adding dummy objects to the bids in the OR⋆ language the agents can make several bids incompatible, as in the XOR language, even if they do not overlap otherwise.

Some authors (see for instance the work by Sandholm 1999) have proposed to combine the OR and XOR language to benefit from the expressivity of the XOR language and the compactness of the OR language. Several languages have been developed and used, among which we can mention the OR-of-XOR, XOR-of-OR and OR / XOR languages.

Some other kinds of bidding language have also been developed, among others logical ones. For instance, Boutilier and Hoos (2001) have proposed a language mixing logic and numerical weights (representing utilities) associated to subformulas.

The main interest of such a language is to combine the approach based on objects and the approach based on bundles, by proposing a way to logically combine the weighted formulas (that can be seen as the atomic "bids" of this language).

4.3 The Winner Determination Problem

4.3.1 Formulation and Theoretical Complexity

The Winner Determination Problem (WDP for short) is the central problem in combinatorial auctions. The objective is to decide, among the set of bids, which ones will be selected, coming down to determine which objects will be allocated to which agents. The main allocation criterion used in combinatorial auctions is the utilitarian criterion, that is, we look for the allocation that maximizes the revenue of the auctioneer.

Definition 2 (*Winner Determination Problem*)

- **Input**: A set of agents $\mathcal{N}$, a set of objects $\mathcal{O}$, and a set of utility functions $(u_1, \ldots, u_n)$ expressed as bids in a combinatorial auction language.
- **Output**: An allocation $\overrightarrow{\pi}$ of the objects that maximizes $\sum_{i=1}^{n} u_i(\pi_i)$.

Notice that this formulation of WDP implicitly assumes that the auctioneer can freely dispose objects (in other words, the allocation can be incomplete), which is a common assumption in combinatorial auctions.

The Winner Determination Problem has mainly been studied in the context of OR or XOR bids, for which there is a natural formulation in 0–1 linear programming. The idea is to create a variable $\mathbf{x_i^j} \in \{0, 1\}$ for each atomic bid $\langle \mathcal{S}_j, w_j \rangle \in \mathcal{M}_i$. $\mathbf{x_i^j} = 1$ if and only if this atomic bid is selected in the allocation.

$$
\begin{aligned}
&\textbf{max} \sum_{i \in \mathcal{N}} \sum_{\mathcal{S}_j \in \mathcal{M}_i} w_j \times \mathbf{x_i^j} \\
&\textbf{s.t.}\ \mathbf{x_i^j} \in \{0, 1\} \\
&\quad \sum_{i \in \mathcal{N}} \sum_{\substack{\mathcal{S}_j \in \mathcal{M}_i \\ o \in \mathcal{S}_j}} \mathbf{x_i^j} \leq 1 \text{ for all } o \in \mathcal{O} \text{ (OR constraint)} \\
&\text{or } \sum_{\mathcal{S}_j \in \mathcal{M}_i} \mathbf{x_i^j} \leq 1 \text{ for all } i \in \mathcal{N} \qquad \text{(XOR constraint)}
\end{aligned}
$$

We can notice that this formulation of the OR and XOR Winner Determination Problem makes the problem strictly equivalent to the well-known knapsack problem. It implies that the general decision version of the problem is NP-complete (Rothkopf et al. 1998), but it also remains NP-complete even with very restrictive assumption about the values and the kind of bids allowed, and also on the number of agents (Lehmann et al. 2006).

4.3.2 Optimal Solving

In spite of the complexity of the WDP for OR, XOR languages and their variants, quite large instances can be nevertheless efficiently solved by state-of-the-art linear solvers running on the previous formulation of the WDP. However, the use of *ad hoc* branching approaches (see for instance chapter "Heuristically Ordered Search in State Graphs" of volume 2) that are tailored to this particular problem give even better results.

There are two natural ways of solving the WDP with a branching algorithm. The first possibility is to branch on objects, that is, to choose an object at each node of the search tree and to decide to which bid this object will be allocated. To take into account the free disposal assumption, a classical approach is to create a dummy bid containing all the objects and to which will be allocated all the disposed items. This branching approach can be used in combination with several methods that drastically reduce the size of the search space. For instance, some parts of the search tree can be pruned by only allocating objects to bids that have not already been considered in the previous branches. The second way of solving the WDP is to branch on bids, that is, to choose at each node of the search tree an atomic bid and decide whether it will be satisfied or not. Maintaining a conflict graph between bids, that updates when the bids are selected or discarded, dramatically improves the algorithm efficiency.

5 Conclusion

In this chapter, we have presented the foundations of (mainly) centralized collective decision making.[19]

This domain, that has originally been mostly studied by political scientists and economists, has recently met computer science and more specifically artificial intelligence. This very active scientific domain born from this convergence has been called *computational social choice*. To illustrate the scientific activity in this domain, we have presented in this chapter three prominent centralized collective making problems: voting, fair division and combinatorial auctions. For each of these domains, we have presented the main works related to artificial intelligence.

Centralized collective decision making proceeds by directly aggregating the agents preferences into a collective preferred decision that is not supposed to be changed afterwards. A different approach to collective decision making is to let the agents interact and negotiate: this is the case in *distributed* collective decision making, presented in chapter "Negotiation and Persuasion among Agents" of this volume, that will complement the overview given in this chapter.

[19] The cake-cutting setting, that has also been introduced in this chapter, is an exception: the resolution of this problem is based on interactions between agents, which is by definition not a centralized process.

References

Ailon N, Charikar M, Newman A (2005) Aggregating inconsistent information: ranking and clustering. In: Proceedings of ACM Symposium on Theory of Computing (STOC'05)

Alon N (2006) Ranking tournaments. SIAM J Discret Math 20(1):137–142

Amanatidis G, Markakis E, Nikzad A, Saberi A (2015) Approximation algorithms for computing maximin share allocations. In: ICALP (1). Lecture notes in computer science, vol 9134. Springer, Berlin, pp 39–51

Arrow K (1951) Social choice and individual values, 2nd edn (1963). Wiley, New Jersey

Asadpour A, Saberi A (2007) An approximation algorithm for max-min fair allocation of indivisible goods. In: Proceedings of ACM Symposium on Theory of Computing (STOC'07), pp 114–121

Aziz H (2015) Competitive equilibrium with equal incomes for allocation of indivisible objects. Oper Res Lett 43(6):622–624

Aziz H, Mackenzie S (2016) A discrete and bounded envy-free cake cutting protocol for any number of agents. In: 2016 IEEE 57th annual symposium on foundations of computer science (FOCS). IEEE, pp 416–427

Aziz H, Gaspers S, Mackenzie S, Walsh T (2015) Fair assignment of indivisible objects under ordinal preferences. Artif Intell 227:71–92

Bachrach Y, Betzler N, Faliszewski P (2010) Probabilistic possible-winner determination. In: Proceedings of AAAI conference on artificial intelligence (AAAI'10)

Balinski ML, Young HP (2001) Fair representation : meeting the ideal of one man one vote, 2nd edn. Brookings Institution Press, USA

Bansal N, Sviridenko M (2006) The Santa Claus problem. In: Proceedings of ACM Symposium on Theory of Computing (STOC'06), pp 31–40

Bartholdi J, Orlin J (1991) Single transferable vote resists strategic voting. Soc Choice Welf 8(4):341–354

Bartholdi J, Tovey C, Trick M (1989a) The computational difficulty of manipulating an election. Soc Choice Welf 6(3):227–241

Bartholdi J, Tovey C, Trick M (1989b) Voting schemes for which it can be difficult to tell who won the election. Soc Choice Welf 6(3):157–165

Bartholdi J, Tovey C, Trick M (1992) How hard is it to control an election? Math Comput Model 16(8/9):27–40

Betzler N, Niedermeier R, Woeginger GJ (2011) Unweighted coalitional manipulation under the Borda rule is NP-hard. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'11)

Bezáková I, Dani V (2005) Allocating indivisible goods. SIGecom Exch 5(3):11–18

Boutilier C, Hoos HH (2001) Bidding languages for combinatorial auctions. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'01), pp 1211–1217

Boutilier C, Rosenschein J (2016) Incomplete information and communication in voting. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge

Bouveret S, Endriss U, Lang J (2009) Conditional importance networks: a graphical language for representing ordinal, monotonic preferences over sets of goods. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'09), pp 67–72

Bouveret S, Endriss U, Lang J (2010) Fair division under ordinal preferences: computing envy-free allocations of indivisible goods. In: Proceedings of European Conference on Artificial Intelligence (ECAI'10)

Bouveret S, Fargier H, Lang J, Lemaître M (2005) Allocation of indivisible goods: a general model and some complexity results. In: Proceedings of International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'05)

Bouveret S, Lang J (2008) Efficiency and envy-freeness in fair division of indivisible goods: logical representation and complexity. J Artif Intell Res 32:525–564

Bouveret S, Lemaître M (2009) Computing leximin-optimal solutions in constraint networks. Artif Intell 173(2):343–364
Bouveret S, Lemaître M (2016) Characterizing conflicts in fair division of indivisible goods using a scale of criteria. Auton Agents Multi-Agent Syst 30(2):259–290
Brams SJ, Fishburn P (2004) Voting procedures. In: Arrow K, Sen A, Suzumura K (eds) Handbook of Social Choice and Welfare, Elsevier, Amsterdam
Brams SJ, Taylor AD (1995) An envy-free cake division protocol. Am Math Mon 102(1):9–18
Brams SJ, Taylor AD (1996) Fair division – from cake-cutting to dispute resolution. Cambridge University Press, Cambridge
Brams SJ, Taylor AD (2000) The win-win solution. Guaranteeing fair shares to everybody. W. W. Norton, New York
Brams SJ, Jones MA, Klamler C (2006) Better ways to cut a cake. Not Am Math Soc 53(11):1314–1321
Brams SJ, Kilgour DM, Zwicker W (1998) The paradox of multiple elections. Soc Choice Welf 15:211–236
Brams SJ, Kilgour M, Sanver R (2007) A minimax procedure for electing committees. Public Choice 3–4(132):401–420
Brandt F, Brill M, Harrenstein P (2016a) Tournament solutions. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge
Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (2016b) Handbook of Computational Social Choice. Cambridge University Press, Cambridge
Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (2016c) Introduction to computational social choice. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge
Brânzei S, Caragiannis I, Kurokawa D, Procaccia AD (2016) An algorithmic framework for strategic fair division. In: AAAI, pp 418–424
Brânzei S, Hosseini H, Miltersen PB (2015) Characterization and computation of equilibria for indivisible goods. In: International symposium on algorithmic game theory. Springer, Berlin, pp 244–255
Budish E (2011) The combinatorial assignment problem : approximate competitive equilibrium from equal incomes. J Polit Econ 119(6)
Caragiannis I, Covey JA, Feldman M, Homan CM, Kaklamanis C, Karanikolas N, Procaccia AD, Rosenschein JS (2009) On the approximability of Dodgson and Young elections. In: ACM-SIAM symposium on discrete algorithms (SODA'09), pp 1058–1067
Caragiannis I, Hemaspaandra E, Hemaspaandra LA (2016a) Dodgson's rule and Young's rule. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge
Caragiannis I, Kurokawa D, Moulin H, Procaccia AD, Shah N, Wang J (2016b) The unreasonable fairness of maximum Nash welfare. In: Proceedings of the 2016 ACM Conference on Economics and Computation. ACM, pp 305–322
Chevaleyre Y, Endriss U, Estivie S, Maudet N (2004) Multiagent resource allocation with k-additive utility functions. In: Proceedings of DIMACS-LAMSADE Workshop on Computer Science and Decision Theory, vol 3, pp 83–100
Chevaleyre Y, Endriss U, Maudet N (2007) Allocating goods on a graph to eliminate envy. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI'07)
Chevaleyre Y, Endriss U, Maudet N (2017) Distributed fair allocation of indivisible goods. Artif Intell 242:1–22
Chevaleyre Y, Lang J, Maudet N, Monnot J (2010) Possible winners when new candidates are added: the case of scoring rules. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI'10)
Chevaleyre Y, Lang J, Maudet N, Ravilly-Abadie G (2009) Compiling the votes of a subelectorate. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'09), pp 97–102

Condorcet N (1735) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Imprimerie Royale, Paris
Conitzer V (2006) Computing Slater rankings using similarities among candidates. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI'06)
Conitzer V, Rognlie M, Xia L (2009) Preference functions that score rankings and maximum likelihood estimation. In: Proceedings of IJCAI-09, pp 109–115
Conitzer V, Sandholm T (2002a) Complexity of manipulating elections with few candidates. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI'98)
Conitzer V, Sandholm T (2002b) Vote elicitation: complexity and strategy-proofness. In: Proceedings of AAAI conference on artificial intelligence (AAAI'98), pp 392–397
Conitzer V, Sandholm T (2005) Communication complexity of common votiong rules. In: Proceedings of ACM conference on electronic commerce (EC'05)
Conitzer V, Walsh T (2016) Barriers to manipulation. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge
Cramton P (2006) Simultaneous ascending auctions. In: Cramton P, Shoham Y, Steinberg R (eds) Combinatorial auctions. MIT Press, Cambridge
Cramton P, Shoham Y, Steinberg R (eds) (2006) Combinatorial auctions. MIT Press, Cambridge
Darmann A, Klamler C, Pferschy U (2009) Maximizing the minimum voter satisfaction on spanning trees. Math Soc Sci 58(2):238–250
Davies J, Katsirelos G, Narodystka N, Walsh T (2011) Complexity of and algorithms for Borda manipulation. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI'11)
de Keijzer B, Bouveret S, Klos T, Zhang Y (2009) On the complexity of efficiency and envy-freeness in fair division of indivisible goods with additive preferences. In: Proceedings of International Conference on Algorithmic Decision Theory (ADT'09)
Dickerson JP, Goldman JR, Karp J, Procaccia AD, Sandholm T (2014) The computational rise and fall of fairness. In: AAAI, vol 14, pp 1405–1411
Dunne PE (2005) Multiagent resource allocation in the presence of externalities. In: Proceedings of International Central and Eastern European Conference on Multi-Agent Systems (CEEMAS'2005), pp 408–417
Dunne PE, Wooldridge M, Laurence M (2005) The complexity of contract negotiation. Artif Intell 164(1–2):23–46
Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the web. In: International World Wide Web conference (WWW10), pp 613–622
Elkind E, Faliszewski P, Slinko AM (2010) Cloning in elections. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI'10)
Elkind E, Slinko A (2016) Rationalizations of voting rules. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge
Endriss U (ed) (2017) Trends in computational social choice. AI Access, to appear
Endriss U, Maudet N (2005) On the communication complexity of multilateral trading: extend report. J Auton Agents Multi-Agent Syst 11(1):91–107
Ephrati E, Rosenschein JS (1993) Multi-agent planning as a dynamic search for social consensus. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'93), pp 423–431
Escoffier B, Gourvès L, Monnot J (2013) Fair solutions for some multiagent optimization problems. Auton Agents Multi-Agent Syst 26(2):184–201
Faliszewski P, Rothe J (2016) Control and bribery in voting. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge
Fischer F, Hudry O, Niedermeier R (2016) Weighted tournament solutions. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge

Fujishima Y, Leyton-Brown K, Shoam Y (1999) Taming the computational complexity of combinatorial auctions: optimal and approximate approaches. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'99)

Galand L, Perny P (2006) Search for compromise solutions in multiobjective state space graphs. In: Proceedings of European Conference on Artificial Intelligence (ECAI'06), pp 93–97

Gibbard A (1973) Manipulation of voting schemes: a general result. Econometrica 41:587–601

Golden B, Perny P (2010) Infinite order Lorenz dominance for fair multiagent optimization. In: Proceedings of International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'10), pp 383–390

Harsanyi JC (1955) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. J Polit Econ 63:309–321

Hemaspaandra E, Hemaspaandra LA, Rothe J (1997) Exact analysis of Dodgson elections: Lewis Carroll's 1876 system is complete for parallel access to NP. J ACM 44(6):806–825

Hudry O (2004a) Computation of median orders: complexity results. In: Proceedings of DIMACS-LAMSADE Workshop on Computer Science and Decision Theory, vol 3, pp 179–214

Hudry O (2004b) A note on banks winners in tournaments are difficult to recognize by Woeginger GJ. Soc Choice Welf 23(1):113–114

Klamler C, Pferschy U (2007) The travelling group problem. Soc Choice Welf 3(29):429–452

Konczak K, Lang J (2005) Voting procedures with incomplete preferences. In: Proceedings of IJCAI'05 Multidisciplinary Workshop on Advances in Preference Handling

Kurokawa D, Procaccia AD, Wang J (2015) When can the maximin share guarantee be guaranteed? Technical report, Carnegie Mellon University

Lacy D, Niou E (2000) A problem with referenda. J Theor Polit 12(1):5–31

Lang J, Xia L (2009) Sequential composition of voting rules in multi-issue domains. Math Soc Sci 57(3):304–324

Lang J, Xia L (2016) Voting in combinatorial domains. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge

Laslier JF, Sanver MR (eds) (2010) Handbook on approval voting. Studies in Choice and Welfare. Springer, Berlin

Lehmann D, Mller R, Sandholm TW (2006) The winner determination problem. In: Cramton P, Shoham Y, Steinberg R (eds) Combinatorial auctions. MIT Press, Cambridge

Lemaître M, Verfaillie G, Bataille N (1999) Exploiting a common property resource under a fairness constraint: a case study. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'99), pp 206–211

Lesca J, Perny P (2010) LP solvable models for multiagent fair allocation problems. In: Proceedings of European Conference on Artificial Intelligence (ECAI'10), pp 387–392

Lipton R, Markakis E, Mossel E, Saberi A (2004) On approximately fair allocations of divisible goods. In: Proceedings of ACM Conference on Electronic Commerce (EC'04)

Luss H (1999) On equitable resource allocation problems: a lexicographic minimax approach. Oper Res 47(3):361–378

May K (1952) A set of independent necessary and sufficient conditions for simple majority decisions. Econometrica 20:680–684

McCabe-Dansted J, Pritchard G, Slinko A (2008) Approximability of Dodgson's rule. Soc Choice Welf 31(2):311–330

Moulin H (1988) Axioms of cooperative decision making. Cambridge University Press, Cambridge

Moulin H (2003) Fair division and collective welfare. MIT Press, Cambridge

Nguyen NT, Nguyen TT, Roos M, Rothe J (2014) Computational complexity and approximability of social welfare optimization in multiagent resource allocation. Auton Agents Multi-Agent Syst 28(2):256–289

Nisan N (2006) Bidding languages for combinatorial auctions. In: Cramton P, Shoham Y, Steinberg R (eds) Combinatorial auctions. MIT Press, Cambridge

Ogryczak W (1997) On the lexicographic minimax approach to location problems. Eur J Oper Res 100:566–585
Procaccia AD (2009) Thou shalt covet thy neighbors cake. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'09), pp 239–244
Procaccia AD (2016) Cake cutting algorithms. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge
Procaccia AD, Rosenschein JS (2007) Junta distributions and the average-case complexity of manipulating elections. J Artif Intell Res 28:157–181
Procaccia AD, Wang J (2014) Fair enough: guaranteeing approximate maximin shares. In: Proceedings of 14th ACM conference on economics and computation (EC'14)
Rassenti S, Smith VL, Bulfin RL (1982) A combinatorial auction mechanisms for airport time slot allocation. Bell J Econ 402–417
Rawls J (1971) A theory of justice. Harvard University Press, Cambridge
Robertson J, Webb W (1998) Cake-cutting algorithms: be fair if you can. AK Peters Ltd, USA
Rothe J, Rothe I (2015) Economics and computation: an introduction to algorithmic game theory, computational social choice, and fair division. Springer, Berlin
Rothe J, Spakowski H, Vogel J (2003) Exact complexity of the winner for Young elections. Theory Comput Syst 36(4):375–386
Rothkopf MH, Pekeč A, Harstad RM (1998) Computationally manageable combinatorial auctions. Manag Sci 44(8):1131–1147
Sandholm TW (1998) Contract types for satisficing task allocation: I. Theoretical results. In: Proceedings of AAAI Spring symposium: satisficing models, pp 68–75
Sandholm TW (1999) An algorithm for optimal winner determination in combinatorial auctions. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'99), pp 452–547
Sandholm TW (2002) Algorithm for optimal winner determination in combinatorial auctions. Artif Intell 134:1–54
Satterthwaite MA (1975) Strategy-proofness and arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. J Econ Theory 10(2):187–217
Sen AK (1970) Collective choice and social welfare. North-Holland, Amsterdam
Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. J Financ 16:8–37
Walsh T (2008) Complexity of terminating preference elicitation. In: Proceedings of International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'08), pp 967–974
Walsh T (2010) Online cake cutting. In: Third international workshop on computational social choice
Woeginger GJ (2003) Banks winners in tournaments are difficult to recognize. Soc Choice Welf 20(3):523–528
Xia L, Conitzer V (2008) Determining possible and necessary winners under common voting rules given partial orders. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI'08), pp 196–201
Xia L, Conitzer V (2010) Compilation complexity of common voting rules. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI'10)
Yager RR (1988) On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans Syst Man Cybern 18:183–190
Young HP (1994) Equity in theory and practice. Princeton University Press, Princeton
Zwicker WS (2016) Introduction to the theory of voting. In: Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (eds) Handbook of Computational Social Choice, Cambridge University Press, Cambridge

Formalization of Cognitive-Agent Systems, Trust, and Emotions

Jonathan Ben-Naim, Dominique Longin and Emiliano Lorini

Abstract A cognitive agent is an agent characterized by properties that are generally attributed to humans. Cognition is viewed here as a general mechanism of reasoning (in contrast with reactive agents) about knowledge. Such agents can perceive their environment, reason about fact or epistemic states of other agents, have a decision making process, etc. This article presents the main concepts used in cognitive agents formalizations, and speak about two particular concepts related to humans: trust and emotion. The language used for cognitive agents is here a logical language because it particularly fits well for both knowledge representation and reasoning formalization. But, even if trust and emotion can be both easily formalized by logical languages, we show that some numerical models are also well adapted.

1 Introduction

To characterize an agent is never easy because a lot of languages can be used, the properties attached to this agent can be various, some concepts may have different names in different contexts, the set of concepts that we need in some context must be different of the set needed in another context, etc. In the following, agents are defined as entities having some properties such as: autonomy (they can act without any human action but only with respect to their internal states); reactivity (they can interact with other –human or artificial– agents by using a communication language, or perform some actions that are needed by the environment); pro-activation (they can adopt a behavior following from their goals by taking the initiative); etc. As it is summarized by Wooldridge (2000), agents are viewed here as computer systems "*deciding for themselves* what to do in any given situation".

J. Ben-Naim (✉) · D. Longin · E. Lorini
IRIT-CNRS, Université Paul Sabatier, Toulouse, France
e-mail: Jonathan.Ben-Naim@irit.fr

D. Longin
e-mail: Dominique.Longin@irit.fr

E. Lorini
e-mail: Emiliano.Lorini@irit.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_19

More specifically, in the area of artificial intelligence (AI), the agents properties are often described by using concepts usually associated to humans such as: mental attitudes (belief, knowledge, goal, desire, intention, etc.); social attitudes (commitment, common belief or common intention, acceptance, etc.), time and action. The properties can also be themselves more specific to humans. We can cite for instance: rationality (in a very wide sense, it means that agents do not act in a contradictory manner: they do not believe both something and its converse, they act with respect to their goals, etc.); sincerity (agents do not aim to communicate something they thinks false), etc. These properties depend on the context where agents evolve. For instance, is it suitable to have a sincere agent playing poker or an insincere agent supposed to report weather forecasting? Certainly not. So, all the properties used by system designers are selected depending on a particular application.

In the following, we call "cognitive agent system" (or "cognitive system" for short) a system which has a behavior predictable only from its mental attitudes. So, the problem is to determine the mental attitudes that are needed to formalize the properties that we want to attribute to the agents of the system. An advantage of such systems is that they can describe everything, even functional objects (cars, locks, etc.). These systems are very popular in AI because they have interesting properties: they are philosophically well-founded, the formal tools are mathematically well defined, the high abstraction level that is used allow to distinguish how something works in the real workd from the general concepts that will used to model it. Finally, these systems have a strong explanatory power (an action mathematically following from both their properties and the agents' mental states that are members of these systems).

In the following, we first speak about cognitive agent systems formalization (Sect. 2). Such an agent is supposed to be able to: represent its physical environment (including the other agents); represent the manner that it wants this environment evolves; reason about these representations in the aim to perform an action.[1] Logic is a tool that fits very well both this formalizing task and this reasoning task, and this section will only present logical tools (more precisely, modal logics), including three types of operators: belief or knowledge (environment representation), desires, goals, preferences, etc. (representation of the wished evolution of this environment), action and time.[2]

Finally, we present two particular concepts strongly related to cognition: trust (Sect. 3) and emotion (Sect. 4). We will focus on the cognitive structure of these concepts, that is, on mental states that are necessary to trust or to trigger an emotion. But logic is less appropriate to the representation of their intensity than numerical models. It explains why there are both logical models and numerical models representing trust and emotion. We will give a short overview of these two approaches.

[1] Note that the word *agent* comes from Latin language *agere* and means to act, to do.

[2] These logics are often called BDI logics (for belief, desire, intention). By analogy, we speak also of BDI agents (systems).

2 Cognitive-Agent Formal Systems

2.1 Short History of BDI Systems

One can say that the story of formal systems as they are today is as long as that of philosophy. Indeed, since Aristotle, philosophy investigated a certain number of concepts: modal logics (logics of necessary and possible), epistemic or doxastic logics (belief and knowledge), deontic logics (obligation, interdiction, permission), temporal, conditional, dynamic logics (explicit or implicit actions), etc.

Our main subject is modal logics, that is, logics including operators that are not truth-functional. So, if $\Box$ is a modal operator, then the formula $\Box\varphi$ (where φ is also a formula of the modal logic) is true independently of the truth-value of φ. This $\Box$ operator can represent beliefs, goals, intentions, etc. For example, if Bel_i *sunny* means that Agent i believes it is sunny, then i can believe it is sunny or not, independently of the weather. (See chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" of the same volume for more details about modal logics.)

All these formal works, as well as certain others, in particular in philosophy (see Searle (1983) and especially Bratman (1987)), have contributed to the construction, between end of 80's and beginning of 90's, of the logic BDI of Cohen and Levesque, where: first, intention is defined, in a non-primitive way, from beliefs and goals (Cohen and Levesque 1990); and second, the formal framework is also used to characterize the capacities of the agents with regard to communication (Cohen et al. 1990). One can say that those works have been the corner stone of cognitive-agent systems.[3] Indeed, it suffices to see theories of agents (in particular, those of the language of the agents) as theories of action.[4]

Those works have been followed by those of Rao and Georgeff who, based on the logical principals adopted by Cohen and Levesque, have looked forward to a more rigorous formal framework in a temporal logic accompanied with a semantics and an axiomatization (Rao and Georgeff 1991). It is worth noting that in those works intention is defined in a non-primitive way. In the same research avenue, we can mention the work of Wooldridge, who introduced the logic LORA (LOgic of Rational Agent) in Wooldridge (2000). The goal of Wooldridge was not only to formalize an agent architecture of the type BDI, but also its evolution in time.

Concerning french work, we can mention the work of Sadek (see his PhD thesis or KR'92), who, in a formal framework of the same family, defined rationality rules in order to guide the behaviour of a rational agent in a system of rational interactions. By the way, his theory has influenced a language of agent communication (agent communication language or ACL) that became an international reference, which

[3]Their paper in *Artificial Intelligence* has received the *AAMAS most influential paper award* in 2008.

[4]This explains by the way the success of the theory of linguistic actions (Austin 1962; Searle 1969) in the agent community: in those theories, the language is seen as the accomplishment of actions, facilitating *de facto* the formal union of physical and linguistic actions.

has been used or gave rise to numerous works in the agent community: the FIPA language.[5]

In the mid 90s, more operational languages appeared, in the sense that the goal is not only to have a logical formalism able to capture the concepts useful to construct the agent systems of interest, but also to implement them. So, BDI systems formalized in situation calculus appeared (see for example the works of Shapiro, Lespérance, and Levesque in Toronto). Programming languages based on primitives of the BDI type also appear: one can mention e.g. GOLOG or ConGolog. This community gave rise to what can be called nowadays cognitive robotics, whose laboratory of the same name in Toronto is the most prominent representative.

Certain formalisms also aims at describing normative systems, that is, systems where the agents have not only to consider what they believe (or know) and what are their goals, but also what they must do. This aspect uses (also also inherits theoretical questions from) deontic logic. For example, we can mention the BOID architecture (where O represent the obligation component of the BDI system) of van der Torre et al. (see e.g. the paper published in AGENTS'01).

Next, BDI systems not only manipulate mental attitudes (in addition to time and/or action), but also social concepts or external constraints. Obligation can be seen as an internal norm (it is then formalized by an operator indexed by an agent or a group of agents), or as an external law every agent must obey (it is then formalized by a non-indexed operator).

By the end of 90's, the BDI systems, as they are then formalized, are heavily criticized, because they are based on strong hypotheses about mental states, in particular sincerity. So, in FIPA for example, an agent believes everything it is told by another agent, because it always assumes the latter tells the truth.

To avoid this problem, certain works describe the effect of a linguistic action by separating what the speaker wants to mean from what the listener believes on the basis of hypotheses made by the latter about the sincerity and competence of the former. Other works looked forward to alternative concepts allowing us to free us from those hypotheses about the internal states of the agents. For example, there are numerous works on social commitment aiming at capturing the public commitment of an agent generated by what that agent says. For instance, when someone says something, he (or she) is committed to the truth-value of that proposition: he could not say he did not said it, and cannot say or do something that opposes what he said (see e.g. the work of Singh (Singh 1999) and de Colombetti in Switzerland). Nevertheless, those approaches also have drawbacks: other hypotheses are made (for example, the public aspect of linguistic actions and the fact that they are correctly interpreted by their targets). In addition, it is not obvious that this concept is devoid of links with the mental states of the committed agents.

Finally, this notion has not been studied in a satisfactory way as a non-primitive concept,[6] despite the fact it apparently contains a normative and conventional component, as well as violation condition. In such circumstances, those approaches are

[5] http://www.fipa.org/repository/aclspecs.html.

[6] That is, a concept constructed from lower-level concepts.

almost not BDI systems, since they do not involve mental states: there is an intuitive link, but it has to be formally established.

Other traditional concept have been confronted to that problem, e.g., common belief. The latter is generally defined as the infinite conjunction of the alternative beliefs between agents. For example, if there is common belief between agents i and j about φ, then i believes φ, j believes φ, i believes j believes φ, j believes i believes φ, i believes j believes i believes φ, etc.

Thus, the problem in an implemented system is to decide whether there is common belief without having access to the minds of the agents. At best, we can construct a subjective notion of common belief, i.e., the fact that an agent believes there is common belief (maybe it is not the case). A number of philosophical works are related to this question (see e.g. Gilbert (1989)). They led to notions like acceptance (see e.g. Lorini et al. (2009).

In parallel, certain prior AI problems have been transferred to the BDI framework and gave rise to a rich literature: the frame problem (how to describe environment in a concise and exhaustive fashion?), the problem of characterizing actions (what are the necessary and sufficient conditions to execute a given action?), the problems of revision (how to have an agent's knowledge evolves with time?) and action ramification (how to describe the impact of an action on the domain, including the mental states of the agents). For example, the advent of BDI systems was followed by the problem of revising mental states (see e.g. van der Hoek et al. (2007).

More recently, this problem has become the heart of a branch of the domain: dynamic epistemic logics (see below). Put simply, the goal is to integrate into the semantics of these logics the fact that the beliefs (or knowledge) of an agent can evolve: that agent can change his mind, learn that certain propositions are true, learn that others are false, etc. At the cost of certain technical constraints, the logics of public announcements give an adequate answer to the hard question of mental-states evolution. For an overview on that subject see e.g. van Ditmarsch et al. (2007).

Finally, agent testbeds have been developed, like e.g. AgentSpeak by Rao, Jason by HÃbner and Bordini, or 2APL by Dastani. Those testbeds allow the implementation of agents and multi-agent systems, but do not yet exhaust all the expressive power of the BDI logics. In particular, they are not equipped with a complete set of boolean operators and do not use theorem provers, which by the way already exist for certain (families of) well-known logics.

Concepts proposed in the domain of BDI systems have also been used in other domains of AI. For example, this is the case of argumentation, where e.g. Amgoud used argumentation methods to generate desires and plans in an autonomous agent (Amgoud and Rahwan 2006) (see also chapter "Argumentation and Inconsistency-tolerant Reasoning" of this volume).

2.2 Basic Concepts

In what follows, we present the basic concepts generally used in the formalization of cognitive-agent systems in terms of mental states. Of course, all systems do not

use all concepts simultaneously, because the way an agent system is characterized depends on the domain of that system.

As soon as we need nested operators, modal logics are particularly adequate, because a formula of a modal logic in the range of a modal operator forms a new formula of that logic. So, we can have an arbitrary large degree of nestedness in the formulas of the object language. This property is particularly important in the domain of cognition, because we can have beliefs about almost anything, including other beliefs: Agent i believes Agent j believes Agent k believes Agent i believes p, etc. (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" on knowledge representation of the same volume).

2.2.1 Belief Operators

The notion of belief has been deeply studied in the domain of doxastic and epistemic logics, since the early 60s (see Gochet and Gribomont (2006) for an exhaustive overview). This is probably one of the most studied notion in Logic, in all its forms (classical logic, modal logic, with or without degrees representing the strength of the beliefs or knowledge of an agent.[7])

A commonly used logic is the propositional modal logic without degrees where "Agent i believes φ is true" is denoted by $Bel_i\,\varphi$, where Bel_i (for every agent i) is called the modal operator of Agent i's beliefs, and where φ is some formula. Traditionally, the fact that $Bel_i\,\varphi$ is true in a certain world w_0 is interpreted as the fact that φ is true in all worlds that are accessible from w_0 according to Agent i. Note that i has no certainty that the real world belongs to this set of epistemic worlds (i may be wrong). To represent this, the semantics includes an accessibility relation for every agent. So, the fact that i believes φ is true in the real world w_0 is denoted by $w_0 \Vdash Bel_i\,\varphi$. Semantically, this means φ is true in all worlds that are accessible from w_0 via the relation corresponding to i and denoted by $\mathscr{B}_i$.

There is a consensus in the literature that the logic of beliefs in the normal modal system KD45 (Chellas 1980), even though this logic constitutes an idealization of certain principles. For example, this logic assumes an agent instantly knows all beliefs implied by its own (omniscience) and it is conscious of all those beliefs (positive introspection). Nevertheless, those criticisms are mitigated by the fact that they constitute idealizations (not aberrations), which are not necessarily counter-intuitive for an artificial agent.

Figure 1 shows the semantics of the belief operator of agent i. The set of all worlds that are accessible from w_0 is denoted by $\mathscr{B}_i(w_0)$, where $\mathscr{B}_i$ is the accessibility relation between worlds for Agent i and is graphically represented by arrows.

[7] In the present work, we only consider qualitative approaches to the notion of belief. We do not discuss the quantitative approaches formalizing degrees of belief (see e.g. (Laverny and Lang 2005)).

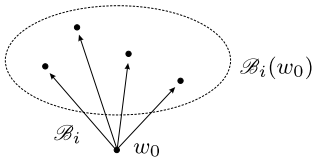


Fig. 1 Kripke semantics of the operator Bel_i

2.2.2 Temporal Operators

There are many temporal logics, depending on the way one wants to represent time (ramified or linear, with or without explicit temporal indexes, etc.). Temporal logics are relatively well-studied in the domain of modal logics and theoretical computer sciences (van Benthem 1991). There semantics is based on transition relations between possible states and are thus equivalent to (potentially infinite) automates (see chapter "Qualitative Reasoning" of the same volume for more detail about temporal reasoning).

Here, we focus on a very simple notion: linear time. Since this notion is combined with the beliefs of the agents, this means that the latter are not about epistemic worlds, but about linearly-ordered sets of worlds called "stories". This allows us to simulate a tree-based nature of time, since each story corresponds to a development of future events (the agent believes possible).

For example, Fig. 2 represents the four stories believed by Agent i. The dots on the stories represent the present moment and the dashes the past and future moments. So, the agent right now believes that p is true ($Bel_i\ p$); it consider the possibility that r is right now true but becomes false thereafter ($\neg Bel_i\ \neg(r \wedge F\neg r)$); etc.

We can define the operators H and P (about the past) in the same way we defined the operators G and F.

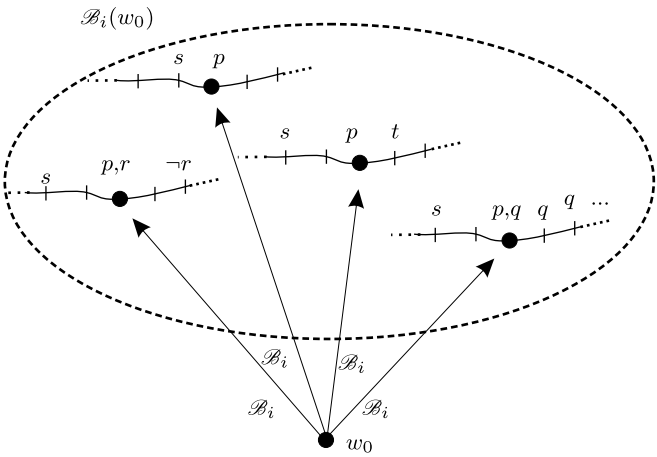


Fig. 2 Linear time and epistemic worlds

Technically, time is defined in a modal logic of linear time of Type $S4.3_t$ (see Burgess (2002) for more details). Nevertheless, those operators can be semantically defined with a tree-based structure (which is by the way what is done in Rao and Georgeff (1991)).

Finally, we sometimes use the two operators X and X^{-1} such that $X\varphi$ means "φ will be true right after the present moment is the considered story" and $X^{-1}\varphi$ means "φ was true right before the present moment in the considered story". Obviously, there exists formal links between those operators and the temporal ones defined previously.

2.2.3 Goal Operators

The notion of goal has been widely studied in the literature and has been used in very different senses (see e.g. the notion of goal in Cohen and Levesque (Cohen and Levesque 1990) or Rao and Georgeff (Rao and Georgeff 1991), the notion of choice in the PhD thesis of Sadek or KR'92). We focus on the notion of *chosen goal* (or *preferred goal*), with regard to the coherent subset of proposition the agent wants to make true. The primitive operators of goal are denoted by $Choice_i$ (where i ranges over all agents) and $Choice_i\ \varphi$ means that "Agent i right now choose to make the goal φ right now true". There is no restriction on the formula φ, so it can represent the present state of affairs. This is the difference with the operators of goals to be achieved (abandoned when the desired state of affairs comes true) or to be maintained (an agent looks forwards to keep a certain state of affairs true). As we did it with beliefs, we interpret $Choice_i\ \varphi$ in a world w_0 as the fact that φ is true in all the preferred world of the agent from w_0. Most generally, goals are partial pre-order, but, for the sake of simplicity, we do not consider this point: we focus on coherent non-ordered binary goals.

A difficult and non-studied question is the following: how those goals emerge? From a cognitive point of view, it looks like they emerge from a deliberative process about more primitive attitudes: desires, ideals, and imperatives (see Rao and Georgeff (1991); Conte and Castelfranchi (1995); Castelfranchi and Paglieri (2007)). The set of goals we characterize is the one obtained from a process of selection of ideals and desires. It is meant to resolve conflicts between those two concept and to eliminate impossible cases. Then, the chosen goals of an agent satisfy the two following fundamental rationality principles: they are consistent (an agent cannot choose two contradictory goals); the chosen goals are related to the beliefs of the agent that chose them. In Cohen and Levesque (1990), the relation between beliefs and goals is an inclusion relation: if an agent right now believes φ is true, then it necessarily right now has φ as a goal (this notion is called *strong realism*). We can also impose a relation of *weak realism*, where it is only required that there is a non-empty intersection between the epistemic worlds that are possible and those that are preferred.

Some recent works aim to explain the goals building process by the way of desires. Desires and goals are often combined (see for instance Dubois et al. (2017).

2.2.4 Ideals

There exist many normative systems in logic with very different characteristics, more or less complex, adapted to a class of problems or another. Those norms may have different origins: state laws, institution rules, moral (be it religious or not), etc.

Certain particular norms, specific to a given agent, are called ideals. We introduce a new set of operators such that $Idl_i\ \varphi$ means: "φ is an ideal state for Agent i". This means that i gives an order to itself, a kind of "must make true" for φ (when φ is false at the present moment) or "must keep true" (when φ is already true) (Castaneda 1975).

There are different ways to explain how a state φ becomes an ideal state for a certain agent. A possible explanation is that ideals are just social norms that have been internalized (or adopted) by this agent (Conte and Castelfranchi 1995). Assume an agent believes in a certain group (or institution) there is a certain norm (e.g. an obligation) saying that a state φ must be true, whilst the agent sees itself as a member of that group. In such a case, the agent adopts this external norm (that does not originate from the agent and has not yet been acknowledge as a norm by the agent) and that norm becomes an ideal for that agent. For example, if Agent i believes in France, it is obligatory to pay taxes and that agent considers himself (or herself) as a French citizen, then he adopts this obligation and pays his taxes.

Semantically, the ideals are represented from the possible worlds considered as ideal by the agent having internalized those ideals. There is no particular relation with the other operators, besides belief, if we assume an agent is conscious of its ideals (see chapter "Norms and Deontic Logic" of the same volume for more details about normative operators). (See also Gabby et al. (2013) pour for more details about normative and deontic systems and Berreby et al. (2015); Lorini (2016) about moral systems.)

2.2.5 Explicit Action

When one tries to define "Agent i is capable of executing Action φ", one has to consider logics of action (see chapter "Reasoning about Action and Change" of the same volume on reasoning about action and change). Generally speaking, those logics formalize actions with state-transition systems. There are essentially to schools of thought, one where action is explicit and one where it is implicit (see the next section).

The main logic of explicit action is propositional dynamic logic (PDL), which studies the interact between an action and its effects (Harel et al. 2000). It has been shown (e.g. in van Linder et al. (1998) that dynamic logic is particularly adapted to the characterization of the concepts of capacity and power. There is a rich literature on the integration of dynamic logic into logics of beliefs and goals (see e.g. epistemic dynamic logic Baltag and Moss (2004) or doxastic dynamic logic Segerberg (1992, 1995)).

PDL distinguishes between actions like α and formulas like φ and ψ, and its set of non-logical constants is constructed from those two categories. The formula $After_\alpha\ \phi$

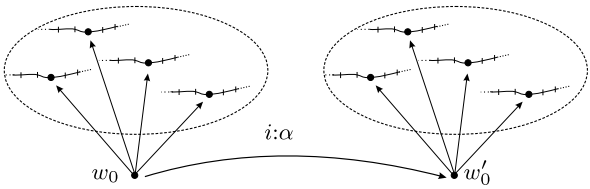


Fig. 3 Transition from the world w_0 to the world w'_0 via the execution of the action $i{:}\alpha$

expresses the fact that φ will be true after any possible execution of Action α. So, $After_\alpha \perp$ means α is not executable.[8]

Several extensions have been proposed where an agent is added to the arguments of the PDL operators. In such extensions, the formula $After_{i:\alpha}\ \phi$ means that φ is true after any possible execution of Action α by Agent i. For any action α and agent i, $After_{i:\alpha}$ is an action modal operator.

Semantically, action is treated as a transition from a real world to a set of other real worlds (certain semantic constraints can force this set of worlds to be a singleton). Figure 3 represents this transition.

In DEON'2008, Lorini and Demolombe have augmented the PDL language with the operators $Does_{i:\alpha}$, where $Does_{i:\alpha}\ \phi$ means "Agent i is about to execute Action α and thereafter φ will be true". This allows us to speak about what an agent can do ($\neg After_{i:\alpha} \perp$) and what an agent does ($Does_{i:\alpha} \top$).

2.2.6 Implicit Action

Action is implicit in logics of agency, which study the interaction between an agent and the effects caused by it. The peculiarity of those logics is that they do not represent the actions that caused the effects (only results matter).

For example, in the logic *STIT* (Belnap et al. 2001), actions are formalized by formulas involving an agent and speaking about the effects caused by that agent. So, the action described in "i buys the product p" is formalized by the following formula of agency: "i sees to it that Product p is bought by Agent i".

Formulas of agency are of the form $STIT_i\ \phi$, which means "The action chosen by Agent i at the present moment ensures that φ is true, independently of what the other agents do". In short, "i sees to it that φ". The modal operator $STIT_i$ is called the operator of agency.

2.2.7 Dynamic of Mental States

Last years, a certain number of researchers working in the domain of logics for autonomous-agent formalization and in multi-agent systems have proposed logics for the dynamic of mental states. They belong to the large family of dynamic epistemic

[8] Besides BDI logics, the operator $After_\alpha$ is often denoted by $[\alpha]$.

logics (DEL), see e.g. Ditmarsch et al. (2007). DEL is a term used in a very large sense to include dynamic extensions of logics of belief and knowledge, but also logics of preferences and norms (deontic logics) (Baltag and Moss 2004; Kooi 2007; van Benthem and Liu 2007). In those logics, modal operators are introduced to describe the effects, on the mental states of the agents, of various types of informative events (transmission of public or private messages, orders, etc.).

Here, we consider the most known dynamic epistemic logic, namely public announcement logic (PAL) (Ditmarsch et al. 2007). Informally, a fact p is publicly announced if and only if: every agent learns that p is true; every agent learns that every agent learns that p is true; every agent learns that every agent learns that every agent learns that p is true, etc., up to infinity. In the PAL logic, public announcements are events that update the beliefs and knowledge of the agent: the role of a public announcement is, first, to reduce the set of possible worlds to the worlds where the publicly announced fact holds, and second, to restrict the epistemic accessibility relations to those worlds. PAL uses the notation $p!$ for the public announcement of p, and introduce modal operators of the form $[p!]$ to describe the effect of a public announcement on the mental states of the agents: the formula $[p!]q$ means that q will be true after the public announcement of p. We take below an example to illustrate those dynamic operators.

Marie, Paul, and Alice are seated around a table on which are laid three cards. The cards are face down, but, on every card, is written a distinct number between 1 and 3. So, the cards can be called Card 1, Card 2, and Card 3. Marie, Paul, and Alice take one card, each. We assume Marie took the card 1 (denoted by m_1), Paul the card 2 (denoted by p_2), and Alice the card 3 (a_3). Each player confidentially looks at his (or her) card and put it of the table face down. Therefore, each player only knows the number written on his card.

In Fig. 4, the model on the left represents the beliefs of Marie, Paul, and Alice in the initial situation. There are 6 possible worlds and the one in grey is the real one. The arrows represent the accessibility relations $\mathscr{B}$ between epistemic worlds, for each player. For example, in the real world, Marie considers as possible the world where Marie has Card 1, Paul Card 2, and Alice Card 3, as well as the world where Marie has Card 1, Paul Card 3, and Alice Card 2. So, in the real world, Marie has no certainty about the card distribution.

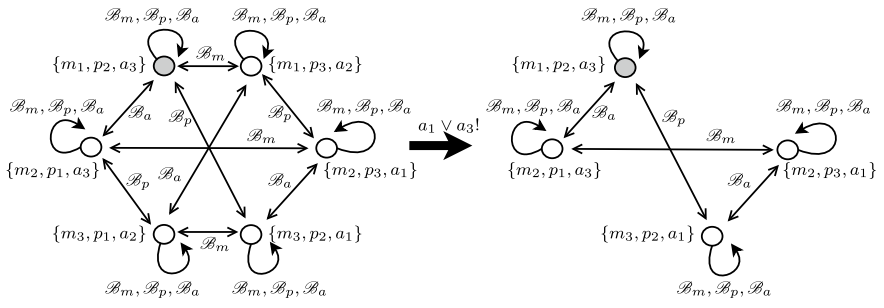


Fig. 4 Example of Cards

Assume it is publicly announced that Alice has a card with an odd number. This announcement is represented by the event $a_1 \vee a_3!$ (Alice has Card 1 or Card 3). In Fig. 4, the model to the right of the arrow represent the beliefs of Marie, Paul, and Alice after this announcement. Thanks to the latter, Marie learns that Paul has Card 2 and Alice Card 3. Indeed, the effect of the public announcement is to reduce the set of possible worlds to those where Alice has an odd card and to restrict the accessibility relations to those worlds. So, in the real world, after the public announcement, Marie knows the card distribution: Marie has Card 1, Paul Card 2, and Alice Card 3. This fact is represented by the formula $m_1 \wedge p_2 \wedge a_3 \wedge Bel_m\,(m_1 \wedge p_2 \wedge a_3)$, which is true in the real world of the model on the right. In contrast, the public announcement does not make Paul and Alice learn anything: after the public announcement they still have no certainty about the card distribution.

Up to know, we gave an overview of the concepts related to cognitive-agent systems and a way to formalize them. In what follows, we present two particular complex concepts that can be described in terms of mental states, time, and action. Cognitive-agent systems are thus very adapted to the formalization of these two concepts. Nevertheless, the latter are also formalized in more numerical ways and, in what follows, we give an overview of this.

3 Formalization of Trust

Trust systems (or trust models) are used in certain multi-agent systems to help users to choose the agents to interact with. Indeed, agents may be incompetent or malicious. But, the agents are typically so numerous that it is impossible for a central authority to test them all. Consequently, the goal of a trust system is to evaluate the agents on the basis of relations between them. More precisely, for a user u, the evaluation of the peers of u is based on two kinds of information:

- the result of past interactions between u and the other agents;
- the feedbacks other agents have provided about their peers.

The value (a score, a position in a ranking, etc.) of an agent a can naturally be seen as the trust of u in a.

Trust systems can be motivated by several large-scale applications where no central authority can test all agents. As examples, we can mention: e-commerce (Ebay, Amazon, etc.), large wikis (Wikipedia, Planetmath, etc.), social networks (Facebook, Tweeter, etc.), webpages and hypertext links, papers and citations.

Various trust systems have been developed. To validate and compare them two kinds of approaches are possible: a theoretical and an experimental one. The first approach consists in establishing desirable properties (or axiom, postulates) that a trust system could satisfy. The second approach consists in developing a testbed where different trust systems can compete.

As far as we know, there are two kinds of trust systems: logic-based systems (essentially modal-logic-based systems) and numeric systems. Two position papers

that cover a large number of models are for example (Sabater and Sierra 2005) and, more recently, (Pinyol and Sabater-Mir 2013).

3.1 Logic-based Trust Models

In the logical approach, the goal is to characterize the notion of trust in a certain formal language. Similarly, the objective is to formalize in such a language the notion of trusting someone, as well as the mental state of an agent trusting someone.

One of the main models of trust is the cognitive one from Castelfranchi et Falcone (denoted by C&F) (Castelfranchi and Tan 2001). Contrary approaches that are more computational, the C&F model is more than subjective probabilities updated in the light of direct interactions with the *trustee* (the agent to be trusted) and feedbacks from interactions between the trustee and other agents.

Informally, the C&F model defines trust as an personal belief of the *truster* (the agent that has to decide whether or not to trust the trustee) that the trustee is reliable with regards to various aspects (capacity, intention, readiness, etc).

According to C&F and the analysis conducted in Herzig et al. (2010), the notion of trust is based on four components: a truster i, a trustee j, an action α of j, and a goal φ of i. According to their definition, "i trusts j that j will perform α in order to achieve φ" if and only if: φ is a goal of i; i believes j is capable of performing φ; i believes that performing φ will makes ϕ true; and i believes that j intends to do α.

For example, assume i trusts j to send a certain product p in order to possess p. Then: possessing p is a goal of i; i believes j is capable of sending p; i believes sending p will make him (or her) possessing p; and i believes j intends to send p.

In other words, trust is formally defined as follows:

$$Trust(i, j, \alpha, \varphi) \stackrel{def}{=} Goal_i\, \varphi \wedge Bel_i\, (Capable_j(\alpha) \wedge After_{j:\alpha}\, \varphi \wedge Intend_j\, (\alpha))$$

where every operator used above is either a basic one or a compound one defined from the basic ones (*cf.* Sect. 2.2):

- $Goal_i\, \varphi \stackrel{def}{=} Choice_i\, F\phi$ means "Agent i chooses to make $F\phi$ true at the present time";
- $Capable_j(\alpha) \stackrel{def}{=} \neg After_{j:\alpha}\, \bot$ means "Agent j is capable of executing Action α if and only if α is already executable";[9]
- $Intend_j\, (\alpha) \stackrel{def}{=} Choice_j\, Does_{j:\alpha}\, \top$ means "Agent i intends to execute Action α if and only if executing α (right here, right now) is a chosen goal of i".

[9] One could think that this should be a sufficient but not necessary condition. Indeed, it suffices that Agent i believes Agent j will be capable of executing Action α in time to achieve Goal φ. Nevertheless, it is worth noting that we formalize a notion of trust "right here, right now", not a notion of potential trust.

A relatively recent paper allowing an agent to reason about its trust model, by providing a method for incorporating a computational trust model into the cognitive architecture of the agent is Koster et al. (2013).

We turn to approaches where the notion of trust is not based on modal logic, but more numeric objects.

3.2 Numerical Models of Trust

Previously, trust was seen essentially as a particular belief of the truster about certain aspects of the trustee. Depending on whether i trusts j or not about a proposition φ, i was in position to decide whether or not to believe what j says about φ.

The situation is similar with numeric approaches. The first question is to decide how to represent trust in a numeric way. Various solutions have been proposed, for example, trust can be represented by a number, an interval, or a fuzzy interval.

First, trust can be represented by a simple number. One of the first approaches of this kind is Marsh (1994). Another important approach is that of *Pagerank* (Page et al. 1998), the system at the basis of the well-known Google search engine. More precisely, a webpage can be seen as an agent and a hypertext link from x to y as a positive feedback. Pagerank associates every agent with a real number between 0 and 1 on the basis of these feedbacks. Theses numbers can be seen as the degrees of trustworthiness of the agents.

It is worth noting that Pagerank evaluates the trustworthiness of an agent for an external user. Most trust systems evaluates the trustworthiness of an agent for another agent x. In such a case feedbacks from direct interactions with x are obviously more important than feedbacks from interactions where x is not involved.

A relatively exhaustive study of questions related to Pagerank and its alternatives can be found in e.g. Langville and Meyer (2005). A version of Pagerank adapted to peer-to-peer systems as been constructed in Kamvar et al. (2003).

In certain approaches, an agent is either trustworthy or not, and a model can associate an agent x with a number indicating the probability that x is trustworthy. In other approaches, a model can associate an agent x with a number indicating the degree of trustworthiness of x. In other words, depending on the model, the same number x is associated with can mean different things. For example, assume x is associated with 0.5. It can mean that x perfectly achieves one goal out of two, as well as x achieves every goal half-successfully.

Concerning links between trust and other important notions, it is described in e.g. Osman et al. (2015) how trust models can be used to distinguish between good and bad advices. Finally, a paper describing how the notion of trust can be integrated with those of negotiation and argumentation is e.g. Bonatti et al. (2014).

3.3 Applications of Trust Systems

We present six examples of multi-agent systems where a user (be it an external entity or an internal agent) needs an evaluation of the trustworthiness of the agents:

E-commerce (Ebay, Amazon, …).

The agents are the buyers and sellers. A user has to choose the agents to make transactions with. But they are numerous, generally unknown to him (or her), far from him, and some agents are malicious or incompetent. So, the user needs an evaluation of the agents. After each transaction, the buyer can rate the seller, and vice versa. So, a trust system can exploit these ratings to construct an evaluation. We can globally admit that the more an agent is trustworthy, the more he tends to provide honest and accurate feedbacks. The same goes for the buyers. So, in case of cycles the trustworthiness of an agent x depends on that of an agent y, and vice versa, which makes the evaluation hard to construct.

Large wikis (Wikipedia, Planetmath, …).

The agents are the contributors of the wiki, that is, those that create, delete, or modify articles. A user has to choose to trust or not the contributions and thus needs an evaluation of the contributors. It is easy to imagine how to modify a wiki so the contributors can provide opinions about their peers, in particular when they participate in long debates about controversial issues. A trust system could exploit these opinions to construct an evaluation. We can admit that the more an agent provides serious contributions, the more he (or she) tends to provide serious opinions about its peers. So, again opinion cycles constitute a difficulty.

Social networks (Facebook, Myspace, …).

The agents are the persons, applications, etc. registered in the network. A user has to choose the agents to establish a formal link with. Such a link gives access to all sorts of personal information about the user. But, some persons or applications are malicious. A friendship link between a and b can be seen as the fact that a recommends b as an honest agent, and vice versa. Those links can exploited to evaluate trustworthiness. There are recommendation cycles and the more an agent is honest, the more it provides honest recommendations.

Web pages and hypertext links.

The agents are the web pages. A hypertext link from a page a to a page b can be seen as a recommendation, that is, as the fact that a provides an opinion that the content of b is important. There are cycles and the more a page a has an important content, the more the hypertext links contained in a are important.

Papers and citations.

An agent is a paper or an author. A citation relation from a paper x to a paper y can be seen as the fact that x supports y. Similarly, an authorship relation between a paper

x and an author a can be seen as a support relation for a. The are no relation cycles, but the more a paper x is trustworthy, the more the citation or authorship relations coming from x are important.

Entity-key bindings and certificates.

In the systems based on public key certificates, there are entities willing to send messages to other entities. Since an entity can listen messages that are not intended for it, they are encrypted and decrypted with keys. So, the system generates a set of keys such that there exists a function f transforming any key K into a key $f(K)$ such that the following holds:

(a) $f(K)$ is the unique key that can decrypt the messages encrypted with K, and it can decrypt only these messages;
(b) the converse is true, that is, K is the unique key that can decrypt the messages encrypted with $f(K)$, and it can decrypt only these messages.

Next, a set of bindings is published. A binding is a pair $\langle E, K \rangle$ where E is an entity and K a key. Such a binding represents a claim that E is the unique entity that knows $f(K)$. If it is indeed the case, then we say that $\langle E, K \rangle$ is valid. So, to send a confidential message to an entity, it suffices to find a binding containing it, and use the corresponding key. By (a), only this entity will be able to decrypt the message. The problem is that a malicious entity F can publish a false binding $\langle E, K \rangle$. In other words, E does not know $f(K)$, but F does. So, if this false binding is used, then F can listen some messages intended for E and decrypt them.

To counter this, a set of public key certificates is published. A certificate is a pair $\langle D, S \rangle$, where D is a quadruplet of the form $\langle E, K, E', K' \rangle$ and S is a digital signature, that is, S is supposed to be the result of encrypting D with $f(K)$. Such a certificate represents a claim that E supports the validity of $\langle E', K' \rangle$. Again, the problem is that false certificates can be published. However, it is possible to formally check that the certificate $\langle D, S \rangle$ was created by an entity knowing $f(K)$. By (b), it suffices to decrypt S with K and then check that the result is indeed equal to D. Only the certificates that pass this test are considered.

Now, we can explain the link with trust systems. An agent is a binding $\langle E, K \rangle$. A user is an entity E that has to choose valid bindings before sending messages. A certificate $\langle \langle E, K, E', K' \rangle, S \rangle$ can be seen as the fact that $\langle E, K \rangle$ supports the validity of $\langle E', K' \rangle$. These support links can be exploited to evaluate the validity of the bindings. Finally, the problem of evaluating the validity (or trustworthiness) of the bindings is difficult in particular because there are cycles of support links and the following holds: if a binding $\langle E, K \rangle$ is valid, then E is the unique entity knowing $f(K)$, thus the certificate $\langle \langle E, K, E', K' \rangle, S \rangle$ was created by E, i.e., this certificate is authentic, so we should attach more importance to it.

4 Formalization of Emotions

There is a rich literature on emotions, be it in philosophy [10] (Gordon 1987), psychology (Lazarus 1991; Ortony et al. 1988), economy (Loewenstein 2000), or cognitive sciences (Lane and Nadel 2000).

In computer sciences, emotions play an important role in multi-agent systems at different levels. Much work focus on the modelization of facial and gestural results of emotions with animated conversational agents (ACA) (see e.g., Gratch and Marsella (2005); Pelachaud (2009)). ACA also use models of emotions to represent those of the users, to show their affective states, or a particular personality.

The goal is to make such agents so realistic that users have the impression to interact with other humans. First, this goal assumes a great realism in the expressive aspects of the agents (facial and corporal movements, intonations, verbal expressions, etc.). Second, it is necessary, for the agents, to be able to recognize and take into consideration user's emotions in their of reasoning (as well as their artificial own). So, agents can speak and act in a most adequate fashion.

Emotions are fundamental to have natural and optimal interactions between agents and users, because nowadays it is known that we constantly communicate information about our emotional states (be them real or not) without explicating them. For example, a "Hello!" accompanied with a smile constitutes a common and short way to express your greetings to someone and to tell him (or her) you are happy to see him (which could be explicated by "Hello, I am happy to see you").

4.1 Logical Formalization of Emotions

Concerning formal models of emotions, we look forward to construct logical frameworks in order to formalize certain specific emotions, their properties, the links between them, etc. (see e.g., Adam et al. (2009); Turrini et al. (2010)). The main objective is to take advantage of logical methods to rigorously specify how to implement emotions into an artificial agent. The design of systems containing such agents (capable of reasoning and expressing certain emotions) can benefit from the fact that logic is a tool particularly adapted to the notion reasoning and forcing the designer to disambiguate the different dimensions of emotions (identified in different psychological models of emotions).

Generally, logical definitions of emotions characterize cognitive structures of emotions, rather than emotions themselves. According the theories of cognitive evaluation (Lazarus 1991), the cognitive structure of an emotion is the configuration of the mental state of an agent when it (artificially or not) feels that emotion. The cognitive structure is just a part of the affective phenomenon. In the sequel, we use the word "emotion" for "the cognitive structure of an emotion".

[10] Plato clearly establishes a distinction between reason, passion, and desire.

We distinguish between simple emotions and what we call complex emotions (Adam et al. 2011; Lorini and Schwarzentruber 2011). The former are those that can be described only with mental attitudes like beliefs, goals, or ideals. The latter are those requiring more complex reasonings like counterfactual conditionals: "I could have made φ true, whilst it is actually false". In that sense, complex emotions are associated with counterfactual reasonings about norms, responsibilities.

For example, the fact that agent i feels joy about Fact φ may be expressed as follows:

$$Joy_i\ \varphi \stackrel{def}{=} Bel_i\ \varphi \wedge Choice_i\ \varphi$$

According to this definition, Agent i feels joy about φ if and only if i believes φ is true and wish φ to be true. For example, Tom feels joy about a certain test, because he thinks he successfully passed it and it is what he wishes. So, Tom is happy because he believes the state of affairs is as he wishes. Joy has a positive valence, that is, when it is felt, it is associated to a state of affairs corresponding to desires. This is not the case of sadness for example whose state of affairs does not correspond to desires.

Concerning complex emotions, we restrict ourselves to those related to the notion of responsibility (be it that of the agent feeling the emotion or another one). The responsibility of Agent i for the fact that φ is true can be defined as follows: φ is true and i could have made φ false. More formally:

$$\mathbf{Resp}_i\varphi \stackrel{def}{=} \phi \wedge \mathbf{Cd}_i\neg\phi$$

Here $\mathbf{Cd}_i$ (i could have made) is a basic operator of the formal language, but can be defined from the implicit action operator STIT (for more details, see Lorini and Schwarzentruber (2011).)

So, when Agent i is responsible for the fact that φ is true, whilst i has $\neg\varphi$ as goal, i feels regret (see e.g., Zeelenberg et al. (1998)). More formally:

$$Regret_i\varphi \stackrel{def}{=} Goal_i\ \neg\varphi \wedge Bel_i\ \mathbf{Resp}_i\varphi.$$

Other emotions can be defined in the same way. Emotions constitute a growing domain, because computer science does not have yet exhausted all their possibilities. Existing and implemented systems can often be reduced to simple labels that can be activated or deactivated. Formal models based on logic force designers to explicate the nature of emotions and thus to better understand them.

4.2 Numerical Models of Emotions

There exist numerical models of emotions that study the quantitative aspects of those affective phenomena. For example, (El-Nasr et al. 2000) proposed a numerical model of emotions called FLAME (Fuzzy Logic Adaptive Model of Emotions) based on

fuzzy logic. The main contribution of this work is a quantification of the intensity of emotions, from appraisal variables like desirability or probability. For example, based on the psychological model of emotions of Ortony, Clore and Collins (Ortony et al. 1988), in the model FLAME, the intensity of hope with regard to a certain event depends on the degree of desirability of that event and its subjective probability. More recently, several researchers in AI have augmented formal models of emotions with quantitative aspects. For example, Meyer et al. (Steunebrink et al. 2008) proposed a model describing how the intensity of emotions decreases with time. Lorini (Lorini 2011) proposed a systematic study of the intensity of emotions on the basis of expectations (hope, fear, disappointment, relief) and the relation between those emotions and the mechanism of belief revision of a cognitive agent.

There also exist numerical models of emotion where the latter is represented by a vector whose numbers correspond to components of emotion. For example, Mehrabian captures mood by a vector representing pleasure, excitation, and dominance (i.e., the capacity of an individual to dominate a stimuli). In other words, mood depends on the values of those three components. We can also mention works on the robot with human-like head WE-4R constructed in the university of Waseda (Japan) by Hiroyasu Miwa and his team. The model of emotion is a space-oriented vector calculated from three components: pleasure, activation, and determination.

4.3 Applications of Emotion Models

Concerning applications, teaching systems have been developed to deal with emotions and thus increase the degree of perseverance and commitment of the students. In parallel, simulators, video games, and ambient-intelligence systems have been developed (see e.g., Adam et al. (2011) for an overview of the literature and applications of emotions in that domain). Among the very large variety of existing ACA, EM[11] is a typical system that simulates the decline of emotions with time for a specific set of emotions corresponding to the goals that generated them. Another example is the system Affective Reasoner of Gratch and Marsella where agents use representations of themselves and others. Finally, GRETA (de Rosis et al. 2003) is an ACA 3D that can be animated in real time and is capable of expressing emotional states.

5 Conclusion

In the present chapter, we have first tackled the formalization of cognitive-agent systems. Such an agent is capable of behaving in an autonomous way, according to its goals. In addition, it is characterized, a minima, by mental attitudes (beliefs, desires, norms, etc.), time, and action. After a brief overview of the great research

[11] It is a system based on the Tok architecture of the project Oz. See http://www.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/.

avenue in this domain, we have presented the fundamental concept of BDI systems, as well as the tools to deal with the well-known AI problem of knowledge evolution. Finally, we used the aforementioned material in order to formalize two concepts used in those systems: trust and emotion. We also showed that those two concepts can also be formalized in a more numerical fashion, which is less fine from the point of view of the definitions of the concepts of interest, but easier to be applied in concrete frameworks.

Of course, there are many other branches in AI about the formalization of cognitive-agent systems. But, some of them are not based on mental states, other are limited to a certain formal language. The peculiarity of the systems presented in this chapter is that they correspond to logic (with both a semantics and an axiomatization) whose properties (in terms of complexity, decidability, and completeness) are also studied. More precisely, those logics are modal logics particularly convenient to represent mental states, as well as relations between those states (beliefs about beliefs, goals, etc. of other agents. The objective is to represent in a fine grain the concepts used by the agents with a logic having "good" logical properties. So, the issues are both computational and mathematical. In addition, they are strongly related to SHS via philosophy and psychology, in particular. It is worth noting that there are studies about the influence of trust on emotions, and vice versa (see e.g., Bonnefon et al. (2009)).

Naturally, trust and emotion are not the only concepts investigated in the literature. In particular, we have not presented non-reductionist social concepts, for example, notions of group belief or acceptance that are reducible to the sum, over all agents of the group, of their beliefs or acceptance. Consequently, it is necessary to capture a group as a unique entity constituting an institution ruled by specific social rules.

The study of formal properties of intelligent agents is thus a first step in the study of multi-agent systems. The latter need to capture the nature of the group constituted by the agents (What unites them? What is the structure of the group represented by them? Is it just a set of agents or a more complex relational structure including e.g. friendship, hierarchy, commerce, etc.?).

References

Adam C, Herzig A, Longin D (2009) A logical formalization of the OCC theory of emotions. Synthese 168(2):201–248. ftp://ftp.irit.fr/IRIT/LILAC/Journaux_internationaux/2009_Adam_et_al_Synthese.pdf

Adam C, Gaudou B, Longin D, Lorini E (2011) Logical modeling of emotions for ambient intelligence. In: Mastrogiacomo F, Chong NY (eds) Handbook of research on ambient intelligence and smart environments: trends and perspectives, IGI Global

Amgoud L, Rahwan I (2006) An argumentation-based approach for practical reasoning. In: Weiss G, Stone P (eds) Proceedings of the international joint conference on autonomous agents and multiagent systems (aAMAS 2006), ACM, pp 347–354

Austin JL (1962) How to do things with words. Oxford University, Oxford

Baltag A, Moss LS (2004) Logics for epistemic programs. Synthese 139(2):165–224

Belnap N, Perloff M, Xu M (2001) Facing the future: agents and choices in our indeterminist world. Oxford University, New York
van Benthem J, Liu F (2007) Dynamic logic of preference upgrade. J Appl Non-Class Log 17(2):157–182
Berreby F, Bourgne G J-G G (2015) Modelling moral reasoning and ethical responsibility with logical programming. In: Logic for programming, artificial intelligence, and reasoning, LNCS, vol 9450, Springer, Berlin, pp 532–548
Bonatti PA, Oliveira EC, Sabater-Mir J, Sierra C, Toni F (2014) On the integration of trust with negotiation, argumentation and semantics. UKnowl Eng Rev 29(1):31–50. https://doi.org/10.1017/S0269888913000064
Bonnefon JF, Longin D, Nguyen MH (2009) A logical framework for trust-related emotions. Electron Commun EASST, Form Methods Interact Syst 22:1–16
Bratman M (1987) Intentions, plans, and practical reason. Harvard University, Cambridge
Burgess JP (2002) Basic tense logic. In: Gabbay D, Guenthner F (eds) Handbook of philosophical logic, vol 7, 2nd edn. Kluwer, pp 1–42
Castaneda HN (1975) Thinking and Doing. D. Reidel, Dordrecht
Castelfranchi C, Paglieri F (2007) The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. Synthese 155:237–263
Castelfranchi C, Tan YH (eds) (2001) Trust and deception in virtual societies. Kluwer Academic Publishers, Dordrecht
Chellas BF (1980) Modal logic: an introduction. Cambridge
Cohen PR, Levesque HJ (1990) Intention is choice with commitment. Artif Intell J 42(2–3):213–261
Cohen PR, Morgan J, Pollack ME (eds) (1990) Intentions in communication. MIT, Cambridge
Conte R, Castelfranchi C (1995) Cognitive and social action. London University College of London, London
van Ditmarsch H, van der Hoek W, Kooi B (2007) Dynamic epistemic logic. Kluwer Academic Publishers
Ditmarsch Hv, der Hoek Wv, Kooi B (2007) Dynamic epistemic logic. Kluwer Academic Publishers
Dubois D, Lorini E, Prade H (2017) The strength of desires: a logical approach. Minds Mach 27(1):199–231. https://doi.org/10.1007/s11023-017-9426-5
El-Nasr MS, Yen J, Ioerger TR (2000) FLAME: fuzzy logic adaptive model of emotions. Auton Agents Multi-Agent Syst 3(3):219–257
Gabbay D, Horty J, Parent X, van der Meyden R, van der Torre L (eds) (2013) Handbook of deontic logic and normative systems. College Publication. http://www.collegepublications.co.uk/downloads/handbooks00001.pdf
Gilbert M (1989) On social facts. Routledge, London
Gochet P, Gribomont P (2006) Epistemic logic. In: Gabbay D, Woods J (eds) Handbook of the history of logic, vol 7. Elsevier, pp 99–195
Gordon R (1987) The structure of emotions. Cambridge University, New York
Gratch J, Marsella S (2005) Lessons from emotion psychology for the design of lifelike characters. J Appl Artif Intell (special issue on Educational Agents - Beyond Virtual Tutors) 19(3–4):215–233
Harel D, Kozen D, Tiuryn J (2000) Dynamic logic. MIT, Cambridge
Herzig A, Lorini E, HÃbner JF, Vercouter L, (2010) A logic of trust and reputation Logic. J IGPL 18(1):214–244
van der Hoek W, Jamroga W, Wooldridge M (2007) Towards a theory of intention revision. Synthese 155(2):265–290
Kamvar SD, Schlosser MT, Garcia-Molina H (2003) The eigentrust algorithm for reputation management in P2P networks. In: 12th international conference on World Wide Web (WWW), ACM, pp 640–651
Kooi B (2007) Expressivity and completeness for public update logic via reduction axioms. J Appl Non-Class Log 17(2):231–253
Koster A, Schorlemmer WM, Sabater-Mir J (2013) Opening the black box of trust: reasoning about trust models in a BDI agent. J Log Comput 23(1):25–58. https://doi.org/10.1093/logcom/exs003

Lane R, Nadel L (eds) (2000) The cognitive neuroscience of emotions., Oxford
Langville AN, Meyer CD (2005) Deeper inside pagerank. Internet Math 1(3):335–400
Laverny N, Lang J (2005) From knowledge-based programs to graded belief-based programs, part ii: off-line reasoning. In: Proceedings of IJCAI'05, Professional book center, pp 497–502
Lazarus RS (1991) Emotion and adaptation. Oxford University, Oxford
van Linder B, van der Hoek WJJC, Meyer, (1998) Formalising abilities and opportunities. Fundamenta Informaticae 34:53–101
Loewenstein G (2000) Emotions in economic theory and economic behavior. Am Econ Rev 90(2):426–432
Lorini E (2011) The cognitive anatomy and functions of expectations revisited. In: Paglieri F, Tummolini L, Falcone R, Miceli M (eds) The goals of cognition: festschfit for cristiano castelfranchi, College Publications, London, to appear
Lorini E (2016) A logic for reasoning about moral agents. Log Anal 58(230):177–218
Lorini E, Schwarzentruber F (2011) A logic for reasoning about counterfactual emotions. Artif Intell 175:814–847
Lorini E, Longin D, Gaudou B, Herzig A (2009) The logic of acceptance: grounding institutions on agents' attitudes. J Log Comput 19(6):901–940. ftp://ftp.irit.fr/IRIT/LILAC/JLC.pdf
Marsh S (1994) Formalising trust as a computational concept. Ph.D. thesis, Department of computing sciece and. mathematics, University of Sterling
Ortony A, Clore G, Collins A (1988) The cognitive structure of emotions. Cambridge University, Cambridge
Osman N, Gutierrez P, Sierra C (2015) Trustworthy advice. Knowl-Based Syst 82:41–59. https://doi.org/10.1016/j.knosys.2015.02.024
Page L, Brin S, Motwani R, Winograd T (1998) The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project
Pelachaud C (2009) Modelling multimodal expression of emotion in a virtual agent. Philos Trans R Soc B 364:3539–3548
Pinyol I, Sabater-Mir J (2013) Computational trust and reputation models for open multi-agent systems: a review. Artif Intell Rev 40(1):1–25. https://doi.org/10.1007/s10462-011-9277-z
Rao AS, Georgeff MP (1991) Modeling rational agents within a BDI-architecture. Morgan Kaufmann Publishers, pp 473–484
de Rosis F, Pelachaud C, Poggi I, Carofiglio V, De Carolis B (2003) From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. Int J Hum-Comput Stud 59:81–118
Sabater J, Sierra C (2005) Review on computational trust and reputation models. Artif Intell 24:33–60
Searle JR (1969) Speech acts: an essay in the philosophy of language. Cambridge
Searle JR (1983) Intentionality: an essay in the philosophy of mind. Cambridge
Segerberg K (1992) Getting started: Beginnings in the logic of action. Studia Logica 51(3–4):347–378
Segerberg K (1995) Belief revision from the point of view of doxastic logic. Log J IGPL 3(4):535–553
Singh MP (1999) An ontology for commitments in multiagent systems. Artif Intell Law 7:97–113
Steunebrink BR, Dastani M, Meyer JJC (2008) A formal model of emotions: integrating qualitative and quantitative aspects. In: Proceedings of the 18th European conference on artificial intelligence (ECAI 2008), IOS, pp 256–260
Turrini P, Meyer JJC, Castelfranchi C (2010) Coping with shame and sense of guilt: a dynamic logic account. J AAMAS 20(3)
van Benthem J (1991) The logic of time. D. Reidel Publishing Company
Wooldridge M (2000) Reasoning about rational agents. MIT, USA
Zeelenberg M, van Dijk WW, Manstead ASR (1998) Reconsidering the relation between regret and responsibility. Organ Behav Hum Decis Process 74:254–272

Negotiation and Persuasion Among Agents

Leila Amgoud, Yann Chevaleyre and Nicolas Maudet

Abstract This chapter presents several techniques allowing agents to come up with an agreement. We start by discussing negotiation among two agents: after having recalled the axiomatic approach of Nash, we present a standard protocol, and point to recent advances in the field. We then discuss issues raised in the multilateral case. Finally, we conclude the chapter by describing an example of persuasion-based negotiation, where agents can put forward justifying reasons through the negotiation, so as to possibly modify preferences over offers or more generally, influence the negotiation process.

1 Introduction

Imagine that a number of robots is to be sent on a remote planet, for exploratory purposes. It is often required to ensure coordination and to allocate different exploration tasks among them. This problem can be tackled as a centralized collective decision problem, as discussed in chapter "Collective Decision Making" of this volume. However, this centralized approach may not always be well suited in our case:

- the computing resources of robots may not allow a designated center to solve the entire problem;
- the amount of information that would need to be communicated to the center could be prohibitive;

L. Amgoud (✉)
IRIT, Université de Toulouse, Toulouse, France
e-mail: leila.amgoud@irit.fr

Y. Chevaleyre
Universit Paris-Dauphine, PSL Research University, CNRS, UMR [7243], LAMSADE, 75016 Paris, France
e-mail: yann.chevaleyre@lamsade.dauphine.fr

N. Maudet
Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, 75005 Paris, France
e-mail: nicolas.maudet@lip6.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_20

- robots may not belong to the same institution (for instance, different countries may have contributed to this group of robots), making it impossible to delegate the allocation task to only one of these institutions.

It is thus sometimes desirable that the allocation occurs as the outcome of a decentralized process (Wellman 1996). More generally, as soon as a problem involves agents owned and programmed by different individuals or institutions, such approaches are appropriate, even though the task is collaborative (Rosenschein and Zlotkin 1994). From the perspective of the designer of the agent, the aim is to come up with a strategy that fills her objective, given the constraints of the system and the (anticipated) strategies of the other agents. From the perspective of the designer of the system, the aim is to design *rules of encounter* (Rosenschein and Zlotkin 1994), that is, rules governing the interaction, in such a way that the system exhibits good properties despite the potentially self-interested behavior of the agents. Such rules constitute the protocol of the system. Some important questions then occur:

- is it possible to guarantee the termination of the process?
- can we provide guarantees regarding the outcome?
- is it possible to give bounds on the amount of computational resources or communication required?

In this chapter we focus on negotiation approaches, which can be seen as decentralized techniques to reach consensus among agents. In Sect. 2 we detail the parameters of the negotiation setting. After having recalled the axiomatics of bilateral negotiation, due to John Nash, we detail basic protocols and strategies, where agents are assumed to have full knowledge of preferences of others Sect. 3.2. In Sect. 4 we discuss negotiation settings involving more than two agents. We conclude the chapter by sketching some aspects of argumentation-based negotiation (Sect. 5), where agents can exchange expressive messages during the negotiation.

2 Parameters of the Negotiation Process

In this chapter, a set of *agents* $\mathscr{A}$ negotiate over a set of alternatives $\mathscr{X}$. Both sets $\mathscr{X}$ and $\mathscr{A}$ may be large. For example, $\mathscr{X}$ may be defined as the cartesian product of a set of attributes (see chapter "Collective Decision Making" of this volume). We assume these agents have quantitative preferences: the satisfaction of agent i for alternative x is given by an utility function $u_i : \mathscr{X} \to \mathbb{R}$.

2.1 Money

An important feature of the negotiation is whether it involves *money* (as opposed to be simply based on exchange of bundles). When this is the case, a common assumption made in economy is that utilities are quasi-linear, which means that they are linear

in the monetary component. More specifically, consider two situations s and s' in which both agents hold the same goods, but such that robot 1 (resp. robot 2) holds p_1 (resp. p_2) more in s'. If utilities are indeed quasi-linear, this would yield utilities:

$$u_1(s) = u_1\left(s'\right) + p_1$$
$$u_2(s) = u_2\left(s'\right) + p_2$$

We say utilities are *transferable* since an agent can transfer a part of his utility by means of money.

2.2 Domains of Negotiation

There are several typical classes of domains on which the negotiation can occur. Rosenschein and Zlotkin (1994) define three main classes: *worth-oriented domains* (WOD), *state-oriented domains* (SOD) and *task-oriented domains* (TOD), ranging from the most general to the most specific. In all of these domains, the agents attempt to find an agreement on the allocation of indivisible tasks (or goods). Using notations introduced in chapter "Collective Decision Making" of this volume, if the set of goods (or tasks) is denoted $\mathscr{O}$, an allocation $\overrightarrow{\pi}$ gives to each agent i a bundle $\pi_i \subseteq \mathscr{O}$. We will focus here on non-shareable goods or tasks. In TODs and SODs, since an assumption of *symmetric abilities* is made, agents share the same valuation for the different states of the world (usually through the same utility function). However, in TODs, each agent is concerned only with her own goods: her preferences may thus be specified on $2^{\mathscr{O}}$ instead of the joint plans of agents, because actions of others have no consequences on her own actions. Furthermore, while in TODs and SODs, the overall goal is assessed as "all-or-nothing" (e.g., robots must explore all the sites together), in WODs, agents have different valuations for different states of the world.

Let us illustrate these ideas with our multi-robot exploration scenario (this example is in part borrowed from Koenig et al. 2006).

Example 1 Consider three robots (r_1, r_2, and r_3) belonging to the same team. These robots have to explore and probe various locations $s_1, \ldots, s_8$, in order to determine if these locations may be drilled later. Here, $\mathscr{X}$ is defined as the set of all different partitions over $s_1, \ldots, s_8$ (so its cardinality is exponential in the number of locations). The cost assigned by a robot to a set of locations is the (minimal) time taken by this robot to visit each location once. Observe that this is typically not *modular*, e.g. the cost visiting several sites may not the sum of visiting each site separately. If the robots are identical and all initially positioned on the same location, we can assume their valuations are identical. In this case, if agents negotiate over the set of locations they wish to visit before the mission starts, the negotiation problem is a TOD. Assume now that r_1 and r_2 need to meet during their mission, in order to obtain supplies from one another. The well-being of r_1 does not depend only on the locations she will visit, but also on the behavior of r_2. This negotiation problem is a SOD. Lastly, if

the agents have different interests regarding the locations to explore, the negotiation problem is a WOD.

2.3 Number of Agents

The number of agents taking part in the negotiation, and the constraints over their interactions are important parameters. In the simplest case, two agents seek an agreement. This is a *bilateral negotiation*. Multilateral negotiation raises other important issues. For example, an agent may be contacted by other agents while a negotiation is already in progress. Designing algorithms that scale up to real world applications is a major issue. Note that these applications often induce specific constraints on the possible interactions: in the case of our robot scenario, we can imagine that the communication system will allow each robot to interact only with its close neighbors. Agents may also have different roles, with specific rights. There are several possible approaches to tackle the problem of multilateral negotiation that we shall describe later. However, even when the number of agents is large, it is often useful to rely on simple negotiation building blocks involving two agents.

2.4 Deadlines

Finally, an important parameter of the negotiation is whether it involves a *deadline* upon which either an agreement is found, or a default outcome (conflict point) is implemented. We will leave this aspect aside in this chapter–the interest reader may find more details in books dedicated to negotiation, e.g. Fatima et al. (2014).

3 Bilateral Negotiation

The simplest setting involves only two agents negotiating together.

3.1 The Axiomatic Perspective

Let us consider the situation of two foraging robots holding in their cart a given amount of resource, and which find out, together, a new deposit they will have to share. Here are some of the options they have:

(o_1) robots share evenly the amount
(o_2) one of the robot exploits 50% while the other exploits 25%

(o_3) one exploits everything, while the other gets nothing
(o_4) one exploits everything, and furthermore steals the content of the other cart

Intuitively, only the first option sounds reasonable. However, if one considers as acceptable any transaction which increases the utility of both robots, only the two last ones should be avoided. This means that we need further principles to define what a "reasonable outcome" of a negotiation for both parties should be.

The approach initiated by Nash (1950) is to start from a set of well-identified axioms:

- *Social rationality (SR)* — the sum of agents' utilities must increase after the transaction. In our example, all the options are socially rational.
- *Pareto-optimality (PO)* — there does not exist any transaction which would be at least as good for both agents, and better for one of them. In our example, option o_2 is not Pareto-optimal, since the second agent could as well take half of the deposit.

These two axioms are minimal conditions, but as we have seen they only exclude one option out of the four. Let us add a condition on the individual utility of agents:

- *Individual Rationality (SR)* — the transaction does not decrease the utility of any of the parties. In our example, option o_4 is not IR since one agent looses the content of its cart without gaining anything. Note that an IR transaction must be SR.

We are left with options o_1 and o_3. We now introduce two further axioms which may be discussed.

- *Scale independence (SI)* — the utility of an agent captures the satisfaction degree for a given situation. Suppose the utility of our two robots vary, depending on circumstances, from 0 to 10 for the first one, and from 0 to 1000 for the second one. Suppose a transaction leads to an outcome of 9 for the first agent, and 500 for agent 2. The fact the the utility of agent 2 is above that of agent 1 does not mean that agent 2 is more satisfied, since the scales used are different. Scale-independence means the selected outcome do not depend of the scale used.
- *Zero-independence (ZI)* — suppose robot 1 has range $[0, \ldots, 0]$ for his utility, while robot 2 has range $[1, \ldots, 10]$. The designer may decide to simply add 1 to the utility of robot 1, so as to have both robots share the same range. A transaction is zero independent iff it is possible to add or retract a constant to the utility function of one of the agent without affecting the outcome of negotiation.

Now denote s_0 the *conflict point*, that is the current, default, situation which will occur if the negotiation fails. The two agents have respective utility $u_1(s_0)$ and $u_2(s_0)$ in this situation (note that this may not be symmetric). What Nash (1950) showed is that any transaction procedure satisfying the aforementioned axioms (as well as some others, more technical) would have to pick the outcome maximizing value

$$(u_1(s_i) - u_1(s_0)) \times (u_2(s_i) - u_2(s_0))$$

In our example, option o_1 would thus be selected. It is noteworthy that SI alone constrains heavily the set of possible outcomes. More generally, if utilities have a

common scale, we can admit they are pairwise comparable. In this case, when an agent has utility higher than that of another agent, this means that this agent is more satisfied than the other one. If we can assume this (which is the case in particular in TOD where utility are identical), then the last two axioms are useless, and we can look for other solutions. For instance, we may seek the transaction which will maximize the utilitarian social welfare, in the sense of the sum of agents' utilities. This solution will only satisfy in general RS, PO, and ZI (IR is thus not guaranteed any longer). If we are concerned with *fairness* notions, then we may consider egalitarian social welfare, which guarantees that the worst-off agent is as satisfied as possible. Unfortunately this solution does not satisfy any of the axioms. It should be noted that Nash was not the only one to adopt an axiomatic perspective on this problem. Most notably, Kalai and Smorodinsky (1975) came up with an alternative solution by considering a slightly different set of axioms.

Example 2 Let us now consider a transaction about minerals m_1 and m_2. The utility of agents is depicted in the following table:

	u_1	u_2
$\emptyset$	0	0
$\{m_1\}$	2	4
$\{m_2\}$	4	2
$\{m_1, m_2\}$	9	9

In situation s_0, agent 1 holds item m_1 and robot 2 item m_2, and they both have the same amount of money. Let s_1 be the situation where both agents have exchanged their items wrt s_0. Let s_2 be the situation where one of the robot holds both goods. Notice that $u_1(s_0) + u_2(s_0) = 4$, that $u_1(s_1) + u_2(s_1) = 8$, and that $u_1(s_2) + u_2(s_2) = 9$. When no money is involved, s_2 is maximizing utilitarian social welfare, but it is not egalitarian optimal, and does not satisfy the Nash criteria. Furthermore, a transaction from s_0 to s_2 does not satisfy IR. These limitations can be circumvented when money is involved in the transaction. Indeed, the transaction from s_0 to s_2 induces a gain of utility of 5 $(9 - 4)$. This constitutes the surplus generated by the transaction, which can be redistributed by the beneficiary agent to the agent whose utility decreases to compensate the loss of utility. More precisely, if we define $s_{2'}$ as the situation where items are allocated similarly, but in which an amount of 4.5 is given by the robot holding both items to the agent with no item. Let us compare $s_{2'}$ to s_0 and s_2.

	u_1	u_2	$u_1 + u_2$	$\min\{u_1, u_2\}$	$u_1 \times u_2$
s_0	2	2	4	2	4
s_2	9	0	9	0	0
s_2'	$9 - 4.5 = 4.5$	$0 + 4.5 = 4.5$	9	4.5	20.25

Thanks to money transfer, the state $s_{2'}$ is now the best possible in the sense of Nash, as well as for the utilitarian and egalitarian social welfare. Here the surplus has been evenly divided among agents, which maximizes both efficiency and equity.

3.2 *Protocols and Strategies for Bilateral Negotiation*

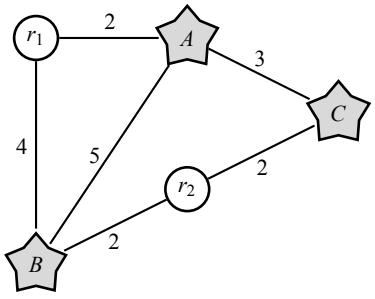
3.2.1 Negotiation Under Complete Information

In this section, we assume agents that may have different preferences, but still full knowledge of all other agent's preferences (this setting is not realistic in a competitive scenario).

Example 3 Consider again a multi-robot problem in which a set of locations have to be visited by robots (see Fig. 1) now starting from different locations. Robots are not required to return to their initial location once their mission is over. Following Rosenschein and Zlotkin (1994), the utility a robot assigns to a bundle of locations will be the difference between the cost of visiting all locations alone for this robot, minus the cost of visiting locations in that bundle.

Possible allocations (assuming each location is visited exactly once) are (with the utility vectors for r_1 et r_2): $o_1 : \langle \emptyset, \{a, b, c\}\rangle = \langle 9, 0\rangle$, $o_2 : \langle \{a\}, \{b, c\}\rangle = \langle 7, 3\rangle$, $o_3 : \langle \{b\}, \{a, c\}\rangle = \langle 5, 4\rangle$, $o_4 : \langle \{c\}, \{a, b\}\rangle = \langle 4, 2\rangle$, $o_5 : \langle \{a, b\}, \{c\}\rangle = \langle 2, 7\rangle$, $o_6 : \langle \{a, c\}, \{b\}\rangle = \langle 4, 7\rangle$, $o_7 : \langle \{b, c\}, \{a\}\rangle = \langle 1, 4\rangle$, et $o_8 : \langle \{a, b, c\}, \emptyset\}\rangle = \langle 0, 9\rangle$. We observe that outcomes o_4, o_5, and o_7 are Pareto dominated. The negotiation will take place on the remaining outcomes.

Let us start with a simple but interesting protocol, described in many AI textbooks: the *monotonic negotiation protocol*. This description follows the ones from Rosenschein and Zlotkin (1994), Wooldridge (2009), and Vidal (2007). The protocol is based on a sequence of *simultaneous* offers made by the agents. At each round t, two offers o_i^t and o_j^t are made, respectively by agent i and agent j. If one offer is



sites	cost r_1	cost r_2	u_1	u_2
$\emptyset$	0	0	9	9
$\{a\}$	2	5	7	4
$\{b\}$	4	2	5	7
$\{c\}$	5	2	4	7
$\{a,b\}$	7	7	2	2
$\{a,c\}$	5	5	4	4
$\{b,c\}$	8	6	1	3
$\{a,b,c\}$	9	9	0	0

Fig. 1 Two robots r_1 et r_2 and three locations A, B, C

satisfactory for the other agent (in other words, if this offer is at least as good for him as what he is himself offering), the protocol stops on an agreement.

$$u_i(o_j^t) \geq u_i(o_i^t) \text{ or } u_j(o_i^t) \geq u_j(o_j^t) \tag{1}$$

Otherwise, another round starts, and each agent is required either to propose the same offer (*stick*), or to *concede* (in other words to make an offer providing a better utility to his partner than the previous offer). If no agent concedes, the protocol ends on a conflict. We assume that in this case, the utility of the agents is the one given at the conflict point o_c.

A possible strategy, in this setting, is suggested by Zeuthen (1930). Intuitively, this strategy consists in assessing the risk of not conceding at some time of the negotiation, by computing the ratio between the loss in utility when conceding (and acception the other agent's offer), and not conceding (and possibly heading towards a conflict). Technically, the propensity of agent i for risking conflict during round t of protocol (noted Z_i^t) is:

$$Z_i^t = \begin{cases} 1 & \text{if} \quad u_i(o_i^t) = 0 \\ \frac{u_i(o_i^t) - u_i(o_j^t)}{u_i(o_i^t) - u_i(o_c)} & \text{else} \end{cases}$$

Each agent is able to compute its propensity for risking conflict, as well as that of its partner. A value close to 1 indicates intuitively that the agent has not much to lose with the conflict, and a value close to 0 shows that the agent fears conflict. The agent which will concede will be the one with the lowest value (or both agents, if their value is the same). The concession to make has to be as small as possible, but high enough to make the other agent concede during next round.

Example 4 Let us go back to our example:

round	offer from r_1	offer from r_2	$u_1(o_{r_1}^t), u_1(o_{r_2}^t)$	$u_2(o_{r_1}^t), u_2(o_{r_2}^t)$	Z_1	Z_2
1	$\langle \emptyset, \{a, b, c\} \rangle$	$\langle \{a, b, c\}, \emptyset \rangle$	9, 0	0, 9	1	1
2	$\langle \{a\}, \{b, c\} \rangle$	$\langle \{a, c\}, \{b\} \rangle$	7, 4	3, 7	$\frac{3}{7}$	$\frac{4}{7}$
3	$\langle \{a, c\}, \{b\} \rangle$	$\langle \{a, c\}, \{b\} \rangle$	4, 4	7, 7	stop	stop

During first round, both agents concede. At second round, the value Z_1 is lower than Z_2, so r_1 will concede. This last offer is the same as the offer r_2 made, but the latter did not concede. In that case, the condition (1) is satisfied for both agents (which is not necessarily the case in general). We observe on this example that the negotiation stops on an offer which maximizes the sum and the product of both utilities. It is also the optimal egalitarian outcome.

However, it is simple to see that the sum of utilities will not always be maximized. Thus, if the two robots must visit two remote sites (see Fig. 2, taken from Rosenschein and Zlotkin 1994), the protocol will lead to a solution where each site will visit a site

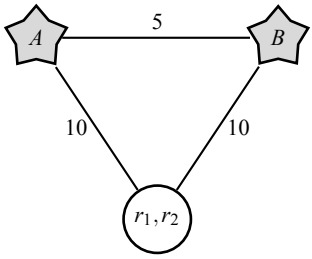


Fig. 2 Two robots r_1 et r_2, with the sites A et B

and return. In this case, as mentioned previously, it would be useful to use money to transfer utility to the agent which made the whole tour.

What are the good properties of this protocol (used in combination with this strategy), in the light of the axiomatic discussion of Sect. 3? Harsanyi (1956) proved that two agents following this protocol will converge to a solution maximizing the product of utilities. However, Nash's axiomatic result assumes the negotiation domain to be convex. Starting from this observation, Zhang (2009) proposes an axiomatic study of TODs: more precisely, he proposes an alternative axiomatization (using additional axioms), which characterize in particular the egalitarian and Nash product solution.

There are of course other protocols to handle bilateral negotiations. In particular, Rubinstein (1982) proposed and analyzed the sequential *alternating offers protocol* where each agent, in turn, either accepts the previous offer of the other agent, or makes a counter-proposal.

3.2.2 Negotiation Under Incomplete Information

In a number of applications (in particular in applications involving self-interested agents), agents can only be assumed to know partially (sometimes, not at all) the preferences of the others. If an a priori distribution can be assumed to be known, then some game-theoretical tools are still available, like for instance the notion of Bayes-Nash equilibrium (we point the reader to Shoham and Leyton-Brown 2009 for a detailed exposure of this and other relevant notions). When it is not the case, the behaviour of the agent must be based on heuristics, ad-hoc strategies which are typically empirically assessed. For instance, it is possible to define different classes of strategies depending on how agents behave as a function of time: a *conceder* agent will be more likely to concede in the first round of negotiation, whereas a *boulware* agent is instead inclined to postpone concessions.

The problem is difficult in particular in multi-issue domains, where agents may have different preferences regarding the different issues at stake. Faratin et al. (1998) distinguish *response strategies* and *compensation strategies*. Responses strategies allow to concede on a single issue (but lead to solution which can be far from being Pareto-optimal). Compensation strategies allow to concede on an issue while

maintaining an overall equivalent utility overall. However, in such settings, an agent cannot even be sure of what constitutes a concession for the other agent. To increase the likelihood of an offer being accepted by the other, an agent should maintain a model of his opponent. One simple intuitive heuristic approach is to seek an offer which is in some sense similar to an offer made by the other agent in the previous round (Faratin et al. 2002). More generally, *opponent modeling* faces three (related) questions (Baarslag et al. 2016): preference estimation, strategy prediction, and opponent classification. For instance, analyzing the opponent's history of concessions may provide valuable information to learn which issues are more important to him (a classical assumption can be that the opponent should concede less easily on these issues). A recent survey discussing the state-of-the-art techniques of the field can be found in Baarslag et al. (2016). Finally, in recent years, the Trading Agent Competition allows agents to compete in a controlled environment, and provides researchers insightful findings regarding heuristics which are most efficient in practice (Wellman et al. 2007).

4 Multilateral Negotiation

We now discuss the case of multilateral settings, where more than two agents are involved in the allocation process. Let us first mention a well-known centralized approach, which relies on *auction mechanisms*. In the application presented in Koenig et al. (2006), a team of robots have to allocate sites to visit, as in our examples. In that case, agents can place *bids* on sites to visit, depending on the cost induced for them. Several types of auctions can be conceived. In principle, each agent could bid on sets of sites to visit. However, observe that the center agent should then solve a combinatorial auction (see chapter "Collective Decision Making" of this volume), and perhaps even more problematically, that each agent must solve a *Traveler Salesman Problem* (TSP) to only evaluate the value of each single set of sites. To circumvent this problem, the authors propose to use *sequential auctions* instead, and show that some performance guarantees can be obtained (Koenig et al. 2006). Even though such protocols are centralized, they can be distributed by letting each agent play the role of the auctioneer, at the price of an overhead of communication. In the rest of this section, we shall discuss other approaches based on negotiation.

4.1 Coordinating Negotiation with a Mediator

A possible approach is to delegate part of the coordination to a designated agent, without requiring this mediator to actually compute the optimal allocation. In the *single text mediated* protocol (Raiffa 1982), the mediator is simply required to make an initial offer (in a multi-issue domain, typically), on which each agent must vote to either *accept* or *refuse* the offer. If *all of the agents* accept, the offer is tagged as

accepted (but the protocol still continues), otherwise it is labelled as *rejected*. Then the mediator looks for another offer, and this process repeats a number of times. Hence, agents never reveal directly anything about their preferences, contrary to an auction. On the other hand, it is clear that without any further information, the mediator will be bound to search blindly in the space of possible outcomes. Each new accepted offer is a Pareto-improvement over the previous one. However, another obvious issue with this protocol is that it doesn't give any guarantee in terms of social welfare. Several works attempt to circumvent these issues, either by employing techniques allowing to escape local optima, like simulated annealing (Klein et al. 2003), or by exploiting the history of interactions, so as to build preference models of the agents, and hence guide the search process (Aydogan et al. 2012).

Another mediated protocol of interest is *fallback bargaining* (Brams and Kilgour 2001), which only assume ordinal preferences over outcomes. In the first round, agents report their preferred outcome to the mediator. If the outcome is the same for everyone, it is chosen. Otherwise, agents report their next preferred outcome, and so on. The protocol stops, at round k, when an outcome at least occurs in the top-k preference of every agent. The outcome is not only Pareto-optimal, but it also provides guarantees in terms of its rank for the least satisfied agent.

4.2 Extending Bilateral Protocols to the Multilateral Setting

A natural question is whether bilateral approaches can naturally be extended to the multilateral case. In particular, let us consider the monotonic concession protocol taken together with the Zeuthen strategy, and see how it can adapted, following Endriss (2006). First, the condition under which the protocol terminates successfully is easily adapted: an agreement is found when an agent makes an offer which at least as good than their current offer, *for all the other agents*. However, things get more complicated when we turn to the definition of what should count as a concession. Indeed, several definitions can be conceived. We may require a concession to be strictly better for all the other agents (*strong concession*), or better for at least one agent (*weak concession*), or increase the sum of utilities of the other agents (*utilitarian concession*), to cite a few examples. Interestingly, these different definitions yield protocols with different properties. For instance, it may not always be possible to avoid *deadlock* situations, in the sense that no agent can make any concession, whereas the outcome is neither an agreement or a conflict.

Example 5 To illustrate that strong concessions can yield deadlock situations, suppose there are only three possible outcomes o_1, o_2 and o_3, yielding the following utilities:

	u_1	u_2	u_3
o_1	2	1	3
o_2	3	2	1
o_3	1	3	2

Take the situation where agents i make offer o_i. No more concessions are possible (all agents enjoy utility 2 in their current offer, so any other makes one of them worse off). Still, no agreement is reached.

4.3 Multilateral Negotiation by Local Deals

Another approach is based on the idea of the *Contract Net protocol* (Smith 1980). Each agent can, depending on the considered task, act as a manager and propose other agents other tasks to be executed. The central idea is thus to allow agents to contract local deals, involving typically a small number of agents. For instance, we may assume each local deal to be bilateral, and thus use different techniques mentioned in this chapter. What can be guaranteed at the global level for such dynamics of local deals? Can we make sure that the system will converge at some point? If so, will the outcome be satisfying? We mention some typical results, based on the work of Sandholm (1998), Dunne et al. (2005), and Chevaleyre et al. (2010). Assuming that agents do not plan ahead, let us assume that deals must be IR. Sandholm (1998) distinguishes in particular the following types of local deals:

- *simple deals*: a single resource is being traded, from one agent to another agent;
- *swap deals*: a resource is swapped against another resource;
- *bilateral deals*: no restriction on resources traded, but involving only two agents.

Bilateral deals thus encompass both simple and swap deals, but are more general. Of course, the complexity of deals can be arbitrarily large. It can be shown that in *modular domains*, any sequence of simple deals (with money transfer is allowed) is guaranteed to reach an outcome maximizing utilitarian social welfare (Endriss et al. 2006). This result is positive in the sense that the class of deals is very simple, but of course modularity remains a strong assumption to be made on the structure of agents' preferences, since it forbids any synergies between resources. In our example involving robots, the condition is unlikely to be satisfied, unless the setting imposes severe restrictions on the possible moves that robots can make: this would be the case if robots were due to return to their initial location after visiting each site. It is thus natural to ask whether any guarantee can be offered on a larger domain. A negative answer can be given, in the following sense: no domain including modular functions can provide the same guarantee on the quality of the outcome. The modular domain is said to be *maximal* for the class of bilateral deals (Chevaleyre et al. 2010). In practice, this means that a designer wishing to use only bilateral deals cannot provide any guarantee unless the domain is certainly modular.

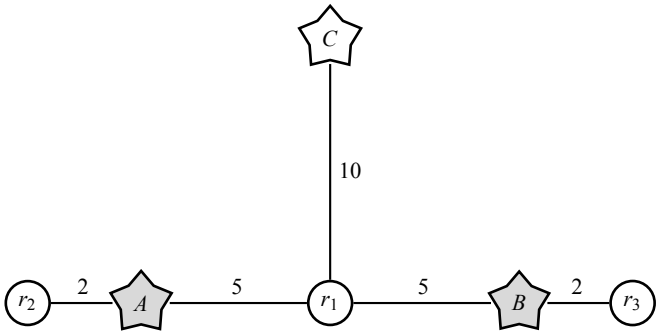


Fig. 3 Robots must take stones from A and B and bring it to C

Example 6 Let us consider three robots r_1, r_2, et r_3. Robot r_1 has a large basket, but r_2 and r_3 both have a small basket. Figure 3 describes the initial situation. Some stones has been discovered on sites A and B, and must be moved to site C. The initial allocation assigns r_2 to A and r_3 to B (total cost $17 + 17 = 34$). Because of their small basket, robots r_2 and r_3 will have to unload the cart to be able to take the stones to the other site (cost: 37). The following allocation would be optimal for the utilitarian social welfare: r_1 should do the whole tour. However, no bilateral deal is possible, since visiting only one of the site is more costly for r_1. Thus, the payment that r_2 or r_3 would require would be too high. This offer can only be reached via a *simultaneous* deal involving all the robots.

Let us now study the speed of convergence of such a protocol, or, to put it differently, the number of deals required to reach an optimal outcome (Endriss and Maudet 2005; Dunne 2005). As a first observation, note that $|\mathscr{O}|^{|\mathscr{A}|}$ is certainly an upper bound, since there are that many allocations, and each deal induces a strict improvement of utilitarian social welafre, which prevents from visiting twice the same allocation. In fact, even restricting the class of deals to single deals does not prevent sequences from being of exponential length (Dunne 2005). However, if we restrict our attention to modular domains, each resource can only visit each other agent (beyond the one which holds it initially): hence we get convergence in $|\mathscr{O}| \times |\mathscr{A}|$ deals in the worst case.

Several other extensions have been studied: other optimality criteria (Dunne 2005), protocols involving richer class of deals (Zheng and Koenig 2009; Chevaleyre et al. 2005), other types of resources (Airiau and Endriss 2010), dealing with underlying graph topology constraining the deals (Nongaillard and Mathieu 2011; de Weerdt et al. 2012; Chevaleyre et al. 2017; Gourvès et al. 2017), or protocols accounting for the limited knowledge agents may have on the preferences of others (Saha and Sen 2007; An et al. 2007).

Finally, a difficulty not mentioned yet is the asynchronous nature of such systems. Indeed, agents can be locked if different deals take place in parallel. In this case, an agent can be tempted to accept a new offer while negotiating with another agent. A proposal to extend the *Contract-Net* protocol to this setting can be found in Aknine et al. (2004). Another concrete example of this approach can be found in An et al.

(2009): the idea is here to use alternating offers so as to allow parallel negotiations. More generally, Sanholm and Lesser (1996) propose to see protocols with limited commitment (uncommitment is allowed, but agents incur a cost when they do) as a manner to allow backtracking in the distributed search performed by a multiagent system.

5 Persuasion-Based Negotiation

Argumentation is a verbal and social activity of reason aimed at increasing (or decreasing) the acceptability of a controversial standpoint for the listener or reader, by putting forward a constellation of propositions intended to justify (or refute) the standpoint before a rational judge. It is also considered as a reasoning model based on the construction and the evaluation of interacting arguments. Those arguments are intended to support statements that can be decisions, opinions, etc (see chapter "Argumentation and Inconsistency-Tolerant Reasoning" of this volume).

Argumentation has developed into an important area of study in artificial intelligence over the last fifteen years, especially in sub-fields such as nonmonotonic reasoning (e.g. Bondarenko et al. 1997; Chesnevar et al. 2000; Dung 1995; Prakken and Vreeswijk 2002; Vreeswijk 1997), multiple-source information systems (e.g. Amgoud and Kaci 2007; Amgoud and Parsons 2002) and decision making (e.g. Amgoud and Prade 2009; Bonet and Geffner 1996; Fox and Parsons 1997). Argumentation has also been extensively used for modeling different kinds of dialogues, in particular persuasion (e.g. Amgoud et al. 2000a; Gordon 1993; Prakken 2005) and inquiry dialogues (e.g. Black and Hunter 2007).

In the nineties, Sycara has emphasized the importance of using argumentation techniques even in negotiation dialogues (Sycara 1990). Since there, several works on argumentation-based negotiation have been done including work by Parsons and Jennings (1996), Reed (1998), Kraus et al. (1998), Tohmé (1997), Amgoud et al. (2000b; 2004), and Kakas and Moraitis (2014; 2006). The basic idea behind an argumentation-based approach for negotiation is to allow agents not only to exchange offers but also reasons that support these offers in order to mutually influence their preferences over offers, and consequently the outcome of the dialogue. Integrating argumentation theory in negotiation provides a good means for supplying additional information and helps agents to convince each other by adequate arguments during a negotiation dialogue. Indeed, an offer supported by a good argument has a better chance to be accepted by an agent, and can also make her reveal her goals or give up some of them. The basic idea behind an argumentation-based approach is that by exchanging arguments, the theories of the agents (i.e. their mental states) may evolve, and consequently, the *status* of offers may change. For instance, an agent may reject an offer because it is not acceptable, then it changes her mind if she receives a strong argument in favor of this offer.

In the literature, there are two categories of works on argumentation-based negotiation. The first category studies the role of argumentation in negotiation. It was

shown in Amgoud and Vesic (2012) that argumentation may improve the quality of solutions reached in negotiation. The same conclusion was also revealed by an empirical study done in Pasquier et al. (2010), where the results of a negotiation model are compared when arguments are exchanged and when they are not allowed.

The second category of works defines concrete negotiation models (Amgoud et al. 2000b; Kakas and Moraitis 2006; Kraus et al. 1998; Parsons and Jennings 1996; Reed 1998; Tohmé 1997). Each model shows how arguments are built from knowledge bases (containing the mental states of agents), and how these arguments are evaluated and then exchanged using a protocol. For instance, the *alternating offers* protocol proposed by Rubinstein (1982) for bargaining between agents was extended in Hadidi et al. (2010) for considering arguments. In what follows, we do not focus on a particular protocol, but rather present the main ideas behind a argumentation-based negotiation model.

5.1 Agent Theory

In what follows, $\mathscr{L}$ will denote a logical language, and $\equiv$ is an equivalence relation associated with $\mathscr{L}$, and θ a symbol not appearing in $\mathscr{L}$. From $\mathscr{L}$, a finite set $\mathscr{O}$ of distinct *offers* is identified (i.e., $\nexists o, o' \in \mathscr{O}$ such that $o \equiv o'$).

Different *arguments* can be built from $\mathscr{L}$. The set $\texttt{Args}(\mathscr{L})$ will contain all those arguments (see chapter "Argumentation and Inconsistency-Tolerant Reasoning" of this volume for a formal definition of argument). The selection of the best offer to propose at a given step of the dialogue is a decision problem. In Amgoud and Prade (2009), it has been shown that in an argumentation-based approach for decision making, two kinds of arguments are distinguished: arguments supporting choices (or decisions), and arguments supporting beliefs. Moreover, it has been acknowledged that the two categories of arguments are formally defined in different ways, and they play different roles. Indeed, an argument in favor of a decision, built both on an agent's beliefs and goals, tries to justify the choice; whereas an argument in favor of a belief, built only from beliefs, tries to destroy the decision arguments, in particular the beliefs part of those decision arguments. Consequently, in a negotiation dialogue, those two kinds of arguments are generally exchanged between agents. In what follows, the set $\texttt{Args}(\mathscr{L})$ is then divided into two subsets: a subset $\texttt{Args}_o(\mathscr{L})$ of arguments supporting offers, and a subset $\texttt{Args}_b(\mathscr{L})$ of arguments supporting beliefs. Thus, $\texttt{Args}(\mathscr{L}) = \texttt{Args}_o(\mathscr{L}) \cup \texttt{Args}_b(\mathscr{L})$ and $\texttt{Args}_o(\mathscr{L}) \cap \texttt{Args}_b(\mathscr{L}) = \emptyset$.

Since knowledge bases from which arguments are built may be inconsistent, arguments may be conflicting too. In what follows, those conflicts will be captured by the binary relation $\mathbb{R}_{\mathscr{L}}$, i.e. $\mathbb{R}_{\mathscr{L}} \subseteq \texttt{Args}(\mathscr{L}) \times \texttt{Args}(\mathscr{L})$. Two assumptions are made on this relation: First the arguments supporting different offers are conflicting. The idea behind this assumption is that since offers are exclusive, an agent has to choose only one at a given step of a dialogue. Note that, the relation $\mathbb{R}_{\mathscr{L}}$ is not necessarily symmetric between the arguments of $\texttt{Args}_b(\mathscr{L})$. The second condition does not allow an argument in favor of an offer to attack an argument supporting a

belief. This avoids wishful thinking. Formally: $\mathbb{R}_{\mathscr{L}} \subseteq \texttt{Args}(\mathscr{L}) \times \texttt{Args}(\mathscr{L})$ is a binary relation such that:

- $\forall a, b \in \texttt{Args}_o(\mathscr{L})$ such that $a \neq b$, $(a, b) \in \mathbb{R}_{\mathscr{L}}$.
- $\nexists (a, b) \in \mathbb{R}_{\mathscr{L}}$ such that $a \in \texttt{Args}_o(\mathscr{L})$ and $b \in \texttt{Args}_b(\mathscr{L})$.

An agent involved in a negotiation, called *negotiating agent*, has a *theory*. It is made of a finite set of arguments, a conflict relation among those arguments, a preference relation between the arguments, and a function that specifies which arguments support offers of the set $\mathscr{O}$. In the literature, an agent is always assumed to be aware of all the arguments of the set $\texttt{Args}(\mathscr{L})$. The agent is even able to express preference $\succeq$ between any pair of arguments, where $a \succeq b$ means that a is at least as preferred as b. The relation $\succeq$ is a (partial or total) preorder, i.e., reflexive and transitive. Its strict version is denoted by $\succ$. Note that the fact that $\succeq$ is defined over $\texttt{Args}(\mathscr{L})$ does not mean that the agent will use all the arguments of $\texttt{Args}(\mathscr{L})$. The assumption rather encodes the fact that when an agent receives an argument from another agent, it can interpret it correctly, and compare it with her own arguments. Similarly, each agent is supposed to be aware of the conflicts between arguments.

Definition 1 A *negotiating agent theory* is a tuple $\mathscr{T} = \langle \mathscr{O}, \mathscr{A}, \mathscr{F}, \succeq, \mathbb{R} \rangle$ such that:

- $\mathscr{O}$ a finite set of offers.
- $\mathscr{A}$ is a finite subset of $\texttt{Args}(\mathscr{L})$.
- $\mathscr{F}: \mathscr{O} \rightarrow 2^{\mathscr{A}}$ such that $\forall o, o' \in \mathscr{O}$ with $o \neq o'$, $\mathscr{F}(o) \cap \mathscr{F}(o') = \emptyset$.
- $\succeq \subseteq \texttt{Args}(\mathscr{L}) \times \texttt{Args}(\mathscr{L})$ is a (partial or total) preorder.
- $\mathbb{R} = \{(a, b) \in \mathbb{R}_{\mathscr{L}} \mid (a, b) \in \mathscr{A} \times \mathscr{A}\}$.

The function $\mathscr{F}$ returns the set of arguments supporting a given offer. Arguments are evaluated using any acceptability semantics from the literature. In what follows, we illustrate a negotiation framework using the stable semantics proposed by Dung (1995).

Definition 2 A set $\mathscr{E} \subseteq \mathscr{A}$ is a stable extension of a theory $\mathscr{T} = \langle \mathscr{O}, \mathscr{A}, \mathscr{F}, \succeq, \mathbb{R} \rangle$ iff i) $\nexists a, b \in \mathscr{E}$ such that $(a, b) \in \mathbb{R}$, and ii) $\forall a \in \mathscr{A} \setminus \mathscr{E}, \exists b \in \mathscr{E}$ such that $(b, a) \in \mathbb{R}$ and not$(a \succ b)$. Let $\texttt{Ext}(\mathscr{T})$ denote the set of all stable extensions of the theory $\mathscr{T}$.

Note that under stable semantics, a theory may have zero, one, or more extensions. From the extensions, a qualitative status is assigned to each argument.

Definition 3 Let $\mathscr{T} = \langle \mathscr{O}, \mathscr{A}, \mathscr{F}, \succeq, \mathbb{R} \rangle$ be an agent theory, and $a \in \mathscr{A}$. If $\texttt{Ext}(\mathscr{T}) = \emptyset$, then a is *undecided*, otherwise:

- a is *accepted* iff $a \in \bigcap_{\mathscr{E} \in \texttt{Ext}(\mathscr{T})} \mathscr{E}$,
- a is *rejected* iff $a \notin \bigcup_{\mathscr{E} \in \texttt{Ext}(\mathscr{T})} \mathscr{E}$,
- a is *undecided* iff $\exists \mathscr{E}, \mathscr{E}' \in \texttt{Ext}(\mathscr{T})$ such that $a \in \mathscr{E}$ and $a \notin \mathscr{E}'$.

Note that $\mathscr{A} = \{a | a \text{ is accepted}\} \cup \{a | a \text{ is rejected}\} \cup \{a | a \text{ is undecided}\}$. From the statuses of arguments, a qualitative status is also assigned to each offer.

Definition 4 Let $\mathscr{T} = \langle \mathscr{O}, \mathscr{A}, \mathscr{F}, \succeq, \mathbb{R} \rangle$ be an agent theory, and $o \in \mathscr{O}$. If $\mathscr{F}(o) = \emptyset$, then o is *non-supported*, otherwise:

- o is *acceptable* iff $\exists a \in \mathscr{F}(o)$ such that a is accepted.
- o is *rejected* iff $\forall\, a \in \mathscr{F}(o)$, a is rejected.
- o is *negotiable* iff $\forall a \in \mathscr{F}(o)$, a is undecided.

Let $\mathscr{O}_a(\mathscr{T})$ (respectively $\mathscr{O}_r(\mathscr{T})$, $\mathscr{O}_n(\mathscr{T})$, $\mathscr{O}_{ns}(\mathscr{T})$) denote the set of acceptable (respectively rejected, negotiable, non-supported) offers in theory $\mathscr{T}$.

Obviously, the above definition provides a partition of the set $\mathscr{O}$ of offers. Indeed, $\mathscr{O} = \mathscr{O}_a(\mathscr{T}) \cup \mathscr{O}_r(\mathscr{T}) \cup \mathscr{O}_n(\mathscr{T}) \cup \mathscr{O}_{ns}(\mathscr{T})$. A preference relation on $\mathscr{O}$ is also defined. The idea is that any acceptable offer is strictly preferred to any negotiable offer, which in turn is more acceptable than any non-supported offer. Non-supported offers are strictly preferred to any rejected one. Let X and Y be two subsets of $\mathscr{O}$. $X \rhd Y$ means that any offer in X is strictly preferred to any offer in the set Y. We can write also for two offers o, o', $o \rhd o'$ iff $o \in X$, $o' \in Y$ and $X \rhd Y$.

Definition 5 Let $\mathscr{T} = \langle \mathscr{O}, \mathscr{A}, \mathscr{F}, \succeq, \mathbb{R} \rangle$ be an agent theory. The following holds: $\mathscr{O}_a(\mathscr{T}) \rhd \mathscr{O}_n(\mathscr{T}) \rhd \mathscr{O}_{ns}(\mathscr{T}) \rhd \mathscr{O}_r(\mathscr{T})$.

5.2 Negotiation Dialogues

A negotiation takes generally place between two or more agents. For simplicity reasons, we assume only two agents P and C. Each agent $i \in \{P, C\}$ is equipped with a theory $\mathscr{T}_i = (\mathscr{O}, \mathscr{A}_i, \mathscr{F}_i, \succeq_i, \mathbb{R}_i)$ which is used for computing the preference relation $\rhd_i$ on the set $\mathscr{O}$ of offers. During a dialogue, the two agents exchange offers and arguments. In what follows, we present a very general definition of dialogue which can then be extended by the rules of any protocol.

Definition 6 A *negotiation dialogue* between two agents P, C over a set $\mathscr{O}$ of offers is a finite sequence of moves $d = \langle m_1, \ldots, m_l \rangle$ such that $m_i = \langle x_i, y_i, z_i \rangle$, where $x_i \in \{P, C\}$, $y_i \in \mathtt{Args}(\mathscr{L}) \cup \{\theta\}$, and $z_i \in \mathscr{O} \cup \{\theta\}$. If $\forall i = 1, \ldots, l$, $y_i = \theta$, then d is said *non-argumentative*. It is *argumentative* otherwise.

At each step t of a dialogue, the theory of each agent may evolve. The original set of arguments is augmented by new ones received from the other party, and the attack relation is modified consequently. Let $\mathscr{T}_i^t = (\mathscr{O}, \mathscr{A}_i^t, \mathscr{F}_i^t, \succeq_i^t, \mathbb{R}_i^t)$ denote the theory of agent $i \in \{P, C\}$ at a step t of a dialogue $d = \langle m_1, \ldots, m_l \rangle$ and $\mathscr{T}_i^0$ her theory before the dialogue. Obviously, the theories of the two agents do not change in case d is non-argumentative.

Property 1 *If a dialogue $d = \langle m_1, \ldots, m_l \rangle$ is non-argumentative, then $\forall i \in \{P, C\}$, $\forall j \in \{1, \ldots, l\}$, it holds that $\mathscr{T}_i^j = \mathscr{T}_i^0$.*

Let us now analyze the different solutions of a negotiation dialogue. The best solution for an agent at a given step of a dialogue is that which suits best her preferences, i.e., an acceptable offer in her own theory. However, an offer may be accepted for one agent but not for the other. Such offer is not suitable as a solution of the dialogue. A *local solution* at a given step is an offer which is accepted for both agents at that step. We use the term "local" because such an offer is accepted locally in time – it may have been rejected before, or may become rejected after several steps. Such a solution does not always exist.

Definition 7 An offer $o \in \mathscr{O}$ is a *local solution* at step t of a negotiation dialogue d iff $o \in \mathscr{O}_a(\mathscr{T}_P^t) \cap \mathscr{O}_a(\mathscr{T}_C^t)$.

A local solution is not necessarily a dialogue outcome since the two agents may miss it. In order to be so, an efficient protocol should be used. Furthermore, it is time-dependent. An offer may, for instance, be a local solution at step t but not at step $t+1$. In what follows, we define two other solutions (one for a single agent and one for a dialogue) which are not time-dependent. They represent respectively the *optimal solution* for an agent and the *ideal solution* of a dialogue. An offer is an optimal solution for an agent iff she would choose that offer if she had access to all arguments owned by all agents.

Definition 8 An offer $o \in \mathscr{O}$ is an *optimal solution* for agent $i \in \{P, C\}$ iff $o \in \mathscr{O}_a(\mathscr{T}_i)$ where $\mathscr{T}_i = (\mathscr{O}, \mathscr{A}_P^0 \cup \mathscr{A}_C^0, \mathscr{F}_i, \succeq_i, \mathbb{R}_i)$ with $\mathbb{R}_i \subseteq (\mathscr{A}_P^0 \cup \mathscr{A}_C^0) \times (\mathscr{A}_P^0 \cup \mathscr{A}_C^0)$.

The following property shows that if an offer is optimal for an agent, then there exists a dialogue in which that solution is accepted for that agent at a given step.

Property 2 *If o is an optimal solution for an agent, then there exists a dialogue d such that o is accepted for that agent at step t.*

If both agents agree when all information has been exchanged, they can obtain an ideal solution.

Definition 9 An offer $o \in \mathscr{O}$ is an *ideal solution* iff $o \in \mathscr{O}_a(\mathscr{T}_P) \cap \mathscr{O}_a(\mathscr{T}_C)$ where $\mathscr{T}_P = \langle \mathscr{O}, \mathscr{A}_P^0 \cup \mathscr{A}_C^0, \mathscr{F}_P^0, \succeq_P, \mathbb{R}_P^0 \rangle$ and $\mathscr{T}_C = \langle \mathscr{O}, \mathscr{A}_P^0 \cup \mathscr{A}_C^0, \mathscr{F}_C^0, \succeq_2, \mathbb{R}_C^0 \rangle$.

The next property shows that if an ideal solution exists, then it is a local solution for a dialogue.

Property 3 *If o is an ideal solution then there exists a dialogue $d = (m_1, \ldots, m_l)$ such that o is a local solution at step l.*

It is natural to expect that for two agents with same beliefs and goals an exchange of arguments can improve the chance of finding a solution.

6 Conclusion

This chapter briefly presented distributed procedures based on negotiation for reaching agreement, for instance regarding task or resource allocation. The reader interested in the robotic scenario used for illustration purpose may find a more exhaustive survey in Dias et al. (2006). Several other applications could have been used to illustrate such distributed approaches: negotiations in smart grids (Vytelingum et al. 2010), or task allocation in the medical domain (Paulussen et al. 2003; Vermeulen et al. 2007). Again, many aspects of negotiations have been left aside. To take a single example, we only briefly mentioned challenges of multi-attribute negotiation, which led to recent theoretical and empirical developments (Fatima et al. 2006, 2014; Lai et al. 2008).

References

Airiau S, Endriss U (2010) Multiagent resource allocation with sharable items: simple protocols and nash equilibria. In: Proceedings of the 9th international conference on autonomous agents and multiagent systems (AAMAS-2010), pp 167–174

Aknine S, Pinson S, Shakun MF (2004) An extended multi-agent negotiation protocol. Auton Agents Multi Agent Syst 8(1):5–45

Amgoud L, Kaci S (2007) An argumentation framework for merging conflicting knowledge bases. Int J Approx Reason 45:321–340

Amgoud L, Parsons S (2002) An argumentation framework for merging conflicting knowledge bases. In: Proceedings of the 8th European conference on logics in artificial intelligence (JELIA'02). LNCS, vol 2424, pp 27–37

Amgoud L, Prade H (2004) Reaching agreement through argumentation: a possibilistic approach. In: Proceedings of the 9th international conference on the principles of knowledge representation and reasoning (KR'04), pp 175–182

Amgoud L, Prade H (2009) Using arguments for making and explaining decisions. Artif Intell J 173:413–436

Amgoud L, Vesic S (2012) A formal analysis of the role of argumentation in negotiation dialogues. J Log Comput 22(5):957–978

Amgoud L, Maudet N, Parsons S (2000a) Modelling dialogues using argumentation. In: Proceedings of the 4th international conference on multiagent systems (ICMAS'00). ACM, New York, pp 31–38

Amgoud L, Parsons S, Maudet N (2000b) Arguments, dialogue, and negotiation. In: Proceedings of the 14th European conference on artificial intelligence (ECAI'00). IOS, Amsterdam, pp 338–342

An B, Miao C, Shen Z (2007) Market based resource allocation with incomplete information. In: Proceedings of the 20th international joint conference on artificial intelligence (IJCAI 2007), pp 1193–1198

An B, Gatti N, Lesser VR (2009) Extending alternating-offers bargaining in one-to-many and many-to-many settings. In: Proceedings of the 2009 IEEE/WIC/ACM international conference on intelligent agent technology (IAT 2009), pp 423–426

Aydogan R, Hindriks KV, Jonker CM (2012) Multilateral mediated negotiation protocols with feedback. In: The fifth international workshop on agent-based complex automated negotiations (ACAN 2012). Valencia, Spain

Baarslag T, Hendrikx MJC, Hindriks KV, Jonker CM (2016) Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. Auton Agents Multi Agent Syst 30(5):849–898. https://doi.org/10.1007/s10458-015-9309-1

Black E, Hunter A (2007) A generative inquiry dialogue system. In: Proceedings of the 6th international joint conference on autonomous agents and multi-agents systems (AAMAS'07)

Bondarenko A, Dung P, Kowalski R, Toni F (1997) An abstract, argumentation-theoretic approach to default reasoning. Artif Intell J 93:63–101

Bonet B, Geffner H (1996) Arguing for decisions: a qualitative model of decision making. In: Proceedings of the 12th conference on uncertainty in artificial intelligence (UAI'96), pp 98–105

Brams SJ, Kilgour DM (2001) Fallback bargaining. Group Decis Negot 10(4):287–316

Chesnevar CI, Maguitman A, Loui RP (2000) Logical models of arguments. ACM Comput Surv 32(4):337–383

Chevaleyre Y, Endriss U, Lang J, Maudet N (2005) Negotiating over small bundles of resources. In: Proceedings of the 4th international joint conference on autonomous agents and multiagent systems (AAMAS-2005). ACM, New York, pp 296–302. https://doi.org/10.1145/1082473.1082518

Chevaleyre Y, Endriss U, Maudet N (2010) Simple negotiation schemes for agents with simple preferences: sufficiency, necessity and maximality. J Auton Agents Multiagent Syst 20(2):234–259

Chevaleyre Y, Endriss U, Maudet N (2017) Distributed fair allocation of indivisible goods. Artif Intell 242:1–22

de Weerdt M, Zhang Y, Klos T (2012) Multiagent task allocation in social networks. Auton Agents Multi Agent Syst 25(1):46–86

Dias MB, Zlot R, Kalra N, Stentz A (2006) Market-based multirobot coordination: a survey and analysis. Proc IEEE 94(7):1257–1270

Dimopoulos Y, Moraitis P (2014) Advances in argumentation-based negotiation. In: Lopes F, Coelho H (eds) Chapter 4, in book "Negotiation and argumentation in multi-agent systems: fundamentals, theories, systems and applications", pp 82–125

Dung PM (1995) On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. Artif Intell J 77:321–357

Dunne PE (2005) Extremal behaviour in multiagent contract negotiation. J Artif Intell Res 23:41–78

Dunne PE, Wooldridge M, Laurence M (2005) The complexity of contract negotiation. Artif Intell 164(1–2):23–46

Endriss U (2006) Monotonic concession protocols for multilateral negotiation. In: Stone P, Weiss G (eds) Proceedings of the 5th international joint conference on autonomous agents and multiagent systems (AAMAS-2006). ACM, New York, pp 392–399. https://doi.org/10.1145/1160633.1160702

Endriss U, Maudet N (2005) On the communication complexity of multilateral trading: extended report. J Auton Agents Multiagent Syst 11(1):91–107

Endriss U, Maudet N, Sadri F, Toni F (2006) Negotiating socially optimal allocations of resources. J Artif Intell Res 25:315–348

Fatima S, Kraus S, Wooldridge M (2014) Principle of automated negotiation. Cambridge University, Cambridge

Faratin P, Sierra C, Jennings NR (1998) Negotiation decision functions for autonomous agents. Robot Auton Syst 24(3–4):159–182

Faratin P, Sierra C, Jennings NR (2002) Using similarity criteria to make issue trade-offs in automated negotiations. Artif Intell 142(2):205–237

Fatima SS, Wooldridge M, Jennings NR (2006) Multi-issue negotiation with deadlines. J Artif Intell Res 27:381–417

Fox J, Parsons S (1997) On using arguments for reasoning about actions and values. In: Proceedings of the AAAI spring symposium on qualitative preferences in deliberation and practical reasoning. Stanford, CA, US

Gordon TF (1993) The pleadings game. Artif Intell Law 2:239–292

Gourvès L, Lesca J, Wilczynski A (2017) Object allocation via swaps along a social network. In: Proceedings of the 26th international joint conference on artificial intelligence (IJCAI-17). Melbourne, Australia, pp 213–219. https://doi.org/10.24963/ijcai.2017/31

Hadidi N, Dimopoulos Y, Moraitis P (2010) Argumentative alternating offers. In: van der Hoek W, Kaminka GA, Lespérance Y, Luck M, Sen S (eds) AAMAS, IFAAMAS, pp 441–448

Harsanyi JC (1956) Approaches to the bargaining problem before and after the theory of games: a critical discussion of Zeuthen's, Hick's and Nash theories. Econometrica 24:144–157

Kakas A, Moraitis P (2006) Adaptive agent negotiation via argumentation. In: Proceedings of the 5th international joint conference on autonomous agents and multi-agents systems (AAMAS'06), pp 384–391

Kalai E, Smorodinsky M (1975) Other solutions to nash's bargaining problem. Econometrica 43(3):513–518. http://www.jstor.org/stable/1914280

Klein M, Faratin P, Sayama H, Bar-Yam Y (2003) Protocols for negotiating complex contracts. IEEE Intell Syst 18(6):32–38

Koenig S, Tovey CA, Lagoudakis MG, Markakis E, Kempe D, Keskinocak P, Kleywegt AJ, Meyerson A, Jain S (2006) The power of sequential single-item auctions for agent coordination. In: Proceedings of the AAAI conference on artificial intelligence (AAAI). AAAI Press, pp 1625–1629

Kraus S, Sycara K, Evenchik A (1998) Reaching agreements through argumentation: a logical model and implementation. J Artif Intell 104:1–69

Lai G, Sycara KP, Li C (2008) A decentralized model for automated multi-attribute negotiations with incomplete information and general utility functions. Multiagent Grid Syst 4(1):45–65

Nash J (1950) The bargaining problem. Econometrica 28:155–162

Nongaillard A, Mathieu P (2011) Reallocation problems in agent societies: a local mechanism to maximize social welfare. J Artif Soc Soc Simul 14(3):5

Parsons S, Jennings NR (1996) Negotiation through argumentation—a preliminary report. In: Proceedings of the 2nd international conference on multi agent systems, pp 267–274

Pasquier P, Hollands R, Rahwan I, Dignum F, Sonenberg L (2010) An empirical study of interest-based negotiation. Auton Agents Multi Agent Syst. To appear

Paulussen TO, Jennings NR, Decker KS, Heinzl A (2003) Distributed patient scheduling in hospitals. In: Gottlob G, Walsh T (eds) IJCAI. Morgan Kaufmann, Burlington, US, pp 1224–1232

Prakken H (2005) Coherence and flexibility in dialogue games for argumentation. J Log Comput 15:1009–1040

Prakken H, Vreeswijk GAW (2002) Logics for defeasible argumentation. In: Handbook of philosophical logic, vol 4. Kluwer Academic, Dordrecht, pp 219–318

Raiffa H (1982) The art and science of negotiation. Harvard university, Cambridge

Reed C (1998) Dialogue frames in agent communication. In: Proceedings of the 3rd international conference on multi agent systems (ICMAS'98), pp 246–253

Rosenschein JS, Zlotkin G (1994) Rules of encounter. MIT, Cambridge, MA, USA

Rubinstein A (1982) Perfect equilibrium in a bargaining mode. Econometrica 50(1):97–109

Saha S, Sen S (2007) An efficient protocol for negotiation over multiple indivisible resources. In: Veloso MM (ed) IJCAI, pp 1494–1499

Sandholm TW (1998) Contract types for satisficing task allocation: I theoretical results. In: Proceeding of the AAAI spring symposium: satisficing models

Sandholm T, Lesser VR (1996) Advantages of a leveled commitment contracting protocol. In: Clancey WJ, Weld DS (eds) AAAI/IAAI, vol 1. AAAI/MIT, Cambridge, MA, USA, pp 126–133

Shoham Y, Leyton-Brown K (2009) Multiagent systems: algorithmic, game-theoetic, and logical foundations. Cambridge University, Cambridge

Smith R (1980) The contract net protocol: high level communication and control in distributed problem solver. IEEE Trans Comput 29:1104–1113

Sycara K (1990) Persuasive argumentation in negotiation. Theory Decis 28:203–242

Tohmé F (1997) Negotiation and defeasible reasons for choice. In: Proceedings of the stanford spring symposium on qualitative preferences in deliberation and practical reasoning, pp 95–102

Vermeulen IB, Bohte SM, Somefun K, Poutré JAL (2007) Multi-agent pareto appointment exchanging in hospital patient scheduling. Serv Oriented Comput Appl 1(3):185–196
Vidal JM (2007) Fundamentals of multiagent systems. http://jmvidal.cse.sc.edu/papers/mas.pdf
Vreeswijk GAW (1997) Abstract argumentation systems. Artif Intell J 90:225–279
Vytelingum P, Ramchurn SD, Voice T, Rogers A, Jennings NR (2010) Trading agents for the smart electricity grid. In: 9th international conference on autonomous agents and multiagent systems (AAMAS 2010), pp 897–904
Wellman MP (1996) Market-oriented programming: some early lessons. In: Clearwater S (ed) Market-based control: a paradigm for distributed resource allocation. World Scientific, Singapore
Wellman MP, Greenwald A, Stone P (2007) Autonomous bidding agents: strategies and lessons from the trading agent competition. MIT, Cambridge
Wooldridge M (2009) An introduction to multiagent systems, 2nd edn. Wiley, New York
Zeuthen F (1930) Problems of monopoly and economic warfare. Routledge, London
Zhang D (2009) Axiomatic characterization of task oriented negotiation. In: Proceedings of the 21st international joint conference on artificial intelligence (IJCAI-09), pp 935–940
Zheng X, Koenig S (2009) K-swaps: cooperative negotiation for solving task-allocation problems. In: Boutilier C (ed) IJCAI, pp 373–379

Diagnosis and Supervision: Model-Based Approaches

Marie-Odile Cordier, Philippe Dague, Yannick Pencolé and Louise Travé-Massuyès

Abstract This chapter is devoted to diagnosis and supervision. It is organized as follows: after a section dedicated to the logical formalization of model-based diagnosis, the focus is made on diagnosis of discrete event systems modeled by automata. In the last section, one presents more succinctly the works that allowed to make the bridge between the approaches proposed by the Artificial Intelligence community and those proposed by the Automatic Control community.

1 Introduction

Diagnosis consists in observing a system (often by using sensors), in detecting from these observations possible dysfunctions or mode change (from normal to abnormal) and in identifying the fault(s) they evoke. Diagnosis can be carried out in the medical field but also in the industrial field or even the environmental, economic ones, etc. The first works in Artificial Intelligence (AI) dealing with diagnosis were, in the 1980s, the expert systems based approaches, which appeared with the application to medical diagnosis and the Mycin system. These approaches relied on general on production rules whose condition part describes observable signs and symptoms and conclusion part the diagnoses they evoke. These associative approaches for diagnosis have continued and gave rise to case-based reasoning approaches (see chapter "Case-Based Reasoning, Analogical Reasoning, and Interpolation" of this volume)

M.-O. Cordier
IRISA, Rennes, France
e-mail: marie-odile.cordier@irisa.fr

P. Dague
LRI, Université Paris-Sud, CNRS, Orsay, France
e-mail: philippe.dague@lri.fr

Y. Pencolé (✉) · L. Travé-Massuyès
LAAS-CNRS, Toulouse, France
e-mail: ypencole@laas.fr

L. Travé-Massuyès
e-mail: louise@laas.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_21

and, for dynamical systems, to chronicles (or scenarios) recognition approaches, where one associates to a set of temporally constrained events the diagnostic situation to which these events correspond.

Despite their success, it has been often reproached associative approaches for coding reasoning shortcuts, left without explanations capabilities relying on the functioning of the system to be diagnosed. This motivated the introduction of model-based approaches, which rely on a description of the behavior of the supervised system, this model being possibly limited to the behavior of the so-called normal behavior of the system studied. It will be seen in the following that it is often relevant to join it, when available, fault models, describing also the behaviors resulting from the occurrence of a fault. One can distinguish between predictive models, which allow prediction of the system's behavior, in particular the values observed by the sensors, and explanatory models, which allow explanation of observations resulting from the faults that occurred.

Diagnosis problem interested a lot researchers in AI. It actually associates a modeling problem, therefore the choice of a formalism (based on logic, graphs or constraints) for behavior representation of the system studied with its uncertainty and complexity, a diagnoses characterization problem and a heuristic algorithmic problem for solving with satisfactory efficiency a task, which is most of the time NP-hard. And this field by the way influenced considerably AI research, since expert systems, default logic, fuzzy logic and non monotonic logics, constraints, causal graphs, qualitative reasoning had often their first applications in this framework. As it will seen in this chapter, this field motivated also largely researchers in the Automatic Control field who, firstly more focused on control, expanded their interest to search of the causes of the dysfunctions detected.

It can be finally noticed that diagnosis is not in general an end per se and that the issue is to "repair" the monitored system, which relates it directly to research in decision theory (see chapters "Multicriteria Decision Making" and "Decision Under Uncertainty" of this volume) and in planning (see chapter "Planning in Artificial Intelligence" of Volume 2). Last, diagnosis depends very directly on the means available for observing the system and research in diagnosis has direct links with systems design and observability and also their repairability if one is interested, as it is most often the case currently, in the design of autonomous and embedded systems, as well with their hardware features as their software ones.

2 Logical Framework for Diagnosis

The formalization of the theory of diagnosis at the end of the eighties has been firstly introduced separately regarding consistency-based diagnosis and regarding abductive diagnosis. In the first case, one requires only for a diagnosis, i.e., an assignment of behavioral modes – normal or abnormal – to each component of the system, to be consistent with the system model and the observations. In the second case, one requires additionally for a diagnosis to "explain", jointly with the system

model, all or some of the observations. Initially, this second case was most often handled in the framework of "naturally abductive" models such as causal graphs or Bayesian models and called on concepts of set covering. It is only a bit later that both approaches converged, the logical framework allowing the whole spectrum from simple consistency to whole abductive to be expressed.

2.1 Consistency-Based Logical Approach

The theory of consistency-based diagnosis was expressed for the first time in a logical framework, which will no longer vary afterwards, in Reiter (1987). This framework claims to be valid for any system structurally described in terms of components, the model of the system being assumed to be given by a first-order theory. One assumes likewise to have available a (sound and complete) first-order solver for checking inconsistency, which, in its whole generality, can be only a semi-algorithm as first-order theory is undecidable. The theory developed is completely independent from the choice of this solver, that we can suppose adapted to such and such actual systems modeling formalism according to their characteristics (but the practical tasks of aid to modeling and to inference algorithms specification are not tackled in this framework). On the other hand, the expression and the computation of diagnoses themselves from the results of the solver come under propositional logic, as the target vocabulary – components normality or abnormality – is propositional.

Definition 1 A *system* is a pair (*SD*, *COMPS*) where *SD*, the *system description*, is a finite set of first-order sentences (with equality) and *COMPS*, the *system components*, is a finite set of constants.

An *observations set OBS* is a finite set of first-order sentences (with equality).

An *observed system* is a triple (*SD*, *COMPS*, *OBS*) where (*SD*, *COMPS*) is a system and *OBS* an observations set.

The elements of *COMPS*, which are the subjects of the diagnosis, appear in *SD* and possibly in *OBS*. The behavioral mode or diagnostic status of each component is represented by a distinguished unary predicate *AB*(.), historically borrowed from the circonscription theory (McCarthy 1986), which is interpreted as signifying *abnormal*. The assumptions about components modes, which determine their behaviors, are thus made explicit in *SD* (nothing forbids *AB*(.) to appear also in *OBS*, but in practice it is always possible to transfer such an occurrence into *SD*). Typically, *SD* formulas code from one side the behavioral models of the generic components (library reusable for any system using the same components), in the form:

```
COMPONENT_TYPE(x) ∧ ¬AB(x) ⇒ Correct_model(x)
    /* correct functioning mode */
COMPONENT_TYPE(x) ∧ AB(x) ⇒
 Fault_model_1(x)∨ ... ∨ Fault_model_n(x) ∨ U(x)
```

```
    /* Fault modes */
¬Correct_model(x) ∨ ¬Fault_model_1(x), ...,
¬Fault_model_1(x) ∨ ¬Fault_model_2(x), ...
    /* exclusion in twos of the different behaviors*/
```

and from the other side the structural description of the system into its components in the form of ground formulas:

```
INVERTER(C1), OR_GATE(C2), =(output(C1), input1(C2))
RESISTOR(C3), =(resistance(C3),150).
```

`Correct_model(x)` is a formula that expresses the normal behavior of component `x`, while `Fault_model_i(x)` is a formula that expresses the behavior of component `x` for the fault mode `i`. The predicate `U(x)` is added to represent the unknown fault mode, thus not accompanied by any model, in order to express that the knowledge of the fault modes cannot claim in general to be exhaustive. It is important to notice that, as the theoretical framework does not assume anything about the nature of the formulas in *SD*, nothing requires modeling faults and one can be satisfied with the only correct functioning models. By the way, it is this idea that prevailed at the origin of model-based diagnosis: show that, unlike all the previous approaches (in particular expert systems) based on the knowledge of the faults and their effects, it was possible to do diagnosis without any prior knowledge of faults and symptoms.

As for the formulas in *OBS*, they describe measurements and are in general ground (but this is not mandatory), for example: `=(port2(C3),2.63)`.

A diagnosis is a mode assignment, normal or abnormal, to each component, which is consistent with both the system description and the observations. According to the context, a diagnosis will be identified either to a subset Δ of components (those that are abnormal) or to a conjunction $D(\Delta)$ of *AB*-literals, where the correspondence between $\Delta \subseteq COMPS$ and $D(\Delta)$ is defined by:

$$D(\Delta) = (\wedge AB(C) | C \in \Delta) \wedge (\wedge \neg AB(C) | C \in COMPS \setminus \Delta).$$

Definition 2 A *diagnosis* for $(SD, COMPS, OBS)$ is a $D(\Delta)$ with $\Delta \subseteq COMPS$ such that: $SD \cup OBS \cup \{D(\Delta)\} \not\models \bot$.

As there are potentially $2^{|COMPS|}$ possible diagnoses, one is often led to apply a parsimony principle and to be interested only in those diagnoses which are *minimal* for set inclusion (the subset of minimal size diagnoses may also be considered, but it is not in general the relevant concept).

Definition 3 A *minimal diagnosis* is a diagnosis $D(\Delta)$ such that $\forall \Delta' \subset \Delta$, $D(\Delta')$ is not a diagnosis.

Remark 1 A diagnosis for $(SD, COMPS, OBS)$ exists if and only if $SD \cup OBS$ is satisfiable, which will be always assumed in the following (otherwise, it means the model has to be revised). $\emptyset$ (i.e., $(\wedge \neg AB(C) | C \in COMPS)$) is a diagnosis (and the only minimal diagnosis) if and only if the observations are consistent with the correct functioning of all the components. Therefore fault detection occurs when $\emptyset$ is no more a diagnosis.

In order to locate the fault(s) after detection, it is natural to be interested in subsets of components – the minimal ones for set inclusion if possible – whose correct modes are by themselves (independently of the modes of the other components) inconsistent with the system model and the observations.

Definition 4 A *conflict set* for $(SD, COMPS, OBS)$ is a set $\mathscr{C} \subseteq COMPS$ such that $SD \cup OBS \cup \{\neg AB(C) | C \in \mathscr{C}\} \models \bot$. A *minimal* conflict set is a conflict set $\mathscr{C}$ such that $\forall \mathscr{C}' \subset \mathscr{C}$, $\mathscr{C}'$ is not a conflict set.

Each conflict set contains thus at least one abnormal component. Consequently a diagnosis Δ must have a nonempty intersection with each conflict set (one can restrain oneself to minimal ones).

Definition 5 Let $\mathscr{K}$ be a sets collection. A *hitting set* for $\mathscr{K}$ is a set $\mathscr{I} \subseteq \cup_{\mathscr{E} \in \mathscr{K}} \mathscr{E}$ such that $\forall \mathscr{E} \in \mathscr{K}$, $\mathscr{I} \cap \mathscr{E} \neq \emptyset$. A *minimal* hitting set is a hitting set $\mathscr{I}$ such that $\forall \mathscr{I}' \subset \mathscr{I}$, $\mathscr{I}'$ is not a hitting set.

Theorem 1 (Characterization of minimal diagnoses) *$\Delta \subseteq COMPS$ is a minimal diagnosis for $(SD, COMPS, OBS)$ if and only if Δ is a minimal hitting set for the collection $\mathscr{K}$ of minimal conflict sets for $(SD, COMPS, OBS)$.*

Theorem 1 provides an operational method for computing minimal diagnoses: one begins by computing all minimal conflict sets, then one computes the minimal hitting sets of the collection obtained in this way. An algorithm has been proposed by Reiter (1987) and corrected by Greiner et al. (1989), based on the construction and pruning of an acyclic direct graph (whose nodes are elements of $\mathscr{K}$ and labels of paths from the root to the leaves are the minimal hitting sets). As for the computation of all minimal conflict sets that involves an unsatisfiability test, it is in all generality a problem which is only semi-decidable; in practice, for real systems models, one deals with decidable fragments but the complexity class is in general NP-hard. An obvious but very inefficient algorithm would be to generate potential conflict sets candidates by a breadth first search of the lattice of subsets of *COMPS*, beginning by *COMPS* (detecting a fault boils down to show that *COMPS* is a conflict set and thus that $\emptyset$ is not a diagnosis), and continue by exploring the subsets of a set each time it has been proved to be a conflict set. This algorithm is improved by coupling conflict sets generation and minimal hitting sets computation: the call to the unsatisfiability checking solver is done at each node of the graph being developed by passing it as argument the conflict set candidate made up of the components that do not appear in the label of the path from the root to the node in question. One takes also advantage of the fact that the solvers (e.g., the resolution-based refutation method) may return,

in case of unsatisfiability, the support of a refutation in the form of a conflict set that is in general strictly included into the conflict set passed as argument, which is used to label the node in question.

Actually, the most popular diagnostic architecture adopted by the majority of real implementations is the GDE (*General Diagnostic Engine*), introduced in De Kleer and Williams (1987) (simultaneously and independently from Reiter 1987). It rests on the coupling of a problem solver and an ATMS (*Assumption-based Truth Maintenance System*). Generally, the solver is based on constraints propagation: it propagates the values provided by *OBS* through the constraints expressing the system model *SD* (such a representation in the form of constraints, in particular equations from physics, is closer from models found in engineering than a first-order logic representation); that way it computes the output values of a component from its behavioral model equations and its input values. In this case the justifications transmitted to the ATMS are Horn clauses and the ATMS handles assumptions (namely the modes $AB(C)$ or $\neg AB(C)$ of each component) management by computing the labels (disjunctions of environments, where each environment is a conjunction of assumptions), supports of each statement inferred by the solver, in particular the *nogoods*, those environments that are the supports of $\bot$, i.e., the inconsistent assumptions sets. The framework of De Kleer and Williams (1987) is limited to the exclusive use of correct functioning modes: in the absence of faults modes, the assumptions are thus all of the type $\neg AB(C)$, which can be simply encoded by the propositional symbol C. In this framework and with this representation of assumptions, one obtains thus an equivalence between *nogoods* and conflict sets.

Property 1 *If only behaviors expressing necessary conditions of correct functioning are modeled in SD and the assumptions $\neg AB(C)$ are coded by the symbols C, then the minimal nogoods computed by an ATMS are exactly the minimal conflict sets.*

Moreover, in the absence of faults modes, one observes that changing, inside a minimal diagnosis Δ, the status $\neg AB(C)$ of a component C in $COMPS \setminus \Delta$ into $AB(C)$ cannot create any inconsistency, as no inference can be done from $AB(C)$. One obtains thus in this case a complete characterization of the set of diagnoses from the set of minimal diagnoses.

Property 2 *If only behaviors expressing necessary conditions of correct functioning are modeled in SD, then any superset of a diagnosis is a diagnosis. The diagnoses are thus exactly all supersets of the minimal diagnoses.*

In general, propagation is not a complete algorithm and one has to resort to more general constraints solvers, which lead to justifications that are no longer necessarily Horn clauses. In this case, and also for the explicit handling of the negation in the assumptions if faults modes are considered, an ATMS is no more sufficient and one has to use a CMS (*Clause Management System*) and to adapt the computation of the hitting sets.

To go further in the characterization of the set of diagnoses in the presence of faults modes, the concept of conflict set has to be generalized. For this, it is beneficial to

move from a set representation of a conflict to a logical representation in the form of a clause, more precisely a *positive AB-clause* (disjonction of positive *AB-literals*).

Remark 2 A conflict set for $(SD, COMPS, OBS)$ identifies with a positive AB-clause $\vee_{C \in COMPS} AB(C)$ entailed by $SD \cup OBS$:

$$SD \cup OBS \models \vee_{C \in COMPS} AB(C).$$

Hence the immediate generalization:

Definition 6 A *conflict* for $(SD, COMPS, OBS)$ is an AB-clause entailed by $SD \cup OBS$, i.e., an AB-clause which is an *implicate* of $SD \cup OBS$. A *positive conflict* is a conflict whose all literals are positive. A *minimal conflict* is a *prime implicate*, i.e., a conflict whose no proper sub-clause is a conflict.

With this definition, the (minimal) conflict sets identify with the (minimal) positive conflicts. Thus the (minimal) hitting sets for the collection of minimal conflict sets identify with the (*prime*) *implicants* of the collection of minimal positive conflicts: one just has to identify the hitting set Δ with the *AB-conjunction* $\wedge_{C \in \Delta} AB(C)$. Moving from the set representation to the logical representation theorem 1 rephrases thus as:

Theorem 2 *$D(\Delta)$ is a minimal diagnosis for $(SD, COMPS, OBS)$ if and only if $\wedge_{C \in \Delta} AB(C)$ is a prime implicant of the collection of positive minimal conflicts for $(SD, COMPS, OBS)$.*

It is important to notice that as and when new observations appear, i.e., the set OBS is growing, the collection of positive conflicts increases as well and as a result some prime implicants do not remain any more in general. That is to say that some minimal diagnoses disappear and are replaced by other ones (involving more abnormal components). This means that the diagnostic process is *non-monotonic* as a function of the observations. This non-monotony is essential and actually it exists a close relationship between the diagnosis theory and the *default logic* (see chapter "Knowledge Representation: Modalities, Conditionals, and Nonmonotonic Reasoning" of this volume): one expresses that the components are correct in the form of (normal) defaults and one obtains a bijection between minimal diagnoses and extensions of the default theory built in this way.

Property 3 *Let $(SD, COMPS, OBS)$ be an observed system. Let DT be the following default theory: $DT = (\{: \neg AB(C)/\neg AB(C) | C \in COMPS\}, SD \cup OBS)$. Then E is an extension of DT if and only if $E = \{\pi | SD \cup OBS \cup D(\Delta) \models \pi\}$ where $D(\Delta)$ is a minimal diagnosis for $(SD, COMPS, OBS)$.*

The logical generalization of the concept of conflict allows one to characterize the set of all diagnoses, and not only of minimal diagnoses. One begins by defining a compact representation of the diagnoses, by considering the partial modes assignments to part of the components, such that all their extensions (by the modes normal or abnormal indifferently) to the rest of the components are diagnoses.

Definition 7 A *partial diagnosis* for $(SD, COMPS, OBS)$ is a satisfiable conjunction P of AB-literals such that, for any satisfiable conjunction P' of AB-literals containing P as a sub-conjunction, $SD \cup OBS \cup \{P'\} \not\models \bot$. A *kernel diagnosis* is a minimal partial diagnosis, i.e., none of its proper sub-conjunctions is a partial diagnosis.

With this definition, the kernel diagnoses provide a compact representation of all the diagnoses, these ones being exactly the total extensions of the kernel diagnoses.

Property 4 (Characterization of the diagnoses) *$D(\Delta)$ is a diagnosis if and only if it exists a sub-conjunction of $D(\Delta)$ which is a kernel diagnosis.*

Theorem 2 that characterizes the minimal diagnoses in terms of the positive conflicts is generalized as a characterization of the kernel diagnoses (and thus of all the diagnoses) in terms of the conflicts.

Theorem 3 (Characterization of the partial and kernel diagnoses) *The partial diagnoses (resp. kernel diagnoses) for $(SD, COMPS, OBS)$ are the implicants (resp. prime implicants) of the collection of minimal conflicts for $(SD, COMPS, OBS)$.*

Note that this theorem shows that the collection (in the disjunctive sense) of the kernel diagnoses, as a disjunctive normal form, is analogous to the collection (in the conjunctive sense) of the minimal conflicts, as a conjunctive normal form.

A sufficient condition guaranteeing that any superset of a diagnosis is a diagnosis has been given by the Property 2. Theorems 1 and 3 allow one to clarify the relationship between this property of closure of the diagnoses collection by the superset operation, and thus the complete characterization of diagnoses in terms of minimal diagnoses, and the nature of the conflicts.

Property 5 *There is a one-to-one correspondence between the kernel diagnoses and the minimal diagnoses (by extending any kernel diagnosis by the normal mode of all the components that it does not contain) if and only if all minimal conflicts are positive. More precisely, the two following statements are equivalent:*

1. *any superset of a minimal diagnosis is a diagnosis, i.e., if $D(\Delta)$ is a minimal diagnosis then $\forall \Delta'$ such that $\Delta \subseteq \Delta' \subseteq COMPS$, $D(\Delta')$ is a diagnosis;*
2. *all minimal conflicts for $(SD, COMPS, OBS)$ are positive.*

Unfortunately one does not know an equivalent of the second statement of this property in terms of a syntactic characterization of $SD \cup OBS$. Only sufficient conditions guaranteeing the positivity of the minimal conflicts do exist, in the form of restrictions on $SD \cup OBS$. The most obvious one is to impose that any occurrence of an AB-literal in $SD \cup OBS$, put in conjunctive normal form, be positive. It is satisfied as soon as only the correct behavior of components is modeled, in the form of necessary conditions, which is the assumption of the Property 2.

Let add that in practice one limits oneself to compute the *preferred* diagnoses, according to a given criterion. It can be for example a *probabilistic* criterion if *prior* probabilities of the components behavioral modes are available. Diagnoses can

thus be generated in decreasing probability rank by using Bayes rule for evaluating conditional probabilities after each observation. One can use quantitative probabilities but also content oneself with relative orders of magnitude between probabilities. It can be also an *explanatory* criterion (see Sect. 2.2). Most of the time the preferred diagnoses are minimal and the selection according to the chosen preference criterion is thus done among minimal diagnoses, even for a model for which it is known that minimal diagnoses are not enough to characterize all diagnoses.

2.2 Abductive Approach

Graphs Based Approach

The very first approaches for diagnosis relied on causal models (see chapter "A Glance at Causality Theories for Artificial Intelligence" of this volume) representing in the form of arcs the causal relationships between the faults situations (D, for defects) that could affect the system and their effects, in particular their observable ones (M, for manifestations). Among these works, one can quote those from Reggia et al. (1983), Peng and Reggia (1990) which propose to use the covering sets theory to characterize the diagnoses. Arcs and nodes are associated to conditional probabilities and a plausibility measure is computed to rank the diagnoses.

Abductive Logical Approach

A limitation of the diagnosis approaches that are exclusively abductive is that they suppose *a priori* the "completeness" of the causal model, which has to describe all the faults and all the manifestations of these faults. An attempt to overcome this limitation is to take into account uncertain causal relationships by distinguishing strong causal link and weak causal link. Another one is, after having analyzed the differences between abductive and consistency-based approaches (Poole 1989), to try to reconcile them (Console and Torasso 1990). The idea is to distinguish among the observations those that the model has to explain (for example, the abnormal observations) from those whose only the consistency with the model is required (for example the exogenous or normal observations). It is examined in the papers (Console and Torasso 1991; Ten Teije and Van Harmelen 1994) which propose a synthesis of the various definitions that may result from it. This can be expressed in the same logical framework than previously by a logical diagnosis theory extending consistency-based diagnosis by abductive diagnosis. Similarly to Sect. 2.1, the following definitions, properties and theorems are obtained.

Definition 8 Let $(SD, COMPS, OBS)$ be an observed system and $OBS = I \cup O$ a partition of OBS, where O are those observations one wants to explain. An *abductive diagnosis* for $(SD, COMPS, I \cup O)$ is a $D(\Delta)$ with $\Delta \subseteq COMPS$ such that: $SD \cup I \cup \{D(\Delta)\} \not\models \bot$ and $SD \cup I \cup \{D(\Delta)\} \models O$.

Definition 9 A *partial abductive diagnosis* for $(SD, COMPS, I \cup O)$ is a satisfiable conjunction P of AB-literals such that, for any satisfiable conjunction P' of AB-literals containing P as a sub-conjunction, $SD \cup I \cup \{P'\} \not\models \bot$ and $SD \cup I \cup \{P'\} \models O$. A *kernel abductive diagnosis* is a minimal partial abductive diagnosis, i.e., such that none of its proper sub-disjunctions is a partial abductive diagnosis.

Property 6 (Characterization of the abductive diagnoses) *$D(\Delta)$ is an abductive diagnosis if and only if it exists a sub-conjunction of $D(\Delta)$ which is a kernel abductive diagnosis.*

Theorem 4 (Characterization of the kernel abductive diagnoses) *Assume that SD, I and O are finite sets of formulas (each one being thus represented by a unique formula resulting from the conjunction of its elements). The kernel abductive diagnoses for $(SD, COMPS, I \cup O)$ are the prime implicants of $\Pi \wedge \{(SD \wedge I) \Rightarrow O\}$, where Π is the conjunction of the minimal conflicts for $(SD, COMPS, I \cup O)$.*

Notice that the logical concept of observations entailment used by the abductive diagnosis is unsuitable as soon as the observations are more precise than the predictions made from the models: one has in this case to resort to an abstraction of the observations (Cordier 1998; Besnard and Cordier 1994), represented by an observations lattice, and extend the definition of abductive diagnosis to that of explanatory diagnosis (explaining at best the observations).

2.3 Extensions

After having presented the formal framework of logical diagnosis, we quote rapidly below the issues that gave rise to later works.

When the number of diagnosis candidates is too large, it is important to use preference criteria to rank them. It is thus possible to generate the most probable diagnoses, from the *prior* faults probabilities (possibly qualitative) and use of Bayes rule (De Kleer 1992, 2006). One may also turn towards the sequential diagnosis, which consists in taking advantage of a succession of observations for reducing gradually the number of diagnoses. Some works had for purpose the choice of the best (in the sense of information theory, i.e., minimizing an entropy function) next observation in the framework of the sequential diagnosis. This issue meets the one of active testing (Feldman et al. 2009; Siddiqi and Huang 2010).

The diagnosis definitions and particularly the preferences (such as the probabilities) used to rank diagnoses are based in general on the assumption of faults independence. Some works are interested in the case of dependent faults such as cascading faults. A category of faults particularly difficult to diagnose is made up of faults affecting the structure (connectivity) of the system. Appear in this category the shortcuts between connections of a printed circuit board that result in hidden interactions (because not taken into account *a priori* in the model).

Rather early, when the application of the theory to real cases has been undertaken, arose the problem of handling uncertainty, both at the level of the model and at the level of the observations. It is especially important as the theory of consistency-based diagnosis only detects and makes explicit the causes of an inconsistency between the model of the system and the real system: inferring from that a malfunction of the system rests thus entirely on the correction of the model. Uncertainty is generally handled by resorting to an abstraction (Torta and Torasso 2003; Chittaro and Ranon 2004), or by qualitative models (that come under another important field in AI, the qualitative reasoning (see chapter "Qualitative Reasoning about Time and Space" of this volume)), or by expressing the values of the model parameters and of the observations by numerical intervals. According to the case, qualitative simulation or interval-based CSP are used as solvers (Dague et al. 1990).

Two research issues that emerged only at a later stage after the seminal works in the domain and are among the most active presently are diagnosability analysis and decentralized diagnostic architectures. The first, diagnosability, appeared around twenty years ago, arises from the assessment that the problem of designing and deploying a diagnostic architecture for a system must be tackled in advance at the very moment of the system design and not once the system has been produced and choices critical for the diagnosis, such as the number and the location of the sensors and thus the observation capacity of the system, have been fixed. For a given set of anticipated faults modeled in addition of the correct functioning of the system and a given set of observable quantities or events, the diagnosability analysis of the model answers the question to know if any occurrence of one of the faults will be always unambiguously identifiable in a finite time thanks to the observations only. Research in the field focused mainly on discrete event systems, modeled by transitions systems such as automata or Petri nets (see Sect. 3.6).

The second, more recent, concerns the diagnostic architectures either decentralized (local diagnosers communicating with a diagnostic supervisor in charge of providing the global diagnosis) or distributed (local diagnosers communicating between them for finding the global diagnosis), essential in particular for diagnosing systems that are by nature distributed (peer-to-peer networks, composite web services, etc.) but also systems made up of proprietary subsystems whose models are private for confidentiality reasons. Distribution may be related to the model, the observations, the algorithms, the software and hardware diagnostic architecture. Such architectures are presented in the case of discrete event systems in the Sect. 3.5.

Among the important problems, one can quote the preventive diagnosis, which consists in being able to detect a problem to come, before it occurs. This issue received attention later, probably because of the difficulty to get predictive models (such as wear models). Approaches different from model-based ones will have probably to be used in this case.

The issue of a tight coupling between diagnosis and repair or reconfiguration, critical in particular for autonomous systems, has been studied by using planning techniques (Sun and Weld 1993; Nejdl and Bachmayer 1993; Friedrich et al. 1994).

A last, important and difficult, problem is taking time into account. It is presented in the Sect. 3.1 and illustrated by the discrete event systems in the Sect. 3.2.

3 Diagnosis of Discrete Event Systems

3.1 Temporal Representation and Diagnosis

The previous section presents a theory that does not handle the representation of time and temporal reasoning. From this theory some extensions have been proposed that deal with several dimensions about time. Brusoni et al. propose in their paper *A spectrum of definitions for temporal model-based diagnosis* (Brusoni et al. 1998) a classification of these different extensions which take into account situations and successive observations as *time-varying contexts*; the system can also evolve between the production of two sets of observations (it is a *time-varying behavior*); faults can also produce observable effects after a given finite duration that can be represented as causal graphs (*temporal behavior*). Most of the time, these extensions can be represented by adding a time variable in the *SD* formulas associated with time constraints. Time is therefore reified. In practice, given a representation of the problem, it is necessary to look for compatible solvers that can manage inference and consistency tests by dealing with the selected representation of time (continuous, discrete or even both in hybrid systems).

Time variations in physical systems that are only due to system inputs do not add any new difficulty as this case can be interpreted as a discretized sequence of statical diagnosis problems. However, most of the systems are actually dynamical, they have internal states that memorize the past so that the behavior of the system not only depends on its current inputs but also on its current state. Time can be represented in a discretized way as a sequence of instantaneous events, in this case, the system is modeled as a discrete event system (see Sect. 3.2). Time can also be seen as a continuous variable that is described in differential equations, typically studied by the control theory community (FDI, see Sect. 4): AI and FDI methods have actually been compared (see Sect. 4). Based on the time granularity that is chosen in a model, continuous time can be symbolically abstracted as a set of instants that can be partially ordered, as time intervals, or as sequences of dates. In this last case, if the space of physical quantities is discrete, a concise representation of the temporal behavior can be done as a set of episodes, otherwise sequence of numerical intervals can be used. Some ATMS extensions are proposed to efficiently deal with these time data structures. It is possible to use the generic diagnosis theory that is described above by using an explicit variable that encodes time. However, the complexity of the model to acquire and the complexity of the inference and consistency test algorithms drastically increase. This theoretical framework can still be applied as long as the faults within the system are permanent (always present). If faults occur at the supervision time and if their effect is permanent after their occurrence, there is no fundamental changes as the evolution of the conflict sets is still monotonic. Dealing with intermittent faults is more difficult and is possible only if the evolution of such intermittent faults is slower than the evolution of the system itself and the speed of observation acquisition.

Most of the contributions, even the ones dealing with time, aim at solving the diagnosis problem based on observation logs after the system has stopped: this is the *off-line diagnosis* problem. Then the AI and FDI communities independently started to develop some works about *on-line diagnosis*. The system is observed at operating time in order to react (repair, control) when a discrepancy with the expected behavior is detected and maintain an operating state that is as satisfactory as possible.

Two types of methods can be distinguished.

1. In *chronicle recognition* methods, the objective is to recognize, within the flow of observations, some observable patterns that characterize faulty situations. A chronicle is a set of events associated with time constraints. Specialized algorithms perform on-line chronicle recognition so that a decision about how to react after a fault has been diagnosed can still be made at operating time (Dousson 1996; Carle et al. 2011).
2. The second type of methods, that is typically model-based, relies on the behavioral description of the system but has to deal with the on-line observation flow incrementally. Assuming that only one fault has occurred or is permanent within the supervision time is not realistic as the supervision time is long. Moreover it must be considered that some faults are repaired during the supervision.

In the next sub-section, we focus on the methods where the system is modeled as a discrete event system (DES). This type of models is particularly relevant when the underlying system reacts to events (reactive systems), such that the opening/closing of a valve, the reception of messages, the occurrence of a fault. This type of models can also be relevant even if the system is continuous but can be discretized as a DES (Lunze 1994). From the initial work from Sampath et al. (1996), a set of contributions are proposed about the diagnosis of discrete event systems in the AI community as well as in the FDI community.

3.2 Models of Discrete Event Systems

A DES is a dynamical system whose state can be described by state variables and the domain of each variable is discrete. The behavior of the DES is characterized by the occurrence of discrete events that instantaneously modify the internal state of the DES. This representation is obviously well-suited to describe systems that are naturally discrete, such as communication networks that aim at receiving, sending messages, automated production line systems that produce objects step by step, etc. But this representation is also well-suited for systems that can be discretized, resulting for example from a qualitative reasoning method (Travé-Massuyès and Dague 2003).

To model DES, several formalisms from the language theory can be used such as the process algebra, Petri nets and automata. With the help of these formalisms, the behavioral language of the DES can be represented in a concise manner. In order to present and illustrate the diagnosis problem of DES, we use here the

formalism of *transition system/automaton* (see chapter "Theoretical Computer Science: Computational Complexity" of Volume 3) which has been used in most of the seminal works of the field.

Definition 10 An *automaton A* is a 5-tuple $\langle Q, E, T, I, F\rangle$ such that

- Q is a finite set of states,
- E is a finite set of events,
- $T \subseteq Q \times E \times Q$ is a finite set of transitions $\langle q, e, q'\rangle$,
- $I \subseteq Q$ is a set of initial states,
- $F \subseteq Q$ is a set of final states.

The event e over the transition $t = \langle q, e, q'\rangle$ triggers the transition. The language $L(A) \subseteq E^\star$ generated by the automaton A is the set of event sequences from E which can be associated with a transition path in A from an initial state q_0 of I to a final state of F, such a path is also called a *trajectory*.

Definition 11 A *trajectory* of an automaton $A = \langle Q, E, T, I, F\rangle$ is a sequence of transitions $traj = q_0 \xrightarrow{e_1} \ldots \xrightarrow{e_m} q_m$ such that: $q_0 \in I$, $q_m \in F$, and $\forall i \in \{1, \ldots, m\}$, $\langle q_{i-1}, e_i, q_i\rangle \in T$. A trajectory can also be denoted as $\langle (q_0, \ldots, q_m), (e_1, \ldots, e_m)\rangle$.

The set of possible behaviors of a system is represented as an automaton *SD*, each behavior being characterized as a trajectory in *SD*.

Definition 12 The *model of the system* is an automaton

$$SD = \langle Q^{SD}, E^{SD}, T^{SD}, I^{SD}, F^{SD}\rangle.$$

As any trajectory $q_0 \xrightarrow{e_1} \ldots \xrightarrow{e_m} q_m$ of the system depends on a previous trajectory of the system $q_0 \xrightarrow{e_1} \ldots \xrightarrow{e_{m-1}} q_{m-1}$, the *SD* automaton can then be such that $F^{SD} = Q^{SD}$ (any state is final). In other words, the language $L(SD)$ is prefix-closed.

In general, a DES can be modeled in a modular way as a set of n components $COMPS = \{C_1, \ldots, C_n\}$ that define the *structural model* of the supervised system. Each component C_i is modeled as an automaton $SD_i = (Q_i, E_i, T_i, I_i, F_i)$. The model of the system is obtained by applying a *synchronized product* on the automata $(SD_i)_{i=\{1,\ldots,n\}}$. The product relies on a set of *synchronisation relations Sync* that are generally a set of constraints $e_i = e_j$ that model the fact that the event e_i of C_i and the event e_j of C_j must always occur at the same time. The global model *SD* is obtained by computing the subset of trajectories from the Cartesian product $\prod_{i=1}^{n} SD_i$ that is restricted to the trajectories when all the constraints of *Sync* are satisfied. This synchronized product is denoted $\otimes_{Sync}$ or simply $\otimes$ when the synchronisation constraints are defined without ambiguity. From this, it follows:

$$SD = SD_1 \otimes \cdots \otimes SD_n.$$

3.3 Faults, Observations and Diagnosis of DES

The automaton *SD* that represents the system, actually models its normal and abnormal behaviors, and especially the behaviors of interest in the monitoring task. The abnormal behaviors are modeled by labeling transitions with *fault events* $e_f \in F \subseteq E^{SD}$ that represent the fact that the system starts to be faulty.

Any diagnosis reasoning requires the observation of the system. In the context of the DES, observations are events, usually resulting from the generation of a piece of information from sensors. In a DES, there are observable events $E^{SD}_{OBS} \subseteq E^{SD}$ and non-observable events $E^{SD}_{\neg OBS} \subseteq E^{SD}$. Among the non-observable events, there are the fault events. Any trajectory τ of the system is then associated with its *observable trace* $\sigma(\tau)$ that is defined as the sequence of observable events that is produced when τ is indeed the trajectory realized by the system (projection of τ on the observable events E^{SD}_{OBS}).

If it is assumed that the observations of the system are perfectly known (no uncertainty about the observed event types and the observed dates), the observation of the system is then defined as *a sequence of observable events*.

Definition 13 The *observation* of the system, denoted *OBS*, is the sequence of observable events that is produced by the system within the time frame of the diagnosis reasoning.

The diagnosis task consists in comparing the effective observation of the system with the prediction of the model as the possible set of observable traces, and then to determine the set of non-observable events (especially the fault events) that explain the current state of the system (Cordier and Thiébaux 1994).

Definition 14 A *diagnosis problem* is described as a 3-uple (SD, OBS, F) where *SD* is the model of the system, *OBS* is the observation of the system and *F* is a set of fault events.

In order to determine the faults, it is firstly necessary to search for the set of system's trajectories in the model *SD* whose observable trace matches *OBS* exactly.

Definition 15 (*Trajectory Diagnosis*) A diagnosis Δ for the problem (SD, OBS, F) is a trajectory of *SD* whose observable trace $\sigma(\Delta)$ is exactly *OBS*.

With this definition, the diagnosis problem does not depend on faults (it can be defined as a couple (SD, OBS)). However, the diagnosis can also be defined in a more concise way as a set of faults. This second definition is closely related to the one for statical systems.

Definition 16 (*Fault Diagnosis*) A diagnosis Δ of the problem (SD, OBS, F) is a set of faults $\Delta \subseteq F$ such that there exists a trajectory τ from *SD* that exactly contains the set of fault events Δ and its observable trace $\sigma(\tau)$ is exactly *OBS*.

It can be noticed that the set of trajectory diagnoses of a system can also be represented as an automaton, more precisely it is a sub-automaton of *SD*, each trajectory in it has an observable trace that is exactly *OBS*.

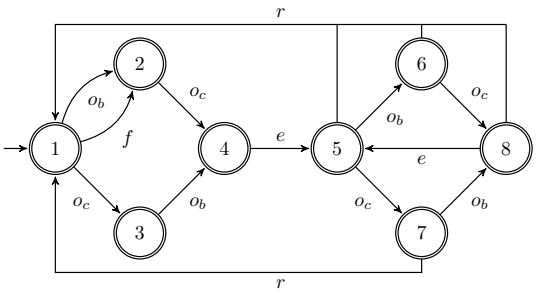


Fig. 1 Model of the system, o_b, o_c, r are the observable events

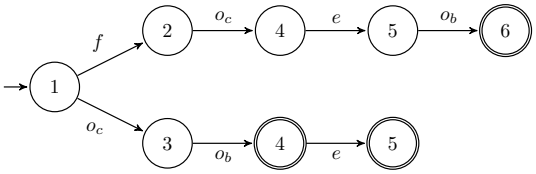


Fig. 2 Diagnoses of the system (Fig. 1) given the observed sequence $OBS_1 = (o_c, o_b)$

Example 1 Figure 1 illustrates a system with a set of observable events o_b, o_c and r. If the observed sequence is $OBS_1 = (o_c, o_b)$, the set of diagnoses are the ones presented as an automaton in Fig. 2, the fault f is not certain and the possible states of the system are 4, 5, 6. If the observed sequence is $OBS_2 = (o_c, o_c)$, the unique diagnosis is presented in Fig. 3: the occurrence of the fault f is indeed certain and the unique possible state is 7.

An observation *OBS* consisting of a sequence of observed events can also be represented as an automaton with one initial state and one final state. The diagnosis can then be computed by performing a *synchronized product* $\otimes$ between the automaton *SD* and the one that describes *OBS*. The synchronization constraints *Sync* are applied on the observable events: an observable event o must occur in *SD* and in *OBS* in the same order. Representing *OBS* this way is interesting as it can be extended to represent uncertain observations. In this case the automaton *OBS* does not represent one sequence of observable events only but several possible sequences (Grastien et al. 2005). From this follows the next theorem:

Theorem 5 *The automaton* $SD \otimes OBS$ *describes the set of trajectory diagnoses from the problem* (SD, OBS).

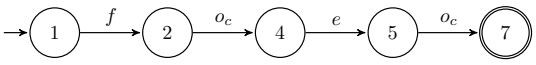


Fig. 3 Diagnosis of the system (Fig. 1) given the observed sequence $OBS_2 = (o_c, o_c)$

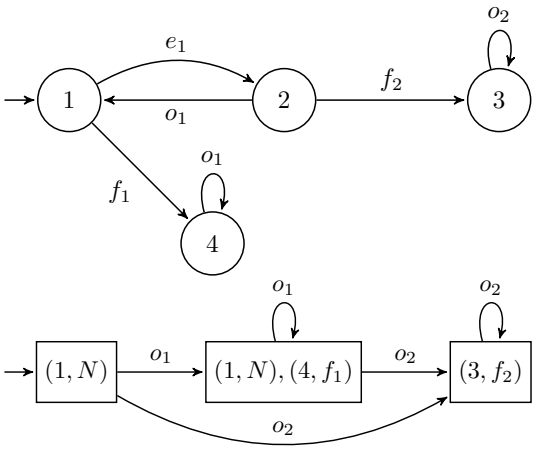


Fig. 4 The global model *SD* (with $E^{SD}_{OBS} = \{o_1, o_2\}$ and $F = \{f_1, f_2\}$) (Top) and its diagnoser (Down). Label *N* (normal) means the absence of any fault

3.4 Diagnoser Approach and Other Centralized Approaches

One of the seminal works to compute diagnosis on DES is in Sampath et al. (1996) and is based on the computation of a *diagnoser* (see Fig. 4). A diagnoser is a deterministic automaton that describes the set of observable behaviors of the system in a similar way as an observer would do. It is built by ε-reduction from the automaton *SD* where ε represents any non-observable event of *SD*. A diagnoser transition is labeled with an observable event. A state of a diagnoser describes the set of states of *SD* that are reachable from its initial states and that are reachable by trajectories that produce the observable sequence. Associated with each state of *SD* the diagnoser state also records sets of fault events that have occurred on such trajectories. For a given sequence of observable events, the diagnoser state thus describes the set of possible reached states and the set of possible faults that have occurred before reaching one of these states.

The diagnoser is a finite state machine that results from the off-line compilation of the diagnosis problem and its use for on-line diagnosis is performed by a simple algorithm. Indeed, the on-line algorithm consists in triggering the observed events of *OBS* in sequence and the result of the algorithm is contained in the diagnoser state that is reached. The problem of this method is about the time/space complexity of the computation of the diagnoser. In Marchand and Rozé (2002), Schumann et al. (2004), other computation methods have been proposed to improve the efficiency on average of the diagnoser computation. These methods rely on binary decision diagrams (*BDD*).

Other methods use different formalisms to build an equivalent diagnoser such as communicating automata, Petri nets, process algebra (Rozé and Cordier 2002; Jiroveanu and Boel 2006; Console et al. 2002). Other works (Lamperti and Zanella 2003) propose specialized data structures and specific algorithms to solve the diagnosis problem. On the other hand, Grastien and Anbulagan (2013) propose the use

of generic SAT techniques and translate the diagnosis problem into a succession of propositional formulas (CNF). It is also possible to use probabilistic models that can model the likelihood of transitions between states. One preference criteria is then to keep the transitions that are the most probable, this can be done for instance by applying the Viterbi algorithm such as in Aghasaryan et al. (1997).

Three extensions of the classical diagnosis problem have been mainly investigated. In the first one, the hypothesis that *OBS* is certain is removed (Lamperti and Zanella 2003; Grastien et al. 2005). It is, in this case, impossible to assert that there is a unique sequence of observations, either because the knowledge about the real order of the observed events is not perfect or because events can be lost or corrupted (noise). This might be due to imprecise or even faulty sensors, or the communication network between the sensors and the diagnoser. One solution then consists in representing the observations as an automaton that contains the set of possible observed trajectories. Then Theorem 5 can be used as in Grastien et al. (2005). The second extension of the problem is about on-line diagnosis that is well-suited for the on-line monitoring of dynamical systems such as communication networks. In this context, *OBS* is partly known (a prefix of *OBS* is known). On-line diagnosis then leads to incremental diagnosis that consists in updating the diagnosis from a previous diagnosis in a new time window when new observed events are available (Pencolé and Cordier 2005; Grastien et al. 2005). Incremental diagnosis has also be extended to deal with large scale systems where it is not possible to efficiently update the diagnosis with the flow of observations. In Su and Grastien (2013), the principle is to compute a diagnosis for a given time window independently from any other time window and Su et al. (2014) analyses the minimal amount of information to retain between time window to assert the diagnosis is correct along the time. Finally, a more recent extension is about the diagnosis of behavioral patterns (Jéron et al. 2006; Pencolé and Subias 2018). In the classical problem, the model represents faults as the occurrence of single events. With behavioral patterns, it is also possible to represent in the model a set of events that might not be considered independently as faulty but some specific ordering of their occurrence can still be abnormal (for instance, in traffic light systems, the sequence of *green, yellow, red* is normal while *green, red, yellow* is not).

3.5 Distributed and Decentralized Approaches

Most of the systems that are monitored and diagnosed have a large size so that a centralized method, as described in the previous sub-section, is not efficient enough. To illustrate this inefficiency, it can be noticed that the synchronized product over the components'models is in $O(2^n)$ where n is the number of components. This complexity makes a centralized approach impossible to implement on a realistic system. Based on the distributed nature of a system as a network of components, it is then possible to design decentralized or even distributed diagnosis methods that are more scalable. The model is then described as a set of components'models

and a set of connections and the global model is not explicitly computed. Several formalisms have been proposed to model the system in a distributed way: automata where the connections are represented by shared events, communicating automata where the connections are represented by messages on input/output ports (Pencolé and Cordier 2005), Petri nets where interactions are modeled by shared transitions or places (Fabre et al. 2005; Jiroveanu and Boel 2006), process algebra (Console et al. 2002) where the synchronization is represented as a cooperation operator.

There exist several methods implementing the collaboration of several local diagnosers to solve a diagnosis problem. Several types of methods can be distinguished depending on the supervision architecture. In a so-called *coordinated* architecture, each diagnoser is in charge of observing local sites and determining a global diagnosis based on its observation sites. Then a coordinator analyzes the global diagnoses of each diagnoser and provides a unique and coordinated one (Debouk et al. 2002). In this type of architecture, local diagnosers must still know the global model of the system so such an architecture has a scalability issue. A second architecture where the local diagnosers do not need to know the global model is the *decentralized* architecture. As opposed to the coordinated architecture, local diagnosers only have a local knowledge about the system (a subset of components, also called a cluster). Local diagnosers perform diagnosis only over the components they know. The local diagnoses, once computed by the local diagnosers, are sent to a global diagnoser that is in charge of checking the global consistency of the local diagnoses (Pencolé and Cordier 2005; Lamperti and Zanella 2003; Grastien et al. 2005; Pencolé et al. 2018) by checking whether the local diagnosed trajectories are globally synchonizable (Fig. 5). The last investigated architecture is the *distributed* architecture. The main difference with the decentralized architecture is that there is no global diagnoser. The result of the diagnosis is not global but only local. Each local diagnoser is in charge of handling the global consistency of its diagnosis by interacting with other local diagnosers (Fabre et al. 2005). The diagnosis algorithms then depend on the selected architecture. Computing a global diagnosis might be a necessity to decide about a global repair or a global reconfiguration of the system, in this case, coordinated or decentralized methods should be used. If the decision is local then a distributed architecture is sufficient.

In the case of distributed or decentralized architectures, the complexity of the algorithms mainly depends on checking the global consistency of the local diagnoses that depends on the number of involved components. To improve this global consistency checking, a BDD-based synchronization algorithm is proposed in Schumann et al. (2010). Another way to increase the average efficiency of the algorithms is to analyze off-line the structural model of the system (the topology) to precompile basic synchronization strategies that can be then applied on-line. For instance in KanJohn and Grastien (2008), this analysis is based on junction trees. In Pencolé et al. (2006), the analysis consists in determining off-line clusters of components based on which a local diagnoser is always *accurate*: it is certain to always have a global consistent diagnosis without any synchronization with other components out of the given cluster.

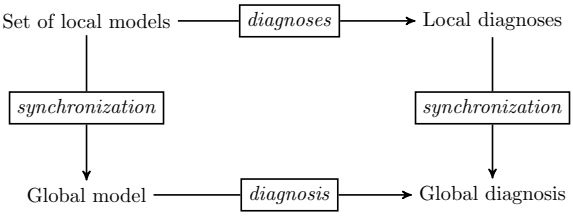


Fig. 5 Principle of a decentralized diagnosis

3.6 Diagnosability

The off-line analyses of DES properties related to the diagnosis problem is essential to implement efficient on-line diagnosis algorithms (such properties like diagnosis accuracy of clusters, topology properties as cited in the sub-section above). Among these properties, *diagnosability* is the most studied one (see Sect. 2.3).

Intuitively, in the context of DES, a system is diagnosable if, in case of an ambiguous diagnosis (faulty or not) at a given time, it is always sufficient to wait for a new finite set of observations to refine the diagnosis and prune the ambiguity and obtain a diagnosis that is certain. The first formal definitions of this property are proposed in Sampath et al. (1995). Several extensions have then been defined, by considering intermittent faults (Contant et al. 2004), or by extending faults to behavioral patterns (Jéron et al. 2006; Gougam et al. 2017; Ye and Dague 2017). A definition that unifies the ones for continuous systems and the DES is proposed in Cordier et al. (2006). Checking the diagnosability is a complex problem. The first solution is presented in Sampath et al. (1995) and consists in checking in the diagnoser (see Fig. 4) whether there is no indeterminate cycles (cycles of states where the diagnosis is ambiguous). Then another solution that is polynomial in the number of states in the global model is based on the synchronization of the model with itself (some *twin models*) where the synchronizations are performed on the observable events only. It consists in checking in this product for infinite sequence of critical pairs (a sequence that represents a faulty sequence in one twin and a non-faulty sequence in the other one) (Jiang et al. 2001; Yoo and Lafortune 2002). Other algorithms improving the efficiency of this method can also be found in Cimatti et al. (2003), Schumann and Pencolé (2007).

One of the objectives of checking diagnosability is to provide a feedback to the design of the system, by essentially adding new sensors (Travé-Massuyès et al. 2001; Ribot et al. 2008), or by respecifying communication protocols between components of the system (Pencolé and Cordier 2005). Some extensions about the diagnosability of distributed systems are also proposed in Provan (2002), Pencolé (2004), Ye and Dague (2012). One method also extends the diagnosability problem to deal with uncertain observations (Su et al. 2016). Another extension is about *self-healability* that combines diagnosability and repairability. A system is said to be self-healing if it is able to perceive its own faults and, without any human intervention to perform necessary actions to recover. Self-healability can hold in a system even if the system

is not fully diagnosable and not fully repairable. The required level of diagnosability is the one that can always make the repair decision certain. In (Cordier et al. 2007), this level of diagnosability is based on a selection of *macrofaults* that are diagnosable and repairable.

4 Bridge Between Model-Based Diagnosis Rooted in AI and in Automatic Control

In the field of DES, the AI community (known as DX community) and the Automatic Control community (known as FDI – *Fault Detection and Isolation*–community) have converged from the start on the same formalisms and jointly developed diagnosis methods. On the contrary, for continuous systems, these communities have worked in parallel for a long time, ignoring their respective results. Although there are common principles, each community has developed its own concepts and methods, guided by different modeling approaches, and relying on analytical models and linear algebra for the first and on logical formalisms for the latter. However, in the 2000s, under the impetus of the BRIDGE group "*Bridging AI and Control Engineering model based diagnosis approaches*" within the Network of Excellence MONET II and its French counterpart, the IMALAIA group "*Integration of Methods Combining Automatic Control and AI*" linked to GDR I3, the French Association for Artificial Intelligence AFIA, and GDR MACS, an increasing number of researchers from these two communities have sought to understand and integrate approaches of their respective fields to provide more effective diagnostic systems (Travé-Massuyès 2014).

First of all, we draw up a panorama of the approaches proposed by the FDI community, and then present a comparative analysis of the concepts and techniques used in the two communities in Sect. 4.2, followed by the works which integrate techniques of both communities in Sect. 4.3.

4.1 FDI Community and Approaches for Continuous Systems: Quick Panorama

Like the methods of the DX community (cf. Sect. 2), the fault detection and diagnosis methods of the FDI community are based on behavioral models that establish the constraints between the system inputs and outputs, i.e., the set of measured variables Z, as well as the internal states, i.e., the set of unknown variables X. The variables $z \in Z$ and the variables $x \in X$ are functions of time. These models are formulated either in the time domain (then known as *state space models*) or in the frequency domain (then known as *transfer functions* in the linear case).

The books (Gertler 1998; Blanke et al. 2003, 2015; Dubuisson 2001) are very good reviews that include the references to original papers, to which the reader can refer.

The central concept of FDI methods is that of *residual* and one of the main problems is the *generation of residuals*. Consider the model of a system under the form of a set of differential and/or algebraic equations $SM(z, x)$ with variables z and x. $SM(z, x)$ is said to be *consistent with the observed trajectory* z, if there exists a trajectory of x such that the equations of $SM(z, x)$ are satisfied.

Definition 17 (*Residual generator for* $SM(z, x)$) A system that takes as input a subset of measured variables $\tilde{Z} \subseteq Z$ and generates as output a scalar r, is a residual generator for the model $SM(z, x)$ if for all z consistent with $SM(z, x)$, $\lim_{t\to\infty} r(t) = 0$.

When the measurements are consistent with the system model, the residuals tend to zero as t tends to infinity, otherwise some residuals may be different from zero. Evaluating the residuals and assigning them a Boolean value – 0 or non-0 – requires statistical tests that account for the statistical characteristics of noises (Dubuisson 2001; Gao et al. 2015). There are three main families of methods for generating residuals.

- The *methods based on testable relations* rely on unknown variables elimination. These methods generate residuals from relations inferred from the model which only involve measured variables and their derivatives. These relations are called *Analytical Redundancy Relations* (ARRs). For linear systems, the so-called *parity space* approach is used to eliminate unknown state variables and obtain ARRs by projection onto a particular space called the parity space (Chow and Willsky 1984). Extensions of this approach to nonlinear systems have been proposed (Staroswiecki and Comtet-Varga 2001). The structural approach (Armengol et al. 2009) allows one to obtain the just determined equation sets of a model from which ARRs can be inferred (Krysander et al. 2008).
- The *methods based on state estimation* are based on estimating unknown variables. They take the form of *observers* or optimized *filters*, such that the Kalman filter, and provide an estimation of the state of the system and its outputs. Numerous diagnosis solutions rely on state estimation, particularly for hybrid systems (cf. Sect. 4.3). In this case, the continuous state is augmented by a discrete state that corresponds to the operation mode (normal or faulty) of the system components.
- The *methods based on parameter estimation* focus on the value of the parameters which directly represent physical characteristics. Fault detection is performed by comparing the estimated value of the parameters to their nominal value. These methods are used for both linear and nonlinear systems.

Note that in the linear case, the equivalence between observers, parity space and parameter estimation has been established (Patton and Chen 1991).

4.2 Comparative Analysis and Concept Mapping for the Model-Based Logical Diagnosis Approach and the Analytical Redundancy Approach

The correspondences in terms of principles, concepts, and assumptions between the model-based diagnostic methods from Automatic Control and those from AI were showed by the French community, concretized by the IMALAIA group mentioned above. This work is recorded in the collective paper (Cordier et al. 2004). The comparative analysis is based on the comparison of the so-called structured residuals approach, or parity space approach (Chow and Willsky 1984), and the logical theory of diagnosis as proposed by Reiter (1987), Kleer et al. (1992) and presented in Sect. 2.

The parity space approach is based on the off-line computation of a set of ARRs from a model SM decomposed in a behavior model BM and an observation model OM. The equations of the model SM are constraints which can be associated with components but this information is not represented explicitly.

The ARRs define constraints for the observable variables O of the system, that is to say the input and output variables, and are obtained by techniques allowing to eliminate state variables that are unknown. Each ARR can be put in the form $r = 0$, where r is called *residual*.

Definition 18 (*ARR for SM* (z, x)) A relation of the form $r(z, \dot{z}, \ddot{z}, \dots) = 0$ is an ARR for the model $SM(z, x)$ if for all z consistent with $SM(z, x)$, the relation is satisfied.

If the behavior of the system satisfies the constraints of the model, then the residuals are zero because the ARRs are satisfied, otherwise some of them may be different from zero and the corresponding ARRs are said violated. Each fault F_j has an associated theoretical signature $FS_j = [s_{1j}, s_{2j}, \dots, s_{nj}]$ given by the binary evaluation (0 or not 0) of each of the residuals. We can then define the *signature matrix FS*.

Definition 19 (*Signature Matrix*) Given a set of n ARRs, the signature matrix FS associated to a set of n_f faults $F = [F_1, F_2, \dots, F_{n_f}]$ is the matrix that crosses ARRs as rows and faults as columns, and whose columns are given by the theoretical signatures of the faults.

Diagnosis consists in the online comparison of the "observed signature", vector of the residuals evaluated with the observations, and the theoretical signatures of the n_f anticipated faults. In the logical theory of diagnosis, the description of the system is component oriented and rests on first order logic in its original version. This has been discussed in detail in Sect. 2.1.

A diagnosis for the system $(SD, COMPS, OBS)$ is a set $\Delta \subseteq COMPS$ such that the assumption that the components of Δ are the only ones to be faulty is consistent with the observations and the description of the system, that is $SD \cup OBS \cup \{AB(C) \mid C \in \Delta\} \cup \{\neg AB(C) \mid C \in COMPS \setminus \Delta\}$ is satisfiable.

Most FDI works do not explicitly use the concept of component given that the behavior model BM represents the global system. When models based on the concept of component are used, topological knowledge is implicitly represented by shared variables. Conversely, the DX approach explicitly represents the topology of the system and the behavior models of the components. The main difference is that the hypothesis of correct behavior of a component, which underlies its model, is represented explicitly by the predicate AB. If $\mathscr{F}$ is a formula representing the correct behavior of a component C, SM contains only $\mathscr{F}$ while SD contains the formula $\neg AB(C) \Rightarrow \mathscr{F}$.

To compare the approaches, the *system representation equivalence* (SRE) property resulting in the fact that SM is obtained from SD by substituting all the occurrences of the predicate $AB(.)$ by $\bot$ is considered true. It is also assumed that the same observation language OBS is used, constituted by a conjunction of equality relations that assign a value v to each observable variable. Finally, the faults relate to the same entities considered as components, without loss of generality. The comparison is based on a theoretical framework to precisely establish the correspondence between the different concepts. This framework is provided by the signature matrix FS, for which each row is associated with an ARR and each column with a component (under the assumption that the faults relate to components). It relies on the concept of *support* of an ARR:

Definition 20 (*ARR Support*) The *support* of an ARR ARR_i, noted $supp(ARR_i)$, is the set of components whose columns in the signature matrix FS have a non zero element on the ARR_i row.

In addition, the following two properties are added:

Property 7 ARR–d–completeness *A set E of ARRs is said to be* d-complete *if:*

- *E is finite;*
- $\forall OBS$, *if* $SM \cup OBS \models \bot$, *then* $\exists ARR_i \in E$ *such that* $\{ARR_i\} \cup OBS \models \bot$.

Property 8 (ARR–i–completeness) *A set E of ARRs is said to be* i-complete *if:*

- *E is finite;*
- $\forall \mathscr{C}$, *set of components such that* $\mathscr{C} \subseteq COMPS$, *and* $\forall OBS$, *if* $SM(\mathscr{C}) \cup OBS \models \bot$, *then* $\exists ARR_i \in E$ *such that* $supp(ARR_i)$ *is included in* $\mathscr{C}$ *and* $\{ARR_i\} \cup OBS \models \bot$.

We then obtain the following result:

Property 9 *Assuming the SRE property and that OBS is the set of observations for the system given by SM (or SD), then:*

1. *If* ARR_i *is violated by OBS, then* $supp(ARR_i)$ *is a conflict set;*
2. *Given E a set of ARRs:*
   - *If E is d-complete, and if there exists a conflict set for* $(SD, COMPS, OBS)$, *then there exists* $ARR_i \in E$ *violated by OBS;*

- *If E is i-complete, then given a conflict set $\mathscr{C}$ for $(SD, COMPS, OBS)$, there exists $ARR_i \in E$ violated by OBS such that $supp(ARR_i)$ is included in $\mathscr{C}$.*

The first result can be intuitively explained by the fact that inconsistencies between model and observations, appraised by the conflicts in the DX approach, are apprehended by ARRs violated by *OBS* in the FDI approach. In consequence, the support of an ARR can be defined as a *potential conflict*. This result echoes the notion of *possible conflict* proposed by Pulido and Gonzalez (2004). The second result provides existence and completeness results, the first referring to detectability and the second to isolability.

We then show below that in the presence of the same assumptions about the manifestation of faults (their observability), commonly called *exoneration assumptions*, and in particular the absence of ARR-exoneration, a result linking the diagnoses on both sides can be obtained.

Definition 21 (*ARR-exoneration*) Given *OBS*, any component in the support of an ARR satisfied by *OBS* is exonerated, i.e., considered as normal.

This assumption states that faults having no observable manifestation through a non-zero residual are exonerated.

Theorem 6 *Under the i-completeness assumption, the diagnoses obtained by the FDI approach in the case of no ARR-exoneration are identical to the (non empty) diagnoses obtained by the DX approach.*

Let us note that the assumptions generally adopted by the two communities are different, the FDI community implicitly adopting the ARR-exoneration assumption. In addition, the computation of fault signatures limits the number of anticipated faults. Conventionally, only single faults are considered. Conversely, in the logical diagnosis theory, no assumption is made *a priori* about the number of faults, even if preferences can be introduced to privilege minimal or highest probability diagnoses. This ensures logically correct results. It can also be noted that in the FDI approach, computation of ARRs and fault signatures is done offline and only a consistency test is required online. This can be advantageous if computational time constraints come into play. In the logical theory diagnosis, all the processing is done online, the advantage being that only the models are to be updated if the system undergoes changes. Note that the two approaches can be combined to take advantage of both. One can cite DX works which adopt the FDI idea of offline generation of the RRAs (Loiez and Taillibert 1997; Washio et al. 1999; Pulido and Gonzalez 2004). One can also cite the works, presented in more detail in Sect. 4.3, which take advantage of explicitly representing the causal influences underlying the model of the system and those concerned with diagnosis of hybrid systems.

4.3 Approaches Taking Advantage of Techniques of Both Fields

Diagnosis Based on Influence Graphs/Causal Graphs

In the 1990s, the synergies between the Qualitative Reasoning community (Travé-Massuyès and Dague 2003; Weld and De Kleer 1989) (see also chapter "Qualitative Reasoning about Time and Space" of this volume) and the Model-Based Diagnosis community concretized in a set of works proposing to use *causal models* for diagnosis reasoning (see chapter "A Glance at Causality Theories for Artificial Intelligence" of this volume). Unlike causal graphs pointed in Sect. 2.2, influence graphs rely on a structure expressing the dependencies between variables in the model of the system explicitly, known as *influences* thus making it possible to provide explanations as to why normal or abnormal values of variables. This structure is commonly called a *causal graph*. Dependencies are obtained directly from expert knowledge (Gentil et al. 2004) or from causal ordering techniques (Travé-Massuyès et al. 2001; Pons et al. 2015) or also from *bond graph* models (Dague and Travé-Massuyès 2004; Chatti et al. 2014).

The very first works were limited to labeling the causal influences by the signs giving the direction of variation of the cause variable with respect to the effect variable, thus obtaining a *signed oriented graph* (Kramer and Palowitch 1987). Subsequently, the parametrization of influences was sophisticated as they were labeled by quantitative local models, such as those used by the FDI community.

By way of example, the principles of the causal fault detection and isolation method CaEn2 (Travé-Massuyès et al. 2001; Travé-Massuyès and Calderon-Espinoza 2007) are given below. Fault detection is an online process that assesses the consistency of sensor measures with respect to the behavioral model of the system. The detection of a variable as abnormal is interpreted as the violation of the influences implied in the estimation of the variable, i.e., the ascending influences in the causal graph. Each influence being associated with a component, this allows one to characterize a set of components constituting a conflict set. The influences of CaEn2 have a "delay" attribute corresponding to a pure delay in the input-output function associated with the influence. This information is used to generate conflict sets whose components are labeled by a time label indicating the date at the latest at which the fault occurred on the component. Diagnoses are obtained from conflict sets by an incremental algorithm that generates hitting sets while managing time labels (Travé-Massuyès and Calderon-Espinoza 2007).

Diagnosis of Hybrid Systems

The works on hybrid systems have been steadily increasing since the pioneering works in the early 2000s (McIlraith et al. 2000). Hybrid systems make it possible to represent double continuous and discrete dynamics that cohabit in many modern

systems. Most systems are indeed made up of a set of heterogeneous interconnected components, orchestrated by a supervisor whose commands, of discrete nature, induce different operation modes. Hybrid system modeling as well as associated diagnosis algorithms use continuous and discrete mathematics, so that hybrid systems open a predilection area for integrating methods from the two FDI and DX communities.

The NASA *Livingstone* diagnosis engine (Williams and Nayak 1996), which flew onboard the DS-1 probe, was one of the first to qualify as hybrid. This engine was rooted in the AI model-based diagnosis framework, relying on a model written in propositional logic, and behavioral equations accounting for continuous aspects abstracted in the form of logical relations (qualitative constraints). However, qualitative abstraction imposed *monitors* between the sensors and the model to interpret the actual continuous signals in terms of discrete modalities. The difficulty in deciding proper thresholds and the poor sensitivity of the fault detection procedure led subsequent works to consider true hybrid models, associating differential equation and discrete event models. As proposed in (Bayoudh et al. 2008a; Bayoudh and Travé-Massuyès 2014) a hybrid model can be represented in the form of a 6-tuple:

$$S = (\zeta, Q, E, T, K, (q_0, \zeta_0))$$

where:

- ζ is the vector of continuous variables;
- Q is the set of discrete system states, each representing an operating mode of the system;
- E is the set of events corresponding to discrete commands, autonomous mode transitions, or occurrence of faults; events corresponding to autonomous mode transitions are subject to guards that depend on continuous variables;
- $T \subseteq Q \times E \rightarrow Q$ is the transition function; it is possible to attach probabilities to the transitions;
- $K = \cup K_i$ is the set of constraints linking the continuous variables, taking the form of differential and possibly algebraic equations modeling the continuous behavior of the system in the different modes $q_i \in Q$;
- $(\zeta_0, q_0) \in \zeta \times Q$ is the initial condition of the hybrid system.

In the hybrid state (ζ, Q), only the discrete state $q_i \in Q$ is representative of the operating mode of the system and provides the diagnosis. However, the evolution of the discrete state is interlinked to the evolution of the continuous state, which is why the problem of diagnosis is often brought back to the problem of estimating the complete hybrid state.

In theory, hybrid estimation presupposes to consider all the sequences of possible modes with the continuous evolution associated with them, which results in exponential complexity. Consequently, many suboptimal methods have been proposed for which we can distinguish the three following families:

- Methods based on *multimode filtering*, rather anchored in the Automatic Control field (Blom and Bar-Shalom 1988; Hofbaur and Williams 2004; Benazera and Travé-Massuyès 2009), are formulated in a probabilistic framework. They track the different "hypotheses", that is to say the sequences of modes and their associated continuous evolution, over a limited time window and merge the continuous estimates according to a likelihood measure resulting in a *belief state* in the form of a probability distribution over the states at the current time.
- Methods based on *particle filtering* (Arulampalam et al. 2002) are based on sampling and rely on a Bayesian update of the belief state. With enough samples, they approximate the optimal Bayesian estimate but are not well adapted to the problem of the diagnosis because the probabilities of faults are generally very low in comparison with the probabilities of the nominal states of the system.
- Methods that address hybrid aspects in a *dedicated manner* adopt strategies to retrieve the trajectory of the system when it has been discarded due to the approximation of the estimation method (Nayak and Kurien 2000; Benazera and Travé-Massuyès 2003).

Let us note that (Bayoudh et al. 2008b; Vento et al. 2015; Sarrate et al. 2018) propose an alternative approach to complete hybrid diagnosis that only estimates the discrete state, i.e. the operating mode. It combines the parity space approach based on ARRs as defined in Sect. 4.2 for processing the information provided by continuous dynamics with the DES diagnoser method as presented in Sect. 3.4 (Sampath et al. 1995).

Recent works address hybrid system diagnosability integrating a *twin plant* approach as presented in Sect. 3.6 for DES with mode distinguishability methods coming from the FDI community (Grastien et al. 2017). This work is based on abstracting the hybrid automaton model. The continuous dynamics are abstracted remembering only two pieces of information: discernability between modes (when they are guaranteed to generate different observations) and ephemerality (when the system cannot stay forever in a given set of modes). Iterative abstractions can be checked for diagnosability with the standard DES twin plant method that provides a counterexample in case of non-diagnosability. The absence of such a counterexample proves the diagnosability of the original hybrid system. In the opposite case, the counterexample is analyzed to refine the DES. This procedure is referred as a counterexample guided abstraction refinement (CEGAR) scheme. It supports the proposals of Zaatiti et al. (2017, 2018) in which Qualitative Reasoning (see chapter "Qualitative Reasoning about Time and Space" of this volume) is used to compute discrete abstractions. Abstractions as timed automata allow one to handle time constraints that can be captured at a qualitative level.

5 Conclusion

Model-based diagnosis found its formal bases in the 1980s for static systems and in the 1990s with regard to dynamic systems. Since then, developments have been constant and promising, and have in fact become industrialized in several industrial domains such as automotive, aeronautics, space. An important point for the French diagnosis community is the real collaboration of the Automatic Control and AI communities, which brought their respective approaches close together by showing their proximity and their specificities. This has been quite productive on both sides. In the domain of dynamic systems, interest has developed over the last few years on hybrid systems, making it possible to deal with double dynamics, discrete and continuous, and to account for the heterogeneity of current systems. It is a privileged area for the collaborations between the two communities.

One of the current topics is the improvement of the efficiency of existing algorithms to scale up and approach large systems such as those proposed by the DX competition, for instance electronic circuits comprising several thousand components. This involves the use of data structures like BDDs or very efficient algorithms like SAT, taking into account the structure of the systems. This also involves distributed approaches that divide the problem in a set of problems that are as independent as possible.

Another major pathway concerns the properties of the systems from the diagnosis point of view, namely the in depth study of diagnosability, observability, and repairability for enabling the design of systems which can be monitored, diagnosed and repaired optimally. A last line of work concerns the monitoring of distributed systems for which detection, diagnosis, and return to nominal operating conditions requires good collaboration between methods and tools proposed by the FDI and AI communities. This is also true for planning and decision making.

Finally, as in any situation where model and real world coexist, attention must be paid to the problems linked to the quality and precision of the model, compared to the quality of the information (accuracy, precision, etc.) gathered on the real system, through sensors that can be imperfect and subject to faults. For all these developments, it can be noted that this involves dialogue and co-operation with researchers from many fields, in particular those from the AI community. This is obviously a challenge but also an opportunity for reciprocal fertilization.

For detailed references on the topic of diagnosis, it is best to consult the proceedings of the international conference DX (Principles of diagnosis), which brings together every year researchers in the field (DX 2018).

References

Aghasaryan A, Fabre E, Benveniste A, Boubour R, Jard C (1997) A Petri net approach to fault detection and diagnosis in distributed systems. II. Extending Viterbi algorithm and HMM techniques to Petri nets. In: 36th IEEE conference on decision and control, San Diego (CA), USA, pp 726–731

Armengol J, Bregon A, Escobet T, Gelso E, Krysander M, Nyberg M, Olive X, Pulido B, Travé-Massuyès L (2009) Minimal structurally overdetermined sets for residual generation: a comparison of alternative approaches. In: 7th IFAC symposium on fault detection, supervision and safety of technical processes, Barcelona, Spain, pp 1480–1485

Arulampalam MS, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans Signal Process 50(2):174–188

Bayoudh M, Travé-Massuyès L (2014) Diagnosability analysis of hybrid systems cast in a discrete-event framework. Discret Event Dyn Syst 24(3):309–338

Bayoudh M, Travé-Massuyès L, Olive X (2008a) Coupling continuous and discrete event system techniques for hybrid system diagnosability analysis. In 18th European conference on artificial intelligence including prestigious applications of intelligent, Patras, Greece. IOS Press, pp 219–223

Bayoudh M, Travé-Massuyès L, Olive X (2008b) Hybrid systems diagnosis by coupling continuous and discrete event techniques. In: Proceedings of the IFAC world congress, Seoul, Korea, pp 7265–7270

Benazera E, Travé-Massuyès L (2003) The consistency approach to the on-line prediction of hybrid system configurations. In: Analysis and design of hybrid systems 2003 (ADHS 03): a proceedings volume from the IFAC Conference, St. Malo, Brittany, France, 16–18 June 2003. Elsevier Science, pp 241–246

Benazera E, Travé-Massuyès L (2009) Set-theoretic estimation of hybrid system configurations. IEEE Trans Syst Man Cybern. Part B Cybern: Publ IEEE Syst Man Cybern Soc 39(6):1277–1291

Besnard P, Cordier M-O (1994) Explanatory diagnoses and their characterization by circumscription. Ann Math Artif Intell 11:75–96

Blanke M, Kinnaert M, Lunze J, Staroswiecki M (2015) Diagnosis and fault-tolerant control, 3rd edn. Springer, Berlin

Blanke M, Kinnaert M, Schröder J, Lunze J, Staroswiecki M (2003) Diagnosis and fault-tolerant control. Springer, Berlin

Blom H, Bar-Shalom Y (1988) The interacting multiple model algorithm for systems with Markovian switching coefficients. IEEE Trans Autom Control 33:780–783

Brusoni V, Console L, Terenziani P, Dupré DT (1998) A spectrum of definitions for temporal model-based diagnosis. Artif Intell 102:39–79

Carle P, Choppy C, Kervarc R (2011) Behaviour recognition using chronicles. In: 2011 fifth international conference on theoretical aspects of software engineering, pp 100–107

Chatti N, Ould-Bouamama B, Gehin A-L, Merzouki R (2014) Signed bond graph for multiple faults diagnosis. Eng Appl Artif Intell 36:134–147

Chittaro L, Ranon R (2004) Hierarchical model-based diagnosis based on structural abstraction. Artif Intell 1–2:147–182

Chow E, Willsky A (1984) Analytical redundancy and the design of robust failure detection systems. IEEE Trans Autom Control 29(7):603–614

Cimatti A, Pecheur C, Cavada R (2003) Formal verification of diagnosability via symbolic model checking. In: Proceedings of the 18th international joint conference on artificial intelligence IJCAI'03, Acapulco, Mexico, pp 363–369

Console L, Picardi C, Ribaudo M (2002) Process algebra for systems diagnosis. Artif Intell 142:19–51

Console L, Torasso P (1990) Hypothetical reasoning in causal models. Int J Intell Syst 5(1):83–124

Console L, Torasso P (1991) A spectrum of logical definitions of model-based diagnosis. Comput Intell 7:133–141

Contant O, Lafortune S, Teneketzis D (2004) Diagnosis of intermittent faults. Discret Event Dyn Syst: Theory Appl 14(2):171–202

Cordier M, Dague P, Lévy F, Montmain J, Staroswiecki M, Travé-Massuyès L (2004) Conflicts versus analytical redundancy relations: a comparative analysis of the model based diagnosis approach from the artificial intelligence and automatic control perspectives. IEEE Trans Syst Man Cybern Part B 34(5):2163–2177

Cordier M-O (1998) When abductive diagnosis fails to explain too precise observations: an extended spectrum of model-based diagnosis definitions based on abstracting observations. In: Proceedings of DX'98, Cape Cod (MA), USA, pp 24–31

Cordier M-O, Pencolé Y, Travé-Massuyès L, Vidal T (2007) Self-healability = diagnosability + repairability. In: 18th international workshop on principles of diagnosis, Nashville, Tennessee, USA, pp 251–258

Cordier M-O, Thiébaux S (1994) Event-based diagnosis for evolutive systems. In: 5th international workshop on principles of diagnosis (DX-94), New Palz (NY), USA, pp 64–69

Cordier M-O, Travé-Massuyès L, Pucel X (2006) Comparing diagnosability in continuous and discrete-event systems. In: 17th international workshop on principles of diagnosis (DX06), Burgos, Spain, pp 55–60

Dague P, Jehl O, Taillibert P (1990) An interval propagation and conflict recognition engine for diagnosing continuous dynamic systems. In: Expert systems in engineering, pp 16–31

Dague P, Travé-Massuyès L (2004) Raisonnement causal en physique qualitative. Intellectica 38:247–290

De Kleer J (1992) Focusing on probable diagnosis. Readings in model-based diagnosis. Morgan Kaufmann, San Mateo

De Kleer J (2006) Improving probability estimates to lower diagnostic costs. In: 17th international workshop on principles of diagnosis (DX06), Burgos, Spain, pp 55–60

De Kleer J, Williams B (1987) Diagnosing multiple faults. Artif Intell 32(1):97–130

Debouk R, Lafortune S, Teneketzis D (2002) Coordinated decentralized protocols for failure diagnosis of discrete event systems. Discret Event Dyn Syst: Theory Appl 10(1–2):33–86

Dousson C (1996) Alarm driven supervision for telecommunication networks: II -On line chronicle recognition. Annales des Télécommunications 51(9–10):501–508

Dubuisson B (2001) Automatique et statistiques pour le diagnostic. Hermes Science Europe Ltd

DX (2018) Proceedings of the 0th to 29th international workshop on principles of diagnosis, 1989–2018

Fabre E, Benveniste A, Haar S, Jard C (2005) Distributed monitoring of concurrent and asynchronous systems. Discret-Event Dyn Syst: Theory Appl 15(1):33–84

Feldman A, Provan G, Van Gemund A (2009) FRACTAL: efficient fault isolation using active testing. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'09), Pasadena (CA), USA, pp 778–784

Friedrich G, Gottlob G, Nejdl W (1994) Formalizing the repair process - extended report. Ann Math Artif Intell 11(1–4):187–201

Gao Z, Cecati C, Ding SX (2015) A survey of fault diagnosis and fault-tolerant techniques part i: fault diagnosis with model-based and signal-based approaches. IEEE Trans Ind Electron 62(6):3757–3767

Gentil S, Montmain J, Combastel C (2004) Combining FDI and AI approaches within causal-model-based diagnosis. IEEE Trans Syst Man Cybern Part B 34(5):2207–2221

Gertler J (1998) Fault detection and diagnosis in engineering systems. Marcel Deker, New York

Gougam H-E, Pencolé Y, Subias A (2017) Diagnosability analysis of patterns on bounded labeled prioritized Petri nets. J Discret Event Dyn Syst: Theory Appl 27(1):143–180

Grastien A, Cordier M-O, Largouët C (2005) Automata slicing for diagnosing discrete-event systems with partially ordered observations. In: 9th congress of the Italian association for artificial intelligence, Milano, Italy, pp 270–281

Grastien A, Travé-Massuyès L, Puig V (2017) Solving diagnosability of hybrid systems via abstraction and discrete event techniques. IFAC-PapersOnLine 50(1):5023–5028

Grastien Al, Anbulagan An (2013) Diagnosis of discrete event systems using satisfiability algorithms: a theoretical and empirical study. IEEE Trans Autom Control (TAC) 58(12):3070–3083
Greiner R, Smith B, Wilkerson R (1989) A correction to the algorithm in Reiter's theory of diagnosis. Artif Intell 41:79–88
Hofbaur MW, Williams BC (2004) Hybrid estimation of complex systems. IEEE Trans Syst, Man, Cybern-Part B: Cybern 34(5):2178–2191
Jéron T, Marchand H, Pinchinat S, Cordier M-O (2006) Supervision patterns in discrete event systems diagnosis. In: Workshop on discrete event systems, WODES'06, Ann-Arbor (MI), USA, pp 262–268
Jiang S, Huang Z, Chandra V, Kumar R (2001) A polynomial time algorithm for diagnosability of discrete event systems. IEEE Trans Autom Control 46(8):1318–1321
Jiroveanu G, Boel R (2006) A distributed approach for fault detection and diagnosis based on time Petri nets. Math Comput Simul 70(5–6):287–313
KanJohn P, Grastien A (2008) Local consistency and junction tree for diagnosis of discrete-event systems. In: European conference on artificial intelligence (ECAI-08). Patras, Greece, pp 209–213
Kleer J, Mackworth A, Reiter R (1992) Characterizing diagnoses and systems. Artif Intell 56(2–3):197–222
Kramer MA, Palowitch BL (1987) A rule-based approach to fault diagnosis using the signed directed graph. AIChE J 33(7):1067–1078
Krysander M, Åslund J, Nyberg M (2008) An efficient algorithm for finding minimal overconstrained subsystems for model-based diagnosis. IEEE Trans Syst, Man, Cybern-Part A: Syst HumS 38(1):197–206
Lamperti G, Zanella M (2003) Diagnosis of active systems. Kluwer Academic Publishers, Dordrecht
Loiez E, Taillibert P (1997) Polynomial temporal band sequences for analog diagnosis. In: IJCAI-97: proceedings of the fifteenth international joint conference on artificial intelligence, Nagoya, Japan, pp 474–479
Lunze J (1994) Qualitative modelling of linear dynamical systems with quantized state measurements. Automatica 30(3):417–431
Marchand H, Rozé L (2002) Diagnostic de pannes sur des systèmes à événements discrets : une approche à base de modèles symboliques. In: 13ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle. Angers, France, pp 191–200
McCarthy J (1986) Applications of circumscription to formalizing common-sense knowledge. Artif Intell 28:89–116
McIlraith S, Biswas G, Clancy D, Gupta V (2000) Hybrid systems diagnosis. Lecture notes in computer science, pp 282–295
Nayak P, Kurien J (2000) Back to the future for consistency-based trajectory tracking. In: Proceedings of AAAI-2000, Austin (TX), USA, pp 370–377
Nejdl W, Bachmayer J (1993) Diagnosis and repair iteration planning versus n-step look ahead planning. In: 4th international workshop on principles of diagnosis, Aberystwyth, UK
Patton R, Chen J (1991) A re-examination of the relationship between parity space and observer-based approaches in fault diagnosis. Eur J Diagn Saf Autom 1(2):183–200
Pencolé Y (2004) Diagnosability analysis of distributed discrete event systems. In: European conference on artificial intelligence (ECAI'04). Valencia, Spain, pp 43–47
Pencolé Y, Cordier M-O (2005) A formal framework for the decentralised diagnosis of large scale discrete event systems and its application to telecommunication networks. Artif Intell 164:121–170
Pencolé Y, Schumann A, Kamenetsky D (2006) Towards low-cost fault diagnosis in large component-based systems. In: 6th IFAC symposium on fault detection, supervision and safety of technical processes, Beijing, China, pp 1473–1478
Pencolé Y, Steinbauer G, Mühlbacher C, Travé-Massuyès L (2018) Diagnosing discrete event systems using nominal models only. In: 28th international workshop on principles of diagnosis, Brescia, Italy, pp 169–183

Pencolé Y, Subias A (2018) Diagnosis of supervision patterns on bounded labeled petri nets by model checking. In: 28th international workshop on principles of diagnosis, Brescia, Italy, pp 184–199

Peng Y, Reggia JA (1990) Abductive inference models for diagnsotic problem-solving. Springer, Berlin

Pons R, Subias A, Travé-Massuyès L (2015) Iterative hybrid causal model based diagnosis: application to automotive embedded functions. Eng Appl Artif Intell 37:319–335

Poole D (1989) Normality and faults in logic-based diagnosis. In: IJCAI, pp 1304–1310

Provan G (2002) On the diagnosability of decentralized, timed discrete event systems. In: 41st IEEE conference on decision and control, Las Vegas (NV), USA, pp 405–410

Pulido B, Gonzalez C (2004) Possible conflicts: a compilation technique for consistency-based diagnosis. IEEE Trans Syst, Man, Cybern, Part B 34(5):2192–2206

Reggia JA, Nau D, Wang Y (1983) Diagnostic expert systems based on a set covering model. Int J Man-Mach Stud 19:437–460

Reiter R (1987) A theory of diagnosis from first principles. Artif Intell 32(1):57–95

Ribot P, Pencolé Y, Combacau M (2008) Design requirements for the diagnosability of distributed discrete event systems. In: 19th international workshop on principles of diagnosis. Blue Mountains, New South Wales, Australia, pp 347–354

Rozé L, Cordier M-O (2002) Diagnosing discrete-event systems: extending the "diagnoser approach" to deal with telecommunication networks. Discrete-Event Dyn Syst: Theory Appl 12(1):43–81

Sampath M, Sengupta R, Lafortune S, Sinnamohideen K, Teneketzis D (1995) Diagnosability of discrete event system. IEEE Trans Autom Control 40(9):1555–1575

Sampath M, Sengupta R, Lafortune S, Sinnamohideen K, Teneketzis D (1996) Failure diagnosis using discrete-event models. IEEE Trans Control Syst Technol 4(2):105–124

Sarrate R, Puig V, Travé-Massuyès L (2018) Diagnosis of hybrid dynamic systems based on the behavior automaton abstraction. In: Fault diagnosis of hybrid dynamic and complex systems. Springer, Berlin, pp 243–278

Schumann A, Pencolé Y (2007) Scalable diagnosability checking of event-driven system. In: Proceedings of the twentieth international joint conference on artificial intelligence (IJCAI07), Hyderabad, India, pp 575–580

Schumann A, Pencolé Y, Thiébaux S (2004) Diagnosis of discrete-event systems using binary decision diagrams. In: Proceedings of the internationalworkshop on principles of diagnosis (DX'04), Carcassonne, France, pp 197–202

Schumann A, Pencolé Y, Thiébaux S (2010) A decentralised symbolic diagnosis approach. In: 19th European conference on artificial intelligence (ECAI-10). IOS Press, Lisbon, Portugal, pp 99–104

Siddiqi S, Huang J (2010) New advances in sequential diagnosis. In: Proceedings of the twelfth international conference on the principles of knowledge representation (KR'10), Toronto, Canada, pp 17–25

Staroswiecki M, Comtet-Varga G (2001) Analytical redundancy relations for fault detection and isolation in algebraic dynamic systems. Automatica 37(5):687–699

Su X, Grastien Al (2013) Diagnosis of discrete event systems by independent windows. In: 24th international workshop on principles of diagnosis (DX-13), Jerusalem, Israel, pp 148–153

Su X, Grastien Al, Pencolé Ya (2014) Window-based diagnostic algorithms for discrete event systems: what information to remember. In: 25th international workshop on principles of diagnosis (DX-14), Graz, Austria

Su X, Zanella M, Grastien A (2016) Diagnosability of discrete-event systems with uncertain observations. In: 25th international joint conference on artificial intelligence (IJCAI-16), pp 1265–1271

Sun Y, Weld DS (1993) A framework for model-based repair. In: 11th national conference on artificial intelligence, Washington, D.C., USA, pp 182–187

Ten Teije A, Van Harmelen F (1994) An extended spectrum of logical definitions for diagnostic systems. In: Proceedings of DX-94 Fifth International Workshop on Principles of Diagnosis, New Paltz (NY), USA, pp 334–342

Torta G, Torasso P (2003) Automatic abstraction in component-based diagnosis driven by system observability. In: Proceedings of the 18th international joint conference on artificial intelligence - IJCAI03, Mexico, Acapulco, pp 394–400

Travé-Massuyès L (2014) Bridging control and artificial intelligence theories for diagnosis: a survey. Eng Appl Artif Intell 27:1–16

Travé-Massuyès L, Calderon-Espinoza G (2007) Timed fault diagnosis. In: Proceedings of the IEEE European control conference (ECC-07), Kos, Greece, pp 2272–2279

Travé-Massuyès L, Dague P (2003) Modèles et raisonnements qualitatifs. Hermes sciences

Travé-Massuyès L, Escobet T, Milne R (2001) Model-based diagnosability and sensor placement application to a frame 6 gas turbine subsystem. In: Proceedings of the seventeenth international joint conference on artificial intelligence, IJCAI'01, vol 1, pp 551–556

Travé-Massuyès L, Pons R, Tornil S, Escobet T (2001) The CA-En diagnosis system and its automatic modelling method. Computación y Sistemas 5(2):128–143

Vento J, Travé-Massuyès L, Puig V, Sarrate R (2015) An incremental hybrid system diagnoser automaton enhanced by discernibility properties. IEEE Trans Syst, Man, Cybern: Syst 45(5):788–804

Washio T, Motoda H, Niwa Y (1999) Discovering admissible model equations from observed data. In Proceeding of IJCAI99: sixteenth international joint conferenceon artificial intelligence, vol 2, Stockholm, Sweden, pp 772–779

Weld D, De Kleer J (1989) Readings in qualitative reasoning about physical systems. Morgan Kaufmann Publishers Inc

Williams BC, Nayak P (1996) A model-based approach to reactive self-configuring systems. In: Proceedings of the 13th national conference on artificialintelligence (AAAI-96), Portland (OR), USA, pp 971–978

Ye L, Dague P (2012) A general algorithm for pattern diagnosability of distributed discrete event systems. In: ICTAI - 24th international conference ontools with artificial intelligence, Athens, Greece

Ye L, Dague P (2017) An optimized algorithm of general distributed diagnosability analysis for modular structures. IEEE Trans Autom Control 62(4):1768–1780

Yoo T, Lafortune S (2002) Polynomial-time verification of diagnosability of partially-observed discrete-event systems. IEEE Trans Autom Control 47(9):1491–1495

Zaatiti H, Ye L, Dague P, Gallois J-P (2017) Counter example guided abstraction refinement for hybrid systems diagnosability analysis. In: 28th internationalworkshop on principles of diagnosis (DX-17), Brescia, Italy

Zaatiti H, Ye L, Dague P, Gallois J-P, Travé-Massuyès L (2018) Abstractions refinement for hybrid systems diagnosability analysis. In: Diagnosability, security and safety of hybrid dynamic and cyber-physical systems. Springer, Berlin, pp 279–318

Validation and Explanation

Laurent Charnay, Juliette Dibie and Stéphane Loiseau

Abstract Knowledge Based systems (KBS) that succeeded to expert systems are used nowadays to face different decision problems. Their architecture separates the modular and declarative knowledge of an application domain from its control using inference algorithms. This architecture requires a specific validation approach. KBS have been also the basis of many systems for which the explanation of computed results are almost as important as the results themselves. The aim of this chapter is to show the issues and the solutions to valid KBS and their use to explain reasoning.

1 Introduction

The first Knowledge Based systems approaches were developed fourthly years ago (Feigenbaum et al. 1971; Minsky 1975; Shortliffe 1976). These systems were essentially characterized by their application *domains* and their implementation *language*. These systems, called *expert system*, were made to replace human experts. The languages used to implement these systems were high level languages; they were supposed to ease computer coding by providing a formalism easily understandable by a non computer scientist. These high level languages enabled *knowledge* to be expressed in a *Knowledge Base* (KB) of which the use was controlled by an inference engine. The knowledge was traditionally characterized by its *modularity*, the *declarativity* of its information and its *semantics*. That means that each piece of knowledge was independent, easy to understand, and had a clear – often logical – associated semantics. The inference engine relied on algorithms, most of the time

L. Charnay (✉)
Orange Business Services, Paris, France
e-mail: laurent.charnay@orange.com

J. Dibie
UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, France
e-mail: juliette.dibie@agroparistech.fr

S. Loiseau
LERIA, Université Angers, Angers, France
e-mail: stephane.loiseau@univ-angers.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_22

logical deduction algorithms, which were independent of any application domains and based on the knowledge semantics.

These characteristics of knowledge (i.e. modularity, declarativity and clear semantics) enabled on the one hand the expert knowledge acquisition to be simplified and so the KBS construction, *validation* and maintenance and, on the other hand, the knowledge easily use not only to solve problems but also to *explain* the computed results. A revolution was expected: computer systems made to solve problems should become with KBS a way to propose solutions to end-users and to explain these solutions. Regarding validation, the knowledge declarativity made the difference between the domain knowledge and the implemented knowledge thin, so the knowledge validation in the KBS should be easy for a domain expert or a KBS computer developer. The knowledge modularity provided a useful help to find potential invalidities and correct them. Regarding explanation, the knowledge declarativity enabled each piece of knowledge used during the problem resolution to become a good candidate to partially explain the solutions: every end-user could understand a piece of knowledge. Furthermore, thanks to the knowledge modularity, each piece of knowledge used in the problem resolution could be cleverly combined to provide a relevant explanation to the end-user. So, a *trace*, i.e. a set of pieces of knowledge produced by the KBS to identify a problem or to obtain a solution, appeared to be a useful way to validate a KB and explain the solution. Many of these KBS were based on a knowledge representation language using rules (Farreny 1985). These *rule based systems* are the core of numerous KBS today. In this chapter, the basic mechanisms used to validate such rule based systems and to explain computed results are presented.

Nevertheless, new needs, resulting from the reflexion about the small use of expert system, appeared at the end of the twentieth century. Four tendencies then emerged in the KBS researches. First, the separation between knowledge and control showed its limits: the control could not always be conceived independently from the system knowledge. Some *meta-knowledge* were therefore provided to direct the control, but this questions the knowledge modularity. Second, the knowledge representation in real life system showed new semantics needs: taking into account knowledge *uncertainty* (e.g. he will probably win the game), knowledge *imprecision* (e.g. he has about 40 degrees fever) and knowledge *context* (e.g. assume that he has fever, he has the flu). The pragmatic approaches derived from systems such as the "*certainty factors*" of MYCIN (Shortliffe 1976) or the theoretical improvements (Sombe 1988) about non classical logics, enabled the KBS reasoning to be improved but at the expense of the KBS readability and understanding which questions their validation and explanation. Third, several problems needed to take into account different kinds of knowledge with different semantics – such as rules, classes, ontology, networks – to build KBS. To face this, *second generation expert systems* (Steels 1985; David 1995) were conceived. Four, the interactions between different kinds of knowledge and between knowledge and the human took a crucial importance in the systems design. Several works on *human and intelligent machine interaction* were done. They proposed new models of resolution, explanation, interaction and dialog. Nowadays new researches about the explanation appear, for instance on intelligent interactive learning environment. Moreover, the explosion of works on the semantic web and

the web of data questions the validation and explanation methods. In this chapter, we show how those new needs require new approaches of validation and explanation.

This chapter is composed of three sections. Section 2 deals with validation. The principles of KB validation are presented; their applications to rule bases and knowledge models are detailed; the use of explanation to refine knowledge is discussed. Section 3 deals with explanation. The principles of explanation in rule based systems are presented; the notions of reactive explanation and explanation dialog are detailed. Section 4 provides an overview of current issues in validation and explanation in systems engineering, semantic web and ontology.

2 Validation: Issues and Solutions

Research works on KBS validation aim to formalize with properties what KB validity means and to provide algorithms to prove those properties. Those properties depend upon the studied knowledge base or model. Once a KB is proved to be invalid, it must be corrected, i.e. refined. Most of the KBS are rule based systems. A rule has the following form "R_i : If conditions Then conclusion". In a first approach, that means that if the set of all the conditions are true, then the system can deduce by inference that the conclusion is true. For simplicity reason, we suppose that conditions and conclusion are facts.

This section is divided into four subsections. Section 2.1 presents different validation approaches in the literature, illustrated by simple examples on a propositional rule base. Section 2.2 presents the KB coherence; a classical KB validity property is proposed and implemented on rule bases. Section 2.3 presents a solution to valid a knowledge model that is not a rule base. Section 2.4 deals with the KB refinement, relying on the explanation notion.

2.1 Different Validation Approaches

The first research works showed that there exist different kinds of anomaly in a KB: the *redundancy*, the *incompleteness* and the *incoherence* (Nguyen et al. 1985). For each anomaly, formal definitions were proposed: (i) a KB is redundant if the same computed result can be deduced from the KB and a unique valid input, (ii) a KB is incomplete if the expected computed result of a valid input cannot be deduced from the KB and this valid input, (iii) a KB is incoherent if a contradictory computed results can be deduced from the KB and a valid input. Redundancy is ambivalent, it can highlight either problems in the KB or some robustness. Incompleteness focuses on the necessity to complete/enrich the knowledge, which is a difficult problem often studied during the knowledge acquisition phase (see chapter "Knowledge Engineering" of this volume). Incoherence is of great interest because it can prove the invalidity of a KB. Most of the works on validation has consisted in defining one of these

anomalies for a kind of KBS with a formal property and also in giving algorithms to check the property truth. Incoherency was especially studied in the literature.

Example 1 Let us consider the following rule bases RB where Tall is supposed to be a valid input:

1. RB = $\{r_1, r_2, r_3\}$ where r_1 : If Tall Then Strong; r_2 : If Strong Then GoodLooking; r_3 : If Tall Then GoodLooking
   RB is redundant: GoodLooking can be deduced by two distinct ways from RB and the valid input Tall.
2. RB = $\{r_1\}$ where r_1 : If Tall Then Strong
   If a test states that someone who is Tall must be GoodLooking, then RB is incomplete.
3. RB = $\{r_1, r_2\}$ where r_1 : If Tall Then Strong; r_2 : If Tall Then ¬Strong
   RB is incoherent: a Tall person will be in the same time Strong et ¬Strong.

Two kinds of solutions were studied to define good properties to validate a KB. The first solution, called *local solution*, consists in comparing the knowledge two by two. This local solution highlights only superficial anomalies: it does not take into account all the knowledge and the deductive capability of the inference engine. The second solution, called *global solution*, takes into account all the set of knowledge and the semantics of the deductive inference.

Example 2 Let us consider the following rule bases RB where Tall is supposed to be a valid input:

1. RB = $\{r_1, r_2\}$ where r_1 : If Tall Then Strong; r_2 : If Tall Then ¬Strong
   RB is locally incoherent.
2. RB = $\{r_1, r_2, r_3\}$ where r_1 : If Tall Then Strong; r_2 : If Strong Then GoodLooking; r_3 : If Tall Then ¬GoodLooking
   RB is locally coherent, nevertheless the facts GoodLooking and ¬GoodLooking can be deduced from RB and the fact base {Tall}.
   The local approach shows its limits: it cannot detect the incoherence, whereas the global approach gives a solution to detect it.

In the validation of a KB, the KB "*test*" is often distinguished from its "*verification*". Testing a KB consists in applying test cases, provided by the KB designer, on the KB and checking if the computed results are valid or not. A test case represents a valid input often associated with its expected output. It is considered as a reliable knowledge. Verifying a KB is performed without test cases; its main difficulty consists in computing valid inputs for the verification.

Example 3 Let us consider the following rule bases RB where Tall is supposed to be a valid input:

1. RB = $\{r_1\}$ where r_1 : If Tall Then Strong
   If a test case states that someone who is Tall must be GoodLooking, then RB is coherent but incomplete. This is a KB test.

2. RB = $\{r_1, r_2, r_3\}$ where r_1 : If Tall Then Strong; r_2 : If Strong Then GoodLooking; r_3 : If Tall Then ¬GoodLooking
   RB is globally incoherent. This approach does not use test cases: it is a KB verification.

The validation approaches that require specific knowledge for validation are distinguished from the ones that enable a validation without any other supplementary knowledge. The first validation approach is called "*semantic validation*"; the specific validation knowledge is provided in a set of specifications. These specifications are reliable knowledge given as reference, generally test cases or integrity constraints. The second validation approach is called "*syntactic validation*".

Example 4 In the first point of Example 3, the proposed approach is semantics because it uses a test case.
Let us now consider the following rule bases RB= $\{r_1, r_2\}$ where Tall is supposed to be a valid input, r_1 : If Tall Then Strong and r_2 : If Tall Then GoodLooking. Let us also consider the following validation constraint: If Strong And GoodLooking Then $\perp$, where $\perp$ means a contradiction. The semantics validation of RB shows an incoherence.

2.2 Knowledge Base Coherence

Numerous works studied rule bases coherence properties. To be valid, a rule based system (RBS) must have a coherent rule base. The different kinds of existing rule bases and their different semantics make a unique RBS coherency definition quite difficult to obtain. Nevertheless a lot of researches (Loiseau 1998; Pipard 1987; Ginsberg 1988; Rousset 1988; Ayel and Rousset 1990; Beauvieux and Dague 1990; Bouali 1996; Rousset and Levy 1996) considered (i) that rules are logical implications of which the condition part is composed of a conjunction of literals and the conclusion part of a unique literal, and (ii) that the inference engine is an algorithm which implements a modus ponens deduction, denoted $\models$, that is data driven.

Let us assume that **a KB is incoherent if a contradictory computed results can be deduced from the KB and a valid input**, which is a generalization of existing works definitions. Two difficulties must be emphasized to propose a corresponding formal definition. The first difficulty is to characterize what is a valid input; the second one is to deduce the computed results for each valid input.

In this chapter, a valid input is an input that checks a set of constraints. These constraints, explicitly given, can be test cases reformulated in rule or integrity constraints describing contradictory solutions, denoted by $\perp$.

Definition 1 Let KB be a knowledge base and Kb a constraints subset of KB. An input I *checks the constraints* of Kb If $I \cup Kb \not\models \perp$. Such an input I is called a *valid input*.

It is important to notice that an input is conventionally a subset of an input literals set given by the RBS designer.

Example 5 Let consider the rule base RB={r_1, r_2, r_3, r_4, r_5, R_6, R_7} where:
r_1 : If WageEarner ∩ Manager Then ¬FreeTime
r_2 : If ¬FreeTime Then ¬VoluntaryHelper
r_3 : If PartialTime Then VoluntaryHelper
r_4 : If StayAtHomeParent Then VoluntaryHelper
r_5 : If ¬FreeTime Then ¬Sportif
R_6 : If StayAtHomeParent ∩ WageEarner Then ⊥
R_7 : If StayAtHomeParent ∩ HasBaby Then ¬VoluntaryHelper
Let us consider the constraints subsets Rb= { R_6, R_7} where R_6 is an integrity constraints and R_7 a test case.
Let IL = {WageEarner, Manager, PartialTime, HouseWife, HasBaby} be a set of input literals given by the RBS designer.
Then, we have the two following input fact bases checking the constraints of Rb:
FB_1 = {WageEarner, Manager, PartialTime}. As a matter of fact, {WageEarner, Manager, PartialTime} ∪ { R_6, R_7} does not allow a contradiction to be deduced.
FB_2 = {StayAtHomeParent, HasBaby}.
Let us notice that there exist several other input fact bases that check the constraints. We are interested in $FB1$ and $FB2$ for which there exist incoherencies (cf. Example 6).

A KB is coherent if no contradictory computed results can be deduced from the KB and any valid input checking the given constraints.

Definition 2 Let KB be a knowledge base, Kb being a constraints subset of KB. KB is *Vcoherent* if for each input I checking the constraints Kb, I ∪ KB $\not\models \bot$.

Example 6 Let consider FB_1 of Example 5 a valid input fact bases, we have $FB_1 \cup \{r_1, r_2, r_3\} \models \bot$ and so RB is Vincoherent. We also have $FB_2 \cup \{r_4, R_7\} \models \bot$.

Different solutions were proposed to prove the Vcoherence of a KB. In a propositional formalism, a solution is to generate all the input fact bases FBi that check the constraints, and then to compute for each of them the possible deductions FBi ∪ RB to check whether a contradiction can be deduced. Such an approach is cost computed because the number of fact bases may be exponential. The computation can be restricted to the maximal valid fact bases. A valid fact base is maximal if it is not the subset of another maximal valid fact base. Such a solution is also cost computed.

Other methods were proposed. Some use Petri nets, others Clause Management Systems (CMS). CMS provides an elegant solution to obtain, under assumption labels, the minimal fact bases that are sufficient to deduce ⊥. The first CMS (Reiter and de Kleer 1987) were restricted to propositional rule bases. The examples given above are propositional examples: the condition and conclusion are composed of literals which are true or false. In such examples, rules can be seen as logical implications. These approaches were extended to attribute/value rules then to first order

rules, i.e. DATALOG rules. The attribute/value rules are rules of which literals can take any value, not only the true or false values like in propositional rules. The DATALOG rules are rules that use predicates and variables, a subset of Prolog syntax; they are logical implication of predicate logic.

Two reasons can explain the Vincoherence of a KB: either integrity constraints are missing or knowledge of the KB includes contradictions. Let us notice that the Vcoherence definition has two limits. On the one hand, a KB can be Vincoherent despite it is coherent and, on the other hand, a KB can be incoherent despite it is Vcoherent. The first limit, although theoretically awkward, can be balanced by the fact that when a KB is Vincoherent, it can be refined by adding missing integrity constraints such that the input that show the Vincoherence become invalid: the constraints subset Kb was therefore only incomplete. Regarding the second limit, it is impossible to have a perfect formal definition of KB coherence, except if there exists a complete and coherent formal model of the knowledge, which is most of the time unrealistic. The incoherence of a Vcoherent KB can be explained by its incompleteness. So, the Vcoherence of a KB is only a partial guarantee of its coherence and so validity.

2.3 Models Validation

Beside numerous works on RBS, some research works focused on models validation (Haouche and Charlet 1996; Lee et al. 2002; Shanks et al. 2003; MoDeVA 2009) and on semantic networks validation (Hors and Rousset 1996; Rousset and Levy 1998; Dibie-Barthélemy et al. 2006). In this subsection, we show how the validation principles presented above can be applied to conceptual graphs models (Sowa 1984; Mugnier and Chein 1996) (see chapter "Reasoning with Ontologies" of this volume).

The conceptual graph model makes a clear distinction between the terminological knowledge (i.e. the support) and the assertional knowledge (i.e. the conceptual graphs).The "*syntactic validation*" allows one to check that the KB was well built. The "*semantic validation*" allows the coherence and the completeness of the KB to be checked using constraints. Such a semantic validation is partial since it depends upon the positive and negative constraints given by the KB designer. A KB is said Vcoherent if it satisfies all the input negative constraints, which represent the knowledge that must not be in the KB. A KB is said CP-complete if it satisfies all the input positive constraints which represent the knowledge that must be in the KB.

Example 7 Let us consider the KB of Fig. 1 composed of the conceptual graph G. It is incoherent: it does not satisfy the negative constraint NC which means that a cat must not paint. It is also incomplete: it does not satisfy the positive constraint PC which means that a painter must paint at least a painting (e.g. Picasso does not paint a painting).

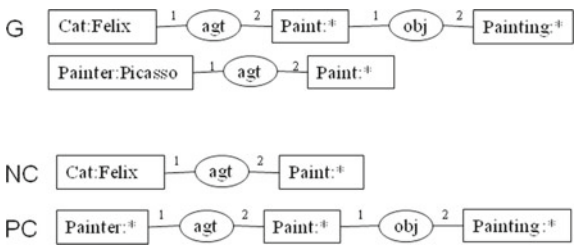


Fig. 1 A conceptual graph G, a negative constraint NC and a positive constraint PC

2.4 Validation, Refinement and Incoherencies Explanation

When using a KBS, an input is given and results are computed using some knowledge of the KB. The trace which enables to follow the reasoning that was made is a couple composed of the input and the used pieces of knowledge. Such a trace is used in explanation as well as in validation to refine a KB. The refinement aims to modify an incoherent KB in order to restore its coherence. To do so, first the different possible reasoning traces which enable the KB incoherence to be proven are computed, then possible explanations of the KB incoherence are build. An explanation is composed of a set of elements, each element being extracted from each computed trace such that it is able to explain the incoherence identified in the trace.

Definition 3 Let KB be a knowledge base, I a valid input and O a valid output such that I ∪ KB ⊨ O. A *trace* of O is composed of a couple (I', KB'), with I' a part of I, KB' a subset of KB which checks I' ∪ KB' ⊨ O. A *trace* (I', KB') is *minimal* if there does not exist another trace (I", KB") of O such that I" ⊨ I' and KB" ⊆ KB'.

The previous definition is not a priori restricted to rule base.

Example 8 Given Example 5, a minimal trace of ¬VoluntaryHelper is ({WageEarner, Manager}, {r_1, r_2}).

As seen in Sect. 2.2, a KB is Vincoherent either because some constraints are missing in the KB or because some knowledge of the KB are invalid and so must be corrected. The traces can be a powerful help to the designer, allowing him to find the missing constraints to add or the pieces of knowledge to remove or correct.

Example 9 Given Example 5, T_1=({WageEarner, Manager, PartialTime}, {r_1, r_2, r_3}) and T_2 = ({StayAtHomeParent, HasBaby}, {r_4, R_7}) are the unique minimal traces of a contradiction.

An explanation of the KB Vincoherence is composed of a set of inputs and a set of invalid pieces of knowledge, such that when each input is added in the condition part of an integrity constraint and each piece of knowledge is removed, the KB becomes Vcoherent.

Definition 4 Let KB be a knowledge base and IT the set of minimal traces of $\perp$ characterizing the Vincoherence. An *explanation* of IT is a couple (I, K) where I is a set of inputs and K a set of knowledge such that $\forall$(It, Kt) $\in$ IT, $\exists$ I_i de I | It $\models$ I_i or K $\cap$ Kt $\neq \emptyset$.

To restore the KB coherence, only minimal explanations that do not contain useless information are considered. The concept of "strong minimality" relies on the fact that constraints are supposed to be reliable and consequently cannot be deleted and cannot contain an invalid condition part. Such assumptions are often made by works on validation to avoid non relevant solutions to be obtained.

Definition 5 Let KB be a knowledge base and Kb a constraints subset of KB. An *explanation* (I, K) is *minimal* if there does not exist another explanation (I', K') such that K' $\subset$ K or such that K'=K and $\forall I_i$' of I', $\exists I_j$ of I such that I_i' $\models$ I_j. An *explanation* (I, K) is *strongly minimal* if I $\cap$ Kb = $\emptyset$ and $\forall I_i$ of I, $\nexists$ a constraint "If Conditions Then..." | Conditions $\models$ I_i.

Example 10 There exist 12 minimal explanations and 4 strongly minimal explanations of IT=$\{T_1, T_2\}$ in Example 5: ({WageEarner, Manager, PartialTime}}, $\{r_4\}$), ({},$\{r_1, r_4\}$), ({}, $\{r_2, r_4\}$), ({}, $\{r_3, r_4\}$). The first explanation means that RB to which is added the integrity constraint "If WageEarner $\cap$ Manager $\cap$ PartialTime Then $\perp$" and removed the rule r_4 is Vcoherent. The second explanation means that RB to which is removed the rules r_1 and r_4 is Vcoherent.

The traces computation can be performed in different manners; it can for instance be obtained with a CMS. The explanations computation can be performed by algorithms extending the hitting set algorithm (Reiter 1987).

3 Explanation: Issues and Solutions

First of all, let us notice that "explanation" in this section consists in justifying the results of a KBS. Unlike the explanation of Vincoherence defined in the previous section, which is a part of a formal validation used by system's designers, the explanation can address various kinds of users – experts, professionals of the domain, and even beginners – in various use cases: decision-making aid, co-design or co-implementation, learning and teaching or coaching. The achievement of these various objectives are based on the system capacity to explain its results and its reasoning in an understandable and oriented way to any user, even interactive.

Expert Systems appeared at the end of the 1960 s, but the problem of explanation really emerged only in mid-1970 s.

This section is composed of four subsections. Section 3.1 describes how explanation arose from the track of reasoning to gradually get loose from it and become, as detailed in Sect. 3.2, a full task and consequently a research domain. Sect. 3.3 shows

how the increasing consideration of the user in the construction of the explanation raises leading to an explanatory dialogue. Finally, Sect. 3.4 deals with the limits associated with such a complex task.

3.1 From the Track/log of Reasoning to the Explanation

The first Expert System to incorporate an explanatory capacity was MYCIN (1972–1980) (Shortliffe 1976) in the medical domain. In 1973, Shortliffe implanted a command "RULE" for the debugging. This caused the display in LISP of the last activated rule. He noticed that a translation in English would facilitate the understanding and could bring information useful to a user who is not a computer specialist. This way every MYCIN rule had a double representation: the coding in LISP used by the inference engine and its translation into English, fit to be displayed to the user. Afterward, Randy Davis endowed MYCIN of a tree of reasoning ("history tree") storing the sequences of activated rules, and he transformed the command "RULE" in "WHY?". A succession of "WHY?" allowed step by step backtracking in the reasoning tree. A "HOW?" command allowed the following of the various branches of the reasoning tree after the resolution.

With MYCIN, the track of the reasoning becomes "explanatory" as reflection of an understandable knowledge, expressed in the vocabulary of the user's vocabulary. The explanation module is independent from the domain and is relatively easy to implement, because the knowledge of the system is represented in a uniform way (e.g. "production rules").

However, the use of MYCIN for learning and teaching Use Cases in the system GUIDON (1977–1981) (Clancey 1986) had to show the "explanatory" limits of the track of reasoning, even if GUIDON was an elaborated system including educational rules and a tutorial module.

The use of a uniform formalism (e.g. production rules) allows the capture of the expertise of a domain under the shape of a set of "items of knowledge" easier to exploit, but it has the inconvenience to translate uniformly in rules the various underlying relations of the domain. Several types of knowledge, even implicit, coexist within the KB: rules of identification, rules of common sense, causal rules... Also, clauses which constitute the premises of rules may have different status: some correspond to the context of rule activation and are present in several rules, others are given to block the rule activation in certain situations, and only few clauses are directly associated to the conclusion. Only these clauses really make sense according to the domain of resolution.

In addition, a knowledge base handling a real domain contains hundreds even thousands of rules which, during a resolution, activate thousands even hundreds of thousand reasoning inferences. Browsing the steps of these inferences, logged in the track of reasoning, lose any explanatory virtue, because the system does not know how to highlight the key elements in its reasoning. The track is too voluminous and flat as well. Also the activation order of rules depends on the resolution engine and

on the KB organization (i.e. rules order in the base, and clauses order in the premises) and does not follow the reasoning of a human expert, what makes it quite difficult to understand and to accept especially for an experienced user.

Furthermore, the rule base specifies only knowledge used for resolution: underlying mechanisms and relations of the domain are not explicit. For example, the rule "If the patient is less than 8 years old Then not to prescribe tetracycline" doesn't justify this contraindication. The explanatory knowledge (Safar 1987) is indeed missing: "use of tetracycline at the childhood → deposit of the drug on bones in development → definitive blackening of teeth → socially unwanted coloring → Do not use a tetracycline for a child".

These limits are specific to expert systems of first generation, ever since, research efforts focus on creating explicit and organized knowledge, as well as elaborating knowledge and specific reasoning in the production of explanation. The explanation took then its autonomy towards the track of the reasoning to become a separate task.

3.2 Explanation as a Specific Task

Making the KB more explicit and better organized can increase its "understandability", and this way facilitates KB creation and update its explanation capacity. These orientations founded the Second Generation of Expert Systems (SGES). Two striking examples are NEOMYCIN (Clancey 1986) and CENTAUR (Aikins 1983). The knowledge is typified and represented differently according to its role: strategic knowledge which controls rules activation; domain heuristics; knowledge of support which justifies heuristic rules. . .

As for the explanatory capacity, an additional line of research exists to improve and/or to modify the track of reasoning to make it more relevant and to adapt it according to the supposed level user expertise, for example in the systems BLAH (Weiner 1980) and XPLAIN (Swartout 1983).

These research directions clarifying specific knowledge allow justifying the reasoning but also taking into account the user (i.e. his supposed level of knowledge) in order to generate a structured explanation in natural language which doesn't follow the track of resolution anymore.

Meanwhile, other works focus on the automatic generation of textual descriptions of domain complex concepts: example TEXT (McKeown 1985), which uses rhetorical plans, or the TAILOR system, which may combine two plans in an explanation (Paris and McKeown 1987) in order to produce a description adapted to the interlocutors expertise level. These reasonings are essentially based on the domain representation, as per the main goal to produce descriptions of specific concepts and domain relations.

As shown in (Kassel 1987; Safar 1987) or (Wick and Thompson 1989), these approaches are complementary to explain a reasoning: it is not only necessary to reason on the resolution to produce a relevant synthesis or to answer negative questions ("why not X?"), but the system has also to reason on the domain in order to explain to

a good level of abstraction (e.g. by neglecting the basic concepts, making explicit the links between concepts, or using analogy with scenarios of breakdown). This way, explanations produced by the EES (Explanatory Expert Systems) are rather close to "natural" explanations supplied by human experts, where line of explanation often diverges from the line of reasoning.

For example, the REX system ("RECONSTRUCTIVE EXPLAINER") (Wick and Thompson 1989) can produce such explanation which seems however rather long: "*I attempted to find the cause of the excessive load of the concrete dam. Based on slow drainage and high uplift pressures, I made a first hypothesis. In studying the causal relationships, I found that settlement of the dam would cause the slow drainage which would in turn create uplift pressures acting on the dam, thereby suggesting settlement as the problem. However, based on the non-uniform damage of the broken pipes in the foundation, I was able to refute this hypothesis. Again in looking at causal processes I noted that settlement would cause crushed-like damage to the drainage pipes whereas erosion of soil would cause the observed selective damage. Therefore, I concluded erosion was causing the excessive load.*"

In order to bring the flexibility to schema based explanation, (Cawsey 1990) proposed the use of plans able to produce varied explanations on complex domains. In such explanation grammars, texts patterns are attached to the leaves of the explanation tree, and their concatenation produces the final explanatory content. These grammars make explicit various hierarchical structural levels of the explanation and allow the use of several strategies of explanation. But, on the other hand they are very dependent on the domain to be explained.

These grammars of explanation launch an approach based on the planning of the explanation which is particularly fruitful (see (Paris et al. 1990), pp. 1–226, the EDGE system (Cawsey 1990), or (PenMan Hovy 1988)). Plans allow a dynamic construction of the explanation by bringing more flexibility than schemes. The hierarchical decomposition of plans produces an explicit representation of the explanation, both on the informative level (i.e. contents) and the explanatory goals (e.g. define a concept).

Among the researches implementing this planning approach, one of the most striking is the EES project Explainable Expert Systems (Neches et al. 1985) in which Cécile Paris and Johanna Moore developed an explanation planner module (Paris et al. 1991). Their system, PEA ("Program Enhancement Advisor"), can answer predefined set of demands concerning the reasoning itself, as well as a system reasoning choice (e.g. rules), and its domain expertise. It is also able to manage justification requests (why?), negative justification (why not?) or method choice (why M1 rather than M2?). The plan contains successive communicative goals, but also rhetorical relations between various elements of explanation translated by two types of plan operators.

Using a deep structure in order to generate explanation in natural language is still a "State-of-the-Art" methodology which can be used in various applications, for example (Bader 2013) who use DRS (Discourse Representation Structure) to produce relevant explanation in smart environment.

A new stage has been archived thanks to this approach, because henceforth the EES (Explanatory Expert systems) are not only able to produce complex and structured explanations, but they also know why they produce these explanations thanks to the explicit operators implementing the goals. They are then capable to explain "again" and in a different manner if the user expresses a lack of understanding (e.g. it will be clear from the feedback following the first explanation supplied by the SEE). They will also be able to answer questions related to it ("follow-up questions"). The explanation which becomes a full specific task appears in the interaction with the user.

3.3 From Reactive Explanation to Explanatory Dialogue

We saw in the previous subsection how the reasoning could be elaborated allowing the EES to plan an explanation. Nevertheless, as underlines (Swartout et al. 1991), it remains complicated to supply, in every case, a "sufficient" explanation. The explanation given by the system can leave unsatisfied needs and raise new needs (e.g. follow-up questions). In the first case it will be necessary to re-explain differently, in the second, to clarify or to bring the necessary precision. As a matter of fact, explanation is an interactive process. Endow EES with capacity of an interactive management of the explanation implies to converge two Domains of Research: the Human Machine Dialogue (HMD) and the Explanation in KBS.

These two AI domains have been developed in parallel since the 60s without explicit convergence. Thus, HMD Systems stay mainly dedicated to simple tasks solving (e.g. Info retrieval) which do not require an explanation elaboration, and, on other hand, KBS supply monolithic explanations and rudimentary interaction. Nevertheless, with projects like EDGE and EES, or thanks to multidisciplinary research groups like "GENE" (Generation of Negotiated Explanations (M.Baker et al. 1995)) a real convergence begins to take shape and to become a reality because of their "retrospectively obvious" complementarity (Charnay 1999).

These approaches are based on analysis of real conversations spontaneous or finalized (i.e. oriented by the achievement of a task). The interlocutors can be either human at the same level of expertise, or in situation of learning/teaching which suggests knowledge transfer, or man-machine with the method of "Wizard of Oz" often used during design phase of an HMD system, in which an expert uses a device in order to simulate the targeted system without users knowing. The corpora of collected dialogues allow various aspects of interaction to be analyzed and this way many linguistic points of view (e.g. lexicons, typology of statements, management of the exchanges, argumentation) as from the point of view of the expressed knowledge and those implicit underlying asserted during the dialogue.

Besides, mutual understanding achievement is more important for the "success" of an interactive explanation than a formal property that the explanation object would satisfy. As a result, it becomes necessary to effectively produce such explanations in order to validate the result of this research with real users, and consequently to

develop testable EES. This brings the entire problem of the evaluation and validation of interactive systems in natural language (Allemandou et al. 2007; Devillers et al. 2004).

These studies of human explanations corpus, also allow the definition of quality criteria for the generation of an explanatory speech, as (Cawsey 1990): (i) cohesion (i.e. surface) and coherence (i.e. semantics) of the text; (ii) adequacy to the communicative intention of the system; (iii) taking into account the knowledge and the goals of the interlocutor; (iv) taking into account the context and the previous speech; (v) interruptions management, the "relevant" questions, the acknowledgments and explicit agreements.

The GENE GROUP (GEneration of Negotiated Explanation 1992–1999) was a multi-disciplinary research group, developing different points of view about a common object: linguistics and computational linguistics, AI, cognitive science, NLP. By working together, trying to interconnect different models and converge different approaches, they demonstrate that Explanatory Dialogue is the result of an interactive co-construction process, in which each co-constructor has his/her own knowledge and tries to achieve various goals. Some of them are shared explicitly or not especially in collaborative dialogues. Sometimes goals of each participant are in opposition, and then dialogue becomes more argumentative. Goals can also be complementary, for example when dialogue deals with a documentary search in a library or in training or teaching situation.

Each interlocutor builds a "model" of his/her interlocutor, including hypothesis regarding his/her knowledge, goals, and psychological profile.

As far as each person has a relative level of expertise about the domain and the task, argumentative phenomena often arise and can turn the dialogue into a real argument. In a Human-Machine-Dialogue system perspective, such co-elaborated explanatory dialogue suggests that the system is able to manage mixed initiative, even if the dialog generally follows a main goal, which should be accomplished by this dialogue (e.g. converge on a medical diagnosis).

Each participant involved in a discussion has knowledge on the domain (i.e. "know-what" and "know-how") which is partially shared but is not necessarily identical. In case of a dialogical explanation he/she pursues an elaborate plan dynamically according to his/her purposes (common or individual) taking into account actions of his/her interlocutor. It includes a set of information shared explicitly, as the processed case, as well as the data necessary for the task (e.g. the clinical data for a medical diagnosis), as well as everything mentioned in the dialogue. All these elements are used for the interaction, but they can also become the object of the discussion. So, a speaker may use a knowledge of the domain-object "theoretically", thus considering it as common to both interlocutors, or to attribute it to his/her interlocutor – "so relegating" it to the rank of belief (e.g. by the use of expressions like "in your mind", "according to you" . . .) –.

Being an observable phenomenon following rules (at least interactive) and manipulating data (knowledge) that is supposed to be formalized, the negotiated explanation can be modeled. Even before being linguistic, it is a phenomenon of communication and interaction. Thus we can design a model of interaction relatively independent

from the medium of interaction (Natural Language, GUI), to obtain a deep representation of the negotiated explanation, in a computing perspective by the system as well as by the user. Such model, in an interactive way, intends to manage jointly according to 3 dialogue plans, the argumentation and the explanation. A proposal of modeling such a SEED (Dialogical Explanatory Expert System) has been made in (Charnay 1999). Such deep modeling facilitates incorporating these functions in a Graphical Interface, as shown in (Baker et al. 1996). Some recent works are following this approach, such as (Walton 2007; Bex et al. 2012) and propose hybrid models to distinguish argument and explanation.

As well as explanation, argumentation is seen here in its interactive dimension, but the link can be made with "argumentative reasoning" as discussed in chapter "Argumentation and Inconsistency-Tolerant Reasoning" of this volume. Such reasoning could contribute to the construction of arguments elaborated from the KB by the system and by the contradictory knowledge brought by the user in the dialogue (e.g. highlight factual elements from which we could deduce in a skeptical manner or argumentative that they oppose to this contradiction).

Such multidisciplinary approach, with multiple perspectives models stay a fruitfull paradigm, as shown in the AAAI/ECAI/IJCAI Workshops "ExaCt Explanation-aware Computing" from 2005 to 2012 (Roth-Berghofer and Schulz 2005; Roth-Berghofer et al. 2007, 2008, 2009, 2010, 2011, 2012).

3.4 The Dialogical Explanation: the Limits of One Paradigm

The increasing elaboration of models designed to manage interactive explanations and even real explanatory dialogues, quickly confronts this research domain by the end of the 1990s, with serious difficulties to acquire, model/design, and maintain these knowledge. On the other hand, application fields in particular the medical domain are confronted with the constant evolution of the knowledge, cf. (Bouaud et al. 2008). Indeed, the simultaneous management of various aspects, each intrinsically complex, dramatically increase the necessary knowledge. Then, the typically proposed multi-expert systems often stay for the greater part at the stage of model or of "Proof of Concept": the passage to the scale processing operationally a real domain stumbles over the human and economic costs of such development.

The explanation intended for end-users can be indeed estimated and validated only in real conditions of use, with real target users. This is even more complex when it is about human-machine interaction (see (Allemandou et al. 2007) for which reference data, i.e. the "good" dialogue to be accomplished, doesn't exist). This adds to the intrinsic complexities of the addressed problems. Paradigm of the classic Explanation of the 70 s in 90 s probably reached its acquisition and representation limits of various kinds of knowledge needed, as well as the operational validation of interactive systems.

The search effort then refocused particularly on these problems of knowledge engineering (i.e. acquisition, modeling and evolutionary maintenance of the KB, in

particular ontologies as discussed in Sect. 4 and chapters "Reasoning with Ontologies" and "Knowledge Engineering" of this volume), also strengthened by new perspectives of knowledge sharing and Big Data.

It is however interesting to notice that research in explanation continued, in particular in the psycholinguistics and cognitive sciences domains, to give rise to purely explanatory models, either interactive phenomena (Dessalles 2008), or the co-construction of the knowledge (Baker 2009). Such multi-dimensional and hybrid approaches, mixing explanation and argumentation in dialogue, are still fruitful (see (Bex-Walton 2016)).

In parallel ExaCt "Explanation-aware Computing" Workshops (2005-2012) show various operational systems using explanation in interactive task combining GUI and natural language (Roth-Berghofer and Schulz 2005; Roth-Berghofer et al. 2007, 2008, 2009, 2010, 2011, 2012).

4 Current Issues

The works on validation and explanation concern nowadays several different computer research fields. Such a spread makes their identification harder but also shows their crucial issues and interests in many different fields. This is due on the one hand to the KBS increasing complexity which are composed of different kinds of knowledge with different semantics and, on the other hand, to the emergence of new issues around knowledge and data management on the semantic web (see chapter "Knowledge Engineering" of this volume). This section gives an overview of validation and explanation current issues. Section 4.1 presents briefly the validation issues in systems engineering. Section 4.2 addresses the validation and explanation issues in semantic web. Finally, Sect. 4.3 deals with the validation and explanation issues in ontology.

4.1 Validation and Systems Engineering

Thanks to the International Council on Systems Engineering (INCOSE[1]), systems engineering is an interdisciplinary approach enabled to ensure the realization of successful systems. It focuses on defining customer needs and required functionality early in the development cycle, documenting requirements, then proceeding with design synthesis and system validation. Its life cycle is usually comprised of the following seven tasks: state the problem, investigate alternatives, model the system, integrate, launch the system, assess performance, and *re-evaluate* (Bahill and Gissing 1998) where 're-evaluate' means observing outputs and use them in order to improve the system, its inputs, the product and the process. The norm ISO/IEC/IEEE

[1] http://www.incose.org/AboutSE/WhatIsSE.

15288:2015 establishes a common framework of process descriptions for describing the life cycle of systems created by humans.

If we focus on the re-evaluation task, the PMBOK guide (Institut 2013), a standard adopted by IEEE, gives in its fourth edition the two following definitions:

- "*Validation*. the assurance that a product, service, or system meets the needs of the customer and other identified stakeholders. It often involves acceptance and suitability with external customers. Contrast with verification."
- "*Verification*. The evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition. It is often an internal process. Contrast with validation."

4.2 Validation, Explanation and Semantic Web

According to the World Wide Web Consortium (W3C),[2] "the term *Semantic Web* refers to W3C's vision of the Web of linked data. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network." (see Chapter Reasoning with Ontologies of Volume 3). The validation issues in semantic web mainly concerns the quality of the web of data and the trust of the web data sources. It supplements the validation issues which could have been taken into account during the web data sources building in a system engineering process. We first present the quality criteria of the web of data and then the trust criteria of the web data sources. Finally, we briefly discuss about the explanation issues in semantic web.

4.2.1 Web of Data Quality

The development and standardization of semantic web technologies has resulted in an unprecedented volume of data being published on the web as Linked Data (LD) which is, unfortunately, of variable quality. Zaveri et al. (2016) proposed a survey of LD quality. Eighteen different data quality dimensions were identified and divided into four main groups: (1) the accessibility dimensions, (2) the intrinsic dimensions, (3) the contextual dimensions and (4) the representational dimensions. Thanks to our validation preoccupation, we focus on the intrinsic dimensions which are independent of the users context and study whether "information correctly (syntactically and semantically), compactly and completely represents the real world and whether information is logically consistent in itself". Five intrinsic dimensions were identified: syntactic validity, semantic accuracy, consistency, conciseness and completeness. The *syntactic validity* is defined as "the degree to which an RDF document conforms to the specification of the serialization format". The metrics identified for syntactic

[2] https://www.w3.org/standards/semanticweb/.

validity can be for instance detecting syntax errors using validators,[3,4] or detecting whether the data conforms to the specific RDF pattern (Kontokostas et al. 2014). The *semantic accuracy* is defined as "the degree to which data values correctly represent the real world facts". The metrics identified for semantic accuracy can be for instance detection of inaccurate values via crowd sourcing (Acosta et al. 2013). "*Consistency* means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms". The metrics identified for consistency is to load the knowledge base into a reasoner and check whether it is consistent. Finally, "*conciseness* refers to the minimization of redundancy of entities at the schema and the data level" and "*completeness* refers to the degree to which all required information is present in a particular data set".

In the LD quality preoccupation, the early draft about the Data Quality Vocabulary[5] proposed by the Data on the Web Best Practices Working Group of the W3C provides a framework in which assertions can be made about a data set quality and appropriateness for given tasks.

4.2.2 Web Data Sources Trust

Building a "web of trust" is a crucial issue of the semantic web: it is essential to build a web that people can trust. Let us notice that the web data source trust corresponds to the trustworthiness dimension of the context dimension group in (Zaveri et al. 2016) defined as "the degree to which the information is accepted to be correct, true, real and credible". The validation of the web data source trust criteria consists in checking that the web data source effectively contains what it pretends to and that its information is reliable. This validation can for instance consists in judging the web data source credibility (i.e. number and quality of input and output hyperlinks), checking whether the web data source property is reliable thanks to its URI or merging and comparing data from different web data sources. One of the most important metrics identified to validate the web data source trust relies on data provenance for which many works were done and led to the PROV model[6] recommended by the W3C Provenance Working Group, chartered to specify a representation of provenance to facilitate its exchange over the Web (Moreau et al. 2015).

4.2.3 Explanation and Semantic Web

The semantic web can be seen on the one hand as a huge KB which allows the development of question answering works (Corby et al. 2006; Lopez et al. 2013; Unger et al. 2015) and, on the other hand, a way to dynamically discover new services

[3] http://w3c.github.io/developers/tools/.

[4] https://validator.w3.org/unicorn/.

[5] https://www.w3.org/TR/2015/WD-vocab-dqv-20150625/.

[6] https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

by the composition of web published services (Rao and Su 2004; Sheng et al. 2014). In the second point, autonomous artificial agents have to browse the web, according to an end-user request, in order to compose relevant services to find the good answers (e.g. a multi-steps and multi-modes travel with cost and date constraints). The web semantic languages such as RDF-S or OWL (see chapter "Semantic Web" of volume 3) allow causal rules and simple inferences to be expressed and are appropriate to produce explanations required to justify the computed results to the end-user (Haynes et al. 2009) or to the "trust negotiation" (Bonatti et al. 2006) between intelligent agents on the web. Moreover, as a huge KB, the semantic web is a powerful means to enable argumentations that are of great interest in social web[7] (Rahwan et al. 2007; Schneider et al. 2013; Cabrio et al. 2013) and decision making contexts (Thomopoulos et al. 2015). Decision-making often requires discussion not just of agreement and disagreement, but also the principles, reasons and explanations driving the choices between particular options. This is also the preoccupation of works on causality as discussed in chapter "A Glance at Causality Theories for Artificial Intelligence" of this volume.

4.3 Validation and Ontology

The development of semantic web has resulted not only in an unprecedented volume of published data but also in an increasing number of ontologies used to model and enrich semantically these data. Such ontologies also appear as an inescapable means for the data and knowledge exchange between the different web data sources (Staab and Studer 2009). The quality of the published data therefore depends upon the quality and validation of the ontologies bringing their semantic, which have become a significant challenge.

The ontology engineering (Sure et al. 2009), inspired from the knowledge engineering (see chapter "Knowledge Engineering" of this volume), is an active research area which propose methods, tools and languages to build ontologies. It deals especially with development process of ontologies and their life cycle. The development process of an ontology can follow a particular methodology like NeOn (Suárez-Figueroa et al. 2015) where knowledge acquisition, documentation, configuration management, evaluation, and assessment should be carried out during the whole ontology network development. The life cycle contains in particular the *evaluation* step of the ontology, which is composed of the sub steps verification and validation. The *verification* sub step deals with the correct building of the ontology whereas the *validation* ones with the building of the "good" ontology. Vrandecic (2009) provided a panorama of current works on ontology verification and Obrst et al. (2007) on its validation. The works on ontology evaluation are inspired from works on KB validation. They study the same anomalies (cf. Sect. 2.1) that are the redundancy, the incompleteness and the incoherence but also some new validation criteria such as the

[7] https://www.w3.org/2005/Incubator/socialweb/XGR-socialweb-20101206/.

semantic validity of ontology according to the domain knowledge, its adaptability, its clarity or its quality thanks to existing ontologies, especially studied in works on ontology alignment (Euzenat and Shvaiko 2007) (see chapter "Semantic Web" of volume 3). Most of the works on ontology evaluation defined a generic quality evaluation framework (Gómez-Pérez 2004; Brewster et al. 2004; Brank et al. 2005; Gangemi et al. 2006; Duque-Ramos et al. 2013; Poveda-Villalón et al. 2014), others dealt with ontology verification such as (Guarino and Welty 2004; Schober et al. 2012) and in recent times methods pattern-based evaluation emerged (Gangemi and Presutti 2009; Djedidi and Aufaure 2010; Presutti et al. 2012). Each work proposes its own methods for evaluating the quality of an ontology and standards remain unfulfilled.

The ontologies may also be used in the conformity control of a job domain and therefore allow the end-user to better understand the reasons of an eventual non-conformity of the system (Yurchyshyna et al. 2008) or enable to better manage the evolution of the domain ontology (Djedidi and Aufaure 2009). It also provides a way to give reasoning and explanation capabilities to intelligent agents, especially in simulation domain for conditioning learning (e.g. "serious game" in military domain or crisis management). The ontologies then become an essential elements of immersive ITS (Intelligent Tutoring Systems) (Lane et al. 2005; Nkambou et al. 2010).

The ontology validation requires nowadays to take into account the evolutionary, dynamic, interdependent and distributed character of the knowledge it models. The ontology takes advantage of knowledge extracted from external sources which constantly evolve and are interdependent (e.g. knowledge from the web of data or social networks). Moreover the interoperability preoccupations of current systems (e.g. exploitation of ontologies in networks, management of ontologies in distributed or pair-to-pair systems) arise new problematics for the study of their coherence. The ontology coherence cannot no more be studied independently, but has to be studied in an ontologies network, each ontology being independently built, taking into account their reasoning power. Some works (Chatalic et al. 2006; Nguyen et al. 2008) proposed solutions using algorithms that reason with inconsistencies.

5 Conclusion

Whereas the software engineering has proposed a response to the crisis of software at the end of the sixties, the knowledge engineering has proposed a response to the crisis of expert systems at the end of the eighties. The validation researches on KBS are part of the knowledge engineering researches. They benefit from the logical semantics often associated with knowledge to provide elegant solutions to validation issues that were studied for several years in software engineering. The final interactive phase of validation to correct KBS, often called refinement phase, benefit from the modularity and declarativity of knowledge to provide original solutions to explain, at a high level, the system invalidities.

The KBS were conceived (e.g. MYCIN) both as a resolution system to obtain solutions and as a system to explain solutions, which opened new perspectives to computer science works. This preoccupations of having understandable and convincing computed results are also at the heart of decision making works at the frontier of applied mathematics and psychology. The interactive tutoring systems and the engineering of knowledge memory are two examples of numerous works requiring further researches in explanation. The tomorrow computer science works will have to act that several problems require the co-resolution of a computer and a human and to think deeper about how computer and human can better exchange their knowledge in a mutual comprehensive way.

Next to those traditional uses of KBS to explain solutions (Chen and Pu 2012; Biran and Cotton 2017; Brinton 2017), a new domain of research, called explainable artificial intelligence, is recently emerging in connection with machine learning models. Its aim is to provide an "intelligence" that can be understood. Machine learning researches focus mainly on how to obtain new knowledge, but the users need to be convinced of the truth and accuracy of those knowledge to solve their problems.

References

Acosta M, Zaveri A, Simperl E, Kontokostas D, Auer S, Lehmann J (2013) Crowdsourcing linked data quality assessment. In: The semantic web - ISWC 2013 - 12th international semantic web conference, Sydney, NSW, Australia, October 21–25, 2013, Proceedings, Part II, pp 260–276

Aikins JS (1983) Prototypical knowledge for expert systems. Artif Intell 20:163–210

Allemandou J, Charnay L, Devillers L, Lauvergne M, Mariani J (2007) Simdial - un paradigme pour évaluer automatiquement des systèmes de dialogue homme-machine en simulant un utilisateur de façon déterministe. *Traitement Automatique des Langues*, 48:1 Principes de l'évaluation en Traitement Automatique des Langues

Ayel M, Rousset M (1990) La cohérence dans les bases de connaissances. Editions Cépaduès

Bader S (2013) Generating explanations for Pro-active assistance from formal action descriptions. Springer International Publishing, Cham, pp 19–31

Bahill AT, Gissing B (1998) Re-evaluating systems engineering concepts using systems thinking. IEEE Trans Syst Man Cybern Part C 28(4):516–527

Baker M (2009). Argumentative interactions and the social construction of knowledge. In: Argumentation and Education: Theoretical Foundations and Practices, San Francisco, CA, USA. N.M. Mirza and A.-N. Perret-Clermont, Springer, Berlin, pp. 127–144

Baker M, Charnay L, Joab M, Lemaire B, Safar B, Schlienger D (1996) Incorporating functionalities of expert medical critiquing dialogues in the design of a graphical interface. In: 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, 0:122

Beauvieux A, Dague P (1990) A general consistency (checking and restoring) engine for knowledge bases. In: European conference of artificial intelligence ECAI'90, pp 77–82

Bex F, Walton D, Budzynska K (2012) Argument and explanation in the context of dialogue. In: Explanation-aware computing, papers from the 2012 ECAI workshop, Montpellier, France, July 28, 2012, pp 6–10

Biran O, Cotton C (2017) Explanation and justification in machine learning: A survey. In: Proceedings of the workshop on explainable AI (XAI), IJCAI, pp 8–13

Bonatti PA, Olmedilla D, Peer J (2006) Advanced policy explanations on the web. In: Proceeding of the 2006 conference on ECAI 2006, Amsterdam, The Netherlands, IOS Press, pp 200–204

Bouali F (1996) Diagnostic, validation et réparation de bases de connaissances : le système KBDR, Université PARIS-XI, centre d'Orsay, PhD thesis

Bouaud J, Séroussi B, Brizon A (2008) Impact de la réactualisation de recommandations de pratiques cliniques sur l'évolution d'une base de connaissances. In: Actes des 19es Journées Francophones d'Ingénierie des Connaissances (IC 2008) 19es Journées Francophones d'Ingénierie des Connaissances (IC 2008), Nancy, France, pp 1–12

Brank J, Grobelnik M, Mladenic D (2005) A survey of ontology evaluation techniques. In: Proceedings of the conference on data mining and data warehouses (SiKDD)

Brewster C, Alani H, Dasmahapatra S, Wilks Y (2004) Data driven ontology evaluation. In: Proceedings of the fourth international conference on language resources and evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal

Brinton C (2017) A framework for explanation of machine learning decisions. In: Proceedings of the workshop on explainable AI (XAI), IJCAI, pp 14–18

Cabrio E, Villata S, Gandon F (2013) A support framework for argumentative discussions management in the web. In: The semantic web: semantics and big data, 10th international conference, ESWC 2013, Montpellier, France, May 26–30, 2013. Proceedings, pp 412–426

Cawsey A (1990) Generating explanatory discourse. In: Current research in natural language generation. Academic Press Professional Inc., San Diego, CA, USA, pp 75–101

Charnay L (1999) Dialogue et explication dans les Systèmes à base de connaissances : ADex, un modèle informatique pour l'énonciation, PhD thesis, Université PARIS-XI, centre d'Orsay, notes et documents LIMSI num, pp 99–19

Chatalic P, Nguyen GH, Rousset M-C (2006) Reasoning with inconsistencies in propositional peer-to-peer inference systems. In: ECAI, pp 352–356

Chen L, Pu P (2012) Critiquing-based recommenders: survey and emerging trends. User Model User-Adapt Interact 22(1–2):125–150

Clancey WJ (1986) From guidon to neomycin and heracles in twenty short lessons. AI Mag 7(3):40–60

Corby O, Dieng-Kuntz R, Faron-Zucker C, Gandon F (2006) Searching the semantic web: approximate query processing based on ontologies. IEEE Intell Syst 21(1):20–27

David J (1995) Les systèmes experts de seconde génération. TSI 14(4):435–471

Dessalles J-L (2008) La pertinence et ses origines cognitives - Nouvelles théories. Hermès-Science, Paris, France

Devillers L, Bonneau-Maynard H, Rosset S, Paroubek P, McTait K, Mostefa D, Choukri K, Charnay L, Bousquet C, Vigouroux N, Bchet, F., Romary, L, Antoine, J-Y, Villaneau, J., Vergnes, M., and Goulian, J (2004) The French MEDIA/EVALDA Project: the Evaluation of the Understanding Capability of Spoken Language Dialogue Systems. In: LREC, Lisbon

Dibie-Barthélemy J, Haemmerlé O, Salvat E (2006) A semantic validation of conceptual graphs. Knowl Based Syst 19(7):26–38

Djedidi R, Aufaure M (2010) ONTO-EVO^{a}L an ontology evolution approach guided by pattern modeling and quality evaluation. In: Foundations of information and knowledge systems, 6th international symposium, FoIKS 2010, Sofia, Bulgaria, February 15–19, 2010, Proceedings, pp 286–305

Djedidi R, Aufaure M-A (2009) Patrons de gestion de changements OWL. In: Ingénierie des connaissances. Hammamet, Tunisia, pp 145–156

Duque-Ramos A, Fernández-Breis JT, Iniesta-Moreno M, Dumontier M, Aranguren ME, Schulz S, Aussenac-Gilles N, Stevens R (2013) Evaluation of the oquare framework for ontology quality. Expert Syst Appl 40(7):2696–2703

Euzenat J, Shvaiko P (2007) Ontology matching. Springer, Berlin

Farreny H (1985) Les systèmes experts. Cepadues

Feigenbaum E, Buchanan B, Lederberg J (1971) On generality and problem solving: a case study using the dendral program. Mach Intell 6:165–190

Gangemi A, Catenacci C, Ciaramita M, Lehmann J (2006) Modelling ontology evaluation and validation. In: The semantic web: research and applications, 3rd European semantic web conference, ESWC 2006, Budva, Montenegro, June 11–14, 2006, Proceedings, pp 140–154

Gangemi A, Presutti V (2009) Ontology design patterns. In: Handbook on ontologies pp 221–243

Ginsberg A (1988) Knowledge-base reduction : a new approach to checking knowledge bases for inconsistency and redundancy. In: National conference of american association of artificial intelligence AAAI'88, pp 585–589

Gómez-Pérez A (2004) Ontology evaluation. In: Handbook on ontologies, pp 251–274

Guarino N, Welty CA (2004) An overview of ontoclean. In: Handbook on ontologies, pp 151–172

Haouche C, Charlet J (1996) KBS validation : a knowledge acquisition perspective. In: European conference of artificial intelligence ECAI'96, pp 433–437

Haynes SR, Cohen MA, Ritter FE (2009) Designs for explaining intelligent agents. Int J Hum Comput Stud 67(1):90–110

Hors P, Rousset M (1996) Modeling and verifying complex objects : a declarative approach based on description logics. In: European conference of artificial intelligence ECAI'96, pp 328–332

Hovy EH (1988) Planning coherent multisentential text. In: Proceedings of the 26th annual meeting of the association for computational linguistics, Association for Computational Linguistics, Buffalo, New York, USA, pp 163–169

Institut PM (2013) Project Management Institut

Kassel G (1987) The use of deep knowledge to improve explanation capabilities of rule-based expert systems. In: Expertensysteme, pp 315–326

Kontokostas D, Westphal P, Auer S, Hellmann S, Lehmann J, Cornelissen R, Zaveri A (2014) Test-driven evaluation of linked data quality. In: 23rd international world wide web conference, www '14, Seoul, Republic of Korea, April 7-11, 2014, pp 747–758

Lane HC, Core MG, van Lent M, Solomon S, Gomboc D (2005). Explainable artificial intelligence for training and tutoring. In: Proceeding of the 2005 conference on artificial intelligence in education, pp 762–764, IOS Press, Amsterdam, The Netherlands,

Lee J, Lai LF, Yang S (2002) A high-level petri nets-based approach to verifying task structures. IEEE Trans. Knowl Data Eng 14(2):316–335

Loiseau S (1998) Validation, mise au point et interaction: quelques solutions pour assister le concepteur de bases de connaissances et l'utilisateur de systèmes informatiques. In: Habilitation à diriger des recherches

Lopez V, Unger C, Cimiano P, Motta E (2013) Evaluating question answering over linked data. J Web Semant 21:3–13

Baker M, Dessalles J-L, Joab M, Raccah P-Y, Safar B, Schlienger D (1995) La génération d'explications négociées dans un système à base de connaissances. In: Actes des 5èmes journées nationales du PRC-GDR intelligence artificielle, pp 297–316

McKeown KR (1985) Discourse strategies for generating natural-language text. Artif Intell 27:1–41

Minsky M (1975) A framework for representing knowledge. In: The psychology of computer vision, pp 211–285

MoDeVA (2009) Workshop model design and validation of models

Moreau L, Groth PT, Cheney J, Lebo T, Miles S (2015) The rationale of PROV. J Web Sem 35:235–257

Mugnier M, Chein M (1996) Représenter des connaissances et raisonner avec des graphes. Revue d'Intelligence Artificielle 7(1):7–56

Neches R, Swartout WR, Moore J (1985) Explainable (and maintainable) expert systems. In: IJCAI'85: proceedings of the 9th international joint conference on artificial intelligence. Morgan Kaufmann Publishers Inc, San Francisco, pp 382–389

Nguyen GH, Chatalic P, Rousset M-C (2008) A probabilistic trust model for semantic peer-to-peer systems. In: ECAI, pp 881–882

Nguyen TA, Perkins WA., Laffrey TJ, Pecora D (1985) Checking an expert system knowledge base for consistency and completness. In: International join conference of artificial intelligence, IJCAI'85, vol 1, pp 375–378

Nkambou R, Bourdeau J, Mizoguchi R (2010) Advances in intelligent tutoring systems. Springer. Berlin
Obrst L, Ceusters W, Mani I, Ray S, Smith B (2007) The evaluation of ontologies. In: Baker CJO, Cheung K-H (eds) Semantic web: revolutionizing knowledge discovery in the life sciences. Springer, Berlin, pp 139–158
Paris C, Cecile L, Swartout WR, Mann WC (eds) (1990) Natural language generation in artificial intelligence and computational linguistics. Kluwer Academic Publishers, Norwell
Paris CL, McKeown KR (1987) Discourse strategies for describing complex physical objects. Springer, Netherlands, pp 97–115
Paris CL, Swartout WR, Mann WC (eds) (1991) Natural language generation in artificial intelligence and computational linguistics. Kluwer, Boston
Pipard (1987) INDE: un système de détection d'inconsistances et d'incomplétudes dans les bases de connaissances. PhD thesis, Université PARIS-XI, centre d'Orsay
Poveda-Villalón M, Gómez-Pérez A, Suárez-Figueroa MC (2014) Oops! (ontology pitfall scanner!): an on-line tool for ontology evaluation. Int J Sem Web Inf Syst 10(2):7–34
Presutti V, Blomqvist E, Daga E, Gangemi A (2012) Pattern-based ontology design. In: Ontology engineering in a networked World, pp 35–64
Rahwan I, Zablith F, Reed C (2007) Laying the foundations for a world wide argument web. Artif Intell 171(10–15):897–921
Rao, J. and Su, X. (2004). A survey of automated web service composition methods. In: Semantic web services and web process composition, first international workshop, SWSWPC 2004, San Diego, CA, USA, July 6, 2004, Revised selected papers, pp 43–54
Reiter R (1987) A theory of diagnosis from first principles. Artif Intell 32:57–95
Reiter R, de Kleer J (1987) Foundations of assumption-based truth maintenance system. In: AAAI, pp 183–188
Roth-Berghofer T, Leake DB, Cassens J (eds) (2012) Explanation-aware computing, papers from the 2012 ECAI Workshop, Montpellier, France, July 28, 2012
Roth-Berghofer T, Schulz S (eds) (2005) Explanation-aware computing, papers from the 2005 AAAI fall symposium, November 4–6, 2005, Arlington, Virginia. AAAI technical report, vol FS-05-04. AAAI Press
Roth-Berghofer T, Schulz S, Bahls D, Leake DB (eds) (2007) Explanation-aware computing, papers from the 2007 AAAI workshop, Vancouver, British Columbia, Canada, July 22–23, 2007. AAAI Technical Report, vol WS-07-06. AAAI Press
Roth-Berghofer T, Schulz S, Leake DB, Bahls D (eds) (2008) Explanation-aware computing, papers from the 2008 ECAI Workshop, Patras, Greece, July 21–22, 2008. University of Patras
Roth-Berghofer T, Tintarev N, Leake DB (eds) (2009) Explanation-aware computing, papers from the 2009 IJCAI workshop, Pasadena, California , USA, July 11–12, 2009
Roth-Berghofer T, Tintarev N, Leake DB (eds) (2011) Explanation-aware Computing, Papers from the 2011 IJCAI Workshop, Barcelona, Spain, July 16–17, 2011
Roth-Berghofer T, Tintarev N, Leake DB, Bahls D (eds) (2010) Explanation-aware Computing, Papers from the 2010 ECAI Workshop, Lisbon, Portugal, August 16, 2010. University of Lisbon, Portugal
Rousset M (1988) On the consistency of knowledge bases: the COVADIS system. In: European conference of artificial intelligence ECAI'88, pp 79–84
Rousset M, Levy A (1996) Verification of knowledge bases based on containment checking. In: National conference of American association of artificial intelligence AAAI'96, pp 585–591
Rousset M, Levy A (1998) Verification of knowledge bases based on containment checking. Artif Intell 101(1–2):227–250
Safar B (1987) Le problème des explications négatives dans les Systèmes Experts: Le système POURQUOI-PAS? PhD thesis, Université PARIS-XI, centre d'Orsay
Schneider J, Groza T, Passant A (2013) A review of argumentation for the social semantic web. Sem Web 4(2):159–218

Schober D, Tudose I, Svatek V, Boeker M (2012) Ontocheck: verifying ontology naming conventions and metadata completeness in protégé 4. J Biomed Sem 3(2):1–10
Shanks G, Tansley E, Weber R (2003) Using ontology to validate conceptual models. Commun ACM 46(10):85–89
Sheng QZ, Qiao X, Vasilakos AV, Szabo C, Bourne S, Xu X (2014) Web services composition: a decade's overview. Inf Sci 280:218–238
Shortliffe EH (1976) Computer-based medical consultations: MYCIN. Kluwer Academic Publishers, Elsevier, New York
Sombe L (1988) Raisonnements sur des informations incomplétes en intelligence artificielle. Revue d intelligence artificielle 2(3–4):9–210
Sowa J (1984) Conceptual structures: information processing in mind and machine. Addison Wesley Publishing Company, Boston
Staab S, Studer R (2009) Handbook on ontologies, 2nd edn. Springer, Berlin
Steels L (1985) Second generation expert systems. Future Gener Comput Syst 1(4):213–221
Suárez-Figueroa MC, Gómez-Pérez A, Fernández-López M (2015) The neon methodology framework: a scenario-based methodology for ontology development. Appl Ontol 10(2):107–145
Sure Y, Staab S, Studer R (2009) Ontology engineering methodology. In: Staab S, Studer R (eds) Handbook on ontologies, 2nd edn. Springer, Berlin, pp 135–152
Swartout W, Paris C, Moore J (1991) Explanations in knowledge systems: design for explainable expert systems. IEEE Expert: Intell Syst Appl 6(3):58–64
Swartout WR (1983) Xplain: a system for creating and explaining expert consulting programs. Artif Intell 21:285–325
Thomopoulos R, Croitoru M, Tamani N (2015) Decision support for agri-food chains: a reverse engineering argumentation-based approach. Ecol Inf 26(2):182–191
Unger C, Forascu C, Lopez V, Ngomo AN, Cabrio E, Cimiano P, Walter S (2015) Question answering over linked data (QALD-5). In: Working notes of CLEF 2015 - conference and labs of the evaluation forum, Toulouse, France, September 8–11, 2015
Vrandecic D (2009) Ontology evaluation. In: Staab S, Studer R (eds) Handbook on ontologies, 2nd edn. Springer, Berlin, pp 293–313
Walton D (2007) Dialogical models of explanation. In: Explanation-aware computing, papers from the 2007 AAAI workshop, Vancouver, British Columbia, Canada, July 22–23, 2007, pp 1–9
Weiner JL (1980) Blah, a system which explains its reasoning. Artif Intell 15(1–2):19–48
Wick MR, Thompson WB (1989) Reconstructive explanation: explanation as complex problem solving. In: IJCAI'89: proceedings of the 11th international joint conference on artificial intelligence. Morgan Kaufmann Publishers Inc, San Francisco, pp 135–140
Yurchyshyna A, Faron-Zucker C, Mirbel I, Sall B, Thanh NL, Zarli A (2008) Une approche ontologique pour formaliser la connaissance experte dans le modèle du contrôle de conformité en construction. Actes d'IC 49–60
Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S (2016) Quality assessment for linked data: a survey. Sem Web 7(1):63–93

Knowledge Engineering

Nathalie Aussenac-Gilles, Jean Charlet and Chantal Reynaud

Abstract Knowledge engineering refers to all technical, scientific and social aspects involved in designing, maintaining and using knowledge-based systems. Research in this domain requires to develop studies on the nature of the knowledge and its representation, either the users' knowledge or the knowledge-based system's knowledge. It also requires the analysis of what type of knowledge sources is considered, what human-machine interaction is envisaged and more generally the specific end use. To that end, knowledge engineering needs to integrate innovation originating from artificial intelligence, knowledge representation, software engineering as well as modelling. This integration enables both users and software systems to manage and use the knowledge for inference reasoning. Other advances are fuelling new methods, software tools and interfaces to support knowledge modelling that are enabled by conceptual or formal knowledge representation languages. This chapter provides an overview of the main issues and major results that are considered as milestones in the domain, with a focus on recent advances marked by the raise of the semantic web, of ontologies and the social web.

1 Introduction

Knowledge engineering (KE) became a research domain in the early 1980s, its research object being designing, maintaining and using knowledge-based systems

N. Aussenac-Gilles (✉)
IRIT-CNRS, Université de Toulouse, Toulouse, France
e-mail: Nathalie.Aussenac-Gilles@irit.fr

J. Charlet
Sorbonne Université, INSERM, Université Paris 13, LIMICS, 75006 Paris, France
e-mail: Jean.Charlet@upmc.fr

J. Charlet
Assistance Publique-Hôpitaux de Paris, DRCI, Paris, France

C. Reynaud
LRI, Université Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France
e-mail: Chantal.Reynaud@lri.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7_23

(KBS). Many of the early expert systems were developed using traditional software engineering methods combined with rapid prototyping. In this context, building conceptual models in the early stages of the process became a major and critical issue. The further population of these models with the appropriate knowledge presented also substantial challenges. The so-called knowledge acquisition bottleneck[1] became the subject of a large amount of research work, Ph.D. theses and international projects, either with a cognitive and methodological perspective (Aussenac 1989) or targeting the definition of new knowledge representations (Cordier and Reynaud 1991; Charlet 1991). In the late 1990s, the perspective broadened and gave birth to KE as a cross-disciplinary research domain. Mainly located in the field of Artificial Intelligence (AI), KE refers to all technical, scientific and social aspects involved in designing, maintaining and using KBS. KE defines the concepts, methods, techniques and tools to support knowledge acquisition, modelling and formalisation in organisations with the aim of structuring the knowledge and making it operational.

KE is expected to address knowledge modelling and sharing issues when designing any KBS that supports human activities and problem solving. Such knowledge intensive applications include knowledge management (KM) systems, Information Retrieval (IR) tools, both semantic or not, document or knowledge browsing, Information Extraction (IE), decision making or problem solving to name but a few. When the Semantic Web (to which the chapter "Semantic Web" of Volume 3 of this book is dedicated) emerged as a promising perspective to turn web data into knowledge and to define more powerful web services, research in KE started waving close relations with this domain. Indeed, the Semantic Web overlaps KE in various ways, both domains use the same languages, standards and tools like ontologies, knowledge representation languages and inference engines.

In the rest of this chapter, we propose a chronological and historical presentation of the major paradigms that marked milestones in KE during the last 25 years in Sect. 2. Then in Sect. 3, we detail the main research issues that KE is dealing with. Section 4 offers a synthetic view of the remaining methodological and representation challenges before we conclude in Sect. 5.

2 Knowledge Modelling

2.1 The Notion of Conceptual Model

Around the 1990s, KE methods proposed to design KBS starting with a knowledge modelling stage that aimed to collect and describe the system knowledge in

[1] Knowledge acquisition refers to the process of gathering expert knowledge (called "knowledge mining" at that time) and representing it in the form of rules and facts in the hope that the KBS behaves like the expert would in a similar situation. The difficulty to precisely collect or capture this knowledge, which is implicit and hard to elicit in many ways, reduces the amount and quality of knowledge actually represented, as the term "bottleneck" illustrates.

an operational form, regardless of the implementation. Knowledge representation in the model was both abstract and with an applicative purpose. It was expected to account for the multiple necessary knowledge features and types to meet the system requirements. Practically, this representation formed the so-called *conceptual model*. A conceptual model should fit the kind of knowledge to be described and would then be formalised using the appropriate formalisms required by the KBS (i.e. inference rules in many applications of the 1990s). Then, conceptual models became key components in knowledge engineering and they significantly evolved over the years to cover a large variety of models depending on the needs they should satisfy, thus being adapted to new approaches and to every recent research work in the field.

The way in which knowledge is described and represented impacts the implementation of the targeted KBS, and even more, the ability to understand or explain its behaviour. Knowledge acquisition and engineering have long referred to A. Newell's notion of *Knowledge Level* (1982). Newell was one of the first to establish a clear separation between the knowledge to be used in a system to produce a behaviour and its formal "in-use" representation in the system implementation. In other words, Newell stressed the necessity to describe the system knowledge at a level that would be independent from the symbols and structure of a programming language, level that he called the *Knowledge Level*. At this level, the system is considered as a rational agent that will use its knowledge to achieve some goals. Such system behaves in a rational way because, thanks to its knowledge, he intends to select the best sequence of actions leading to one of its goals as directly as possible. Newell's Knowledge Level not only prompted researchers to define conceptual models, but it also influenced the structuring of these models in several layers corresponding to various types of knowledge required to guarantee the system behaviour. In conceptual models, *domain knowledge*, that gathers entities or predicates and rules, is distinct from *problem solving knowledge* that consists in actions and goals modelled using *methods* and *tasks*.

2.2 Problem Solving Models

Problem solving models describe in an abstract way, using tasks and methods, the reasoning process that the KBS must carry out. A task defines one or several goals and sub-goals to be achieved by the system, and a method describes one of the ways the task goals can be achieved. A task description also specifies the input and output knowledge, constraints and resources required to perform the task. To describe the way the system should behave to solve a problem, a hierarchy of tasks can be defined, a general task being decomposed into several more specific tasks that specify the sub-goals required to achieve the goal of the main task. Methods make explicit how a goal can be reached thanks to an ordered sequence of operations. Methods that decompose a task into sub-tasks are distinguished from methods that implement a basic procedure to directly reach a particular goal. The distinction between tasks and methods progressively emerged from research works after B. Chandrasekaran

proposed the notion of Generic Task (1983) and L. Steels proposed a componential modelling framework that included three types of components: tasks; methods and domain data models (1990). This distinction has been adopted to account for the reasoning process in many studies (Klinker et al. 1991; Puerta et al. 1992; Schreiber et al. 1994; Tu et al. 1995) because it provides a separate description of the targeted goal and the way to achieve it. Thus, several methods can be defined for one single task, making it easier to explicitly represent alternative ways to reach the same goal. This kind of model is similar to results established in task planning (Camilleri et al. 2008; Hendler et al. 1990) where planning systems implement problem solving models thanks to operational methods and tasks, as it is suggested in the CommonKADS methodology (Schreiber et al. 1999).

2.3 From Conceptual Models to Ontologies

Once solutions had been found to design explicit problem-solving models, building the full conceptual model of an application consisted in reusing and adapting problem-solving components together with an abstract representation of domain data and concepts. Then an analysis of the domain knowledge was needed to establish a proper connection between each piece of the domain knowledge and the roles it played in problem solving (Reynaud et al. 1997). Domain knowledge models include two parts. The *domain ontology* forms the core part; it gathers concepts, i.e. class-sets of domain entities in a class/sub-class hierarchy, and relations between these classes, to which may be associated properties like constraints or rules. The second part extends this core with instances or entities belonging to the concepts classes, and relations between these entities. Thus an ontology defines a logical vocabulary to express domain facts and knowledge, in a formal way so that a system can use it for reasoning. Some concepts, called *primitive concepts*, are defined thanks to their situation in the concept hierarchy and thanks to properties that form necessary conditions for an entity to belong to this class. Other concepts, called *defined concepts*, are defined as classes equivalent to necessary and sufficient conditions that refer to properties and primitive concepts. The word *ontology* used to refer to a sub-field of philosophy. It has been first used in computer science, and particularly in AI, after the *Knowledge Sharing Effort* ARPA project (Neches et al. 1991) introduced it to refer to a structure describing the domain knowledge in a KBS. A little later, Gruber (1993) was the first to propose a definition of ontology in the field of KE. A more recent definition, proposed in Studer et al. (1998), is currently the acknowledged one:

> *An ontology is a formal, explicit specification of a shared conceptualisation.*
>
> *Conceptualisation refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable.*
>
> *Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group.*

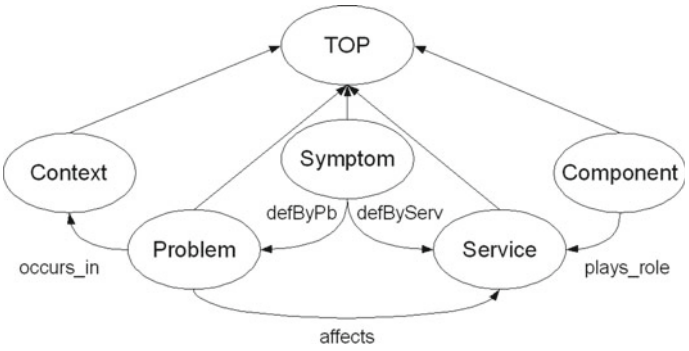


Fig. 1 High level concepts of an ontology used in the domain of electronic fault diagnosis

To sum up, ontologies meet complementary and symmetric requirements: (a) as specifications, they define a *formal semantics* so that software tools may process them; (b) as knowledge models, they reflect a – partial – point of view on a knowledge domain, that designers try to build as consensual as possible, and they provide semantic bridges that connect machine processable representations with their actual meanings for humans – supporting what Rastier calls *interpretative semantics* (2009).

The fact that an ontology be formal is both a strength because it enables to produce inferences (e.g. entity classification) and a limitation, using a formal language for its representation making it more difficulty to build. Figure 1 presents the main high level concepts of an ontology designed for an IR system in the domain of electronic fault diagnosis for cars. The `symptom` concept is defined by the identification of a car `component`, that provides a `service` to the vehicle user, that has been affected by a `problem` in a particular `context`. In the formal representation of this model, cardinality constraints on the `defByPb` and `defByServ` relations contribute to express that an instance of symptom cannot be identified unless a service and a problem have been identified too.

According to how the ontology will be used, it needs to be more or less rich in defined concepts and relations. For instance, if the ontology will be used in a standard information retrieval system, its role will be to structure domain concepts in a hierarchy and to provide labels (terms) for these concepts. This kind of ontology is called a *light-weight ontology*: it contains a concept hierarchy (or taxonomy) and very few defined concepts. When concept labels are represented with a specific formal class and properties, either called (formal) term or lexical entry, this kind of ontology is called *Lexical Ontology*.[2] If the ontology is to be used to produce inferences on domain knowledge, it will generally be larger and it will contain more relations, more axioms involved in the definition of defined concepts or any concept required for reasoning. This second kind of ontology is called a *heavy-weight ontology*.

[2]Whereas the KE English-speaking community uses "lexical ontology", many French research groups refer to Termino-Ontological Resource (TOR) (Reymonet et al. 2007) for very similar knowledge structures.

Due to their genericity and potentially high reusability, ontologies were expected to be easy to design. Several research lines have tried to characterise which parts of an ontology could be generic, and consequently reusable, on the one hand, and which techniques and methods could support the design of the non-generic parts. This distinction led to define the following typology of ontologies, which may also correspond to knowledge levels in a single ontology:

- An *upper level ontology* or *top*-ontology is considered the highest level. It structures knowledge with very general and abstract categories that are supposed to be universal and that are the fruit of philosophical studies on the nature of the main knowledge categories when formally representing human thinking in any domain. The major reference studies about top levels in ontologies are Sowa's *top-level categories*,[3] SUMO,[4] or DOLCE[5] to name a few of them. As concluded by the SUO[6] working group and the joint communiqué from the Upper Ontology Summit,[7] trying to define a unique norm for high level categories is pointless as long as various philosophical schools or trends propose distinct ways to categorise the world entities. Top level ontologies are the anchor point of more specific levels (core ontologies and domain knowledge), and they are generic enough to be shared.
- A *core ontology* or *upper domain ontology* provides a domain description that defines the main concepts of a particular domain, together with properties and axioms applying on these concepts. For instance, a core ontology of medicine would contain concepts such as *diagnosis, sign, anatomic structure* and relations like *localisation* linking a pathology to the affected anatomic structure (cf. GFO-Bio[8]); in Law, the LKIF-Core[9] ontology offers notions like *norm, legal action* and *statutory role*.
- A *domain ontology* describes the domain concepts practically handled by professionals and experts in everyday activities. It is the most specific kind of a knowledge model, and it becomes a knowledge base when instances of domain specific concepts are represented. Nevertheless, there may be no clear frontier between a *core*-ontology and an ontology of the same domain that includes the core one when both of them are designed within the same process. The distinction is more obvious when the domain ontology reuses and specialises an existing core ontology. Domain ontologies or the domain level of ontologies can be designed thanks to text-based approaches and reusing domain thesaurus or terminologies (cf. Sect. 4.1).

[3] http://www.jfsowa.com/ontology/toplevel.htm.

[4] http://www.ontologyportal.org/.

[5] http://www.loa-cnr.it/DOLCE.html.

[6] http://suo.ieee.org/.

[7] http://ontolog.cim3.net/cgi-bin/wiki.pl?UpperOntologySummit/UosJointCommunique.

[8] http://www.onto-med.de/ontologies/gfo-bio/index.jsp.

[9] http://www.estrellaproject.org/lkif-core/.

3 Issues and Major Results

If we consider the KE evolution over the last 30 years, changes have been driven by the diversification of what could be considered as a knowledge source for "intelligent" or AI information systems. This wealth in knowledge sources came together with changes in computers that impacted any software system: the amazing increase in storage capacities and always higher computing performance of computers. Knowledge source diversification offered the advantage to benefit from complementary sources together with available techniques to analyse them. In the following we first outline the various knowledge sources that KE has successively focused on over the years, as well as the research issues raised by the passage from these sources to models. By model, we mean here the different types of knowledge models presented in Sect. 2 used to represent either the knowledge in a KBS (*conceptual models*), some problem-solving process (*problem-solving models*) or domain specific knowledge (*domain models*). Then we show the research paradigms that deal with these issues, as well as the variety of modelling methods and techniques produced in KE to overcome them. We end with the presentation of major results about model reuse and with the connection of this research with the one on knowledge representation.

3.1 Knowledge Sources

Historically, *knowledge* for KBS first referred to human expertise, for which the knowledge base of *expert systems* should account according to a human-inspired paradigm. Knowledge was thus both technical and specialised. It gathered high-level skills and know-how that generally never had been verbalised before, and that were hard to explicit. The expected role of expert systems was to capitalise and make this expertise explicit so that it could be sustained and transferred to the KBS, or to humans via the KBS. Knowledge was then represented with inference rules.[10]

In a second period, expert systems evolved and became *Knowledge-Based systems* because their role was no longer to replace the expert but rather to provide an intelligent help to the end-user. Efficiency was privileged against the accuracy towards human reasoning. Then reference knowledge became shared knowledge, that KBS used for reasoning according to their own problem solving engines.

Today, many applications (i.e. spelling checkers, decision support systems, billing systems, but also chest players or search engines) include some *model-based modules*. Their goal is to perform some of the system tasks either in an autonomous way or in a cooperative way together with other modules or in cooperation with the user, adapting to the use context and to users' profiles. The knowledge required for these support tasks to solve problems or to perform activities includes technical,

[10]For a historical outline on knowledge-based system, one can read Aussenac (1989), Stefik (1995), Aussenac-Gilles et al. (1996), or Charlet et al. (2000).

consensual and shared knowledge, that is modelled as rules or action maps, and as structured and goal-oriented domain models.

The historical evolution of knowledge-based information systems highlights various types of knowledge that were considered over the years: individual expert knowledge, in-use knowledge related to practice, activities and individual usage; knowledge about organisations, consensual and shared knowledge of an application field, common sense knowledge, knowledge related to knowledge integration or distributed knowledge over the Web. It is to capture these various kinds of knowledge that new knowledge sources have been taken into account. Thus, documents have played an increasing role as more digital documents were available. Since the early works on knowledge acquisition for expert systems, KE relies on documents, in particular textual documents, as they convey meaning and may contribute to reveal some knowledge. Documents are exploited for the language and information they contain, which is complementary or an alternative to interviews of domain experts or specialists. Data can also become knowledge sources thanks to knowledge or information extraction processes from data or data mining. Last, components of existing knowledge models can be reused when they convey consensual and shared knowledge. These components can either be *problem solving models*, that can be reused across various domains, like the library of problem solving methods in CommonKADS (this library is one of the major results of the KADS and later CommonKADS[11] European projects Schreiber et al. 1999), or domain models, ontologies, semantic resources like lexical data-bases or thesauri. Ontologies represent domain concept definitions in a formal structure. A lexical data-bases like WordNet[12] registers, classifies and organises, according to semantic and lexical criteria, most of the vocabulary of the English language. Thesauri collect normalised domain vocabularies as structured sets of terms.

3.2 From Knowledge Sources to Models: Research Issues

One of the core and typical issues in KE is to provide or develop tools, techniques and methods that support the transition from the knowledge sources listed in Sect. 3.1 to the models presented in Sect. 2. These techniques not only rely on software systems but also on analysis frameworks or observation grids borrowed to other disciplines. Research in KE actually follows an engineering paradigm in the sense that it requires innovation to design new tools, languages and methods or to select and adapt existing ones. It requires as much innovation to organise them in an appropriate way within methodological guidelines and integrated or collaborative platforms. Expected innovations concern the nature and development of these tools as well as the definition of their use conditions, their synergy and interactions so that they could manage particular knowledge types at each stage of the development process of an application.

[11] http://www.commonkads.uva.nl/.

[12] http://wordnet.princeton.edu/wordnet/.

For the last twenty years, methodological research in KE raised cross-functional issues that have been reformulated and renewed when new knowledge sources were addressed, new types of models were designed or new use-cases and problems had to be solved using these models.

3.2.1 How to Design a Model?

Two complementary methodological streams first defined diverging stages and techniques (Aussenac-Gilles et al. 1992). *Bottom-up methods* privilege data analysis, first driven by the identified users' needs and later guided by the model structure and the components to be filled. Bottom-up approaches focus on tools that support data collection and mining, knowledge identification and extraction, and later on tools that produce abstract representations of knowledge features (classification, structuring and identification of methods and problem solving models). In contrast, the alternative process follows a *top-down approach* that privileges the *reuse* and adaptation of existing knowledge components. Then knowledge gathering starts with the selection of appropriate components, that further guides the extraction of new knowledge and the model instantiation process. A unified view considers that modelling follows a cyclic process where bottom-up and top-down stages alternate. The process moves from stages dedicated to knowledge collection or reuse towards knowledge representation stages using more and more formal languages. Most methods and tools presented in Sect. 3.3 combine both processes, whereas we focus on results about model reuse in Sect. 3.4.

3.2.2 How to Benefit from Complementary Knowledge Sources?

Diversifying knowledge sources and knowledge types is one of the solutions to get more precise and richer models, or to automatically design a part of them. As a consequence, KE methods start with the identification of appropriate knowledge sources. They suggest also a set of relevant tools and techniques that explore and efficiently process these sources. Most of all, they propose methodological guidelines to articulate the use of these tools in a coordinated way that ensures a complementary exploitation of their results to design an appropriate model. Results in Sect. 3.3 illustrate this process.

3.2.3 What Are Models Made of? What is the Optimal Formal Level?

Each model combines various types of knowledge. In a similar way, each KE method questions and makes suggestions on the nature of the models to be designed, on the way to structure them and to collect the appropriate knowledge that feel them as well as on the representation formalism to select, which can be more or less formal as discussed in Sect. 3.5.

3.2.4 How Does Model Engineering Take into Account the Target Use of a Model?

Several research studies have shown that conceptual models were all the more relevant than they were dedicated to a specific range of systems. KE does not restrict its scope to design models; it is highly concerned by their actual use because it is one of the ways to validate the engineering process, and because it is this specific use that determines the model content, its structure and, as a side effect, the way the model is designed. In short, the targeted use of a model has a strong impact on methodological options and on the selection of a knowledge representation in the model (Bourigault et al. 2004).

3.2.5 How to Promote Model Reuse?

The reuse of structured knowledge fragments is often the best option to reduce the cost of knowledge modelling. However, reuse is not possible unless the principles that guided the model design are available, unless models can be compared and combined, and unless the selection of some of their components and their combination are technically feasible and sound. These very same questions also arise in research work about ontology or KB alignment, reuse and composition to build new knowledge bases.

3.2.6 How to Ensure Model Evolution in Relation with the Use Context?

The knowledge models used in KBS are involved in a life cycle that includes their evolution. This parameter became increasingly significant as a consequence of the evolution of the knowledge sources, of domain knowledge and users' needs. Since the early 2000s, ontology evolution is one of the major challenges to be solved to promote their actual use. Various research studies define an evolution life-cycle, several means to identify and to manage changes while keeping the model consistent (Stojanovic 2004; Luong 2007).

3.3 Designing Models: Techniques, Methods and Tools

In order to make practical proposals in getting access to knowledge coming from people or documents deemed to provide indications, KE has its own solutions: techniques and tools that may be integrated into methodologies and frameworks. These solutions are largely inspired by close disciplines, depending on the considered source of knowledge, sequentially covering cognitive psychology, ergonomics, terminology and corpus linguistics since KE emerged as a discipline.

Designing models requires access to knowledge available through various sources. Access techniques depend on the nature of the sources, with potentially generation of new knowledge that had not been made explicit before. *Technique* makes reference here to operating modes requiring specific ways to choose or create knowledge production or use situations, then ways to discover/collect/extract or analyse data, and finally proposals to interpret, evaluate and structure the results of the analysis. We focus on the two knowledge sources that have been most widely used in this process: human expertise and textual documents.

3.3.1 Human Expertise as Knowledge Source

Regarding human expertise, research approaches have evolved from a *cognitivist* perspective, assuming a possible relation between mental and computer representations, to *constructivist* approaches, considering that models as artifacts that enable the system to behave as the human would, and then *situated cognition*, taking into account a contextual or collective dimension. In the first case, the task is to locate, make explicit and represent technical expertise. According to this view, which historically lead to design expert systems, one or several human experts possess the knowledge that has to be made explicit in order to design a system that produces the same reasoning. Cognitive psychology has provided guidelines on how to carry out interviews, on how to analyse them and gave the pros and cons of each form of interview in relation to the study of human cognitive phenomena (Darses and Montmollin 2006). These techniques have been adapted and then used to extract knowledge from experts, as in the works of Aussenac (1989), Shadbolt et al. (1999) or Dieng-Kuntz et al. (2005). We can distinguish the *direct* methods that consist in querying the expert to get him to speak in a more or less guided way and the *indirect* methods as *repertory grids* based on the interpretation of acquired elements as the expert performs tasks using his expertise.

This *cognitivist* perspective has been increasingly brought into question to better satisfy the situated aspect of the knowledge. As expertise is only accessible when applied in problem solving situations, KE has taken up task and activity analysis techniques from the area of ergonomics.

One main result was to lay the foundations of knowledge acquisition as a discipline focusing on knowledge *itself* prior to considering its formalisation and its use within a given system. Both adopting the *constructivist* view and taking into account existing methods in software engineering then led to new methodological proposals guiding the whole knowledge acquisition process. Several methods defined in important projects, mainly European projects, are presented in Sect. 3.3.3.

Knowledge in software aims at better guiding users. By the way, it impacts their working methods. So it raises the need to analyse their practices and the practices of their collaborators, to study their activities and their use of support tools, to consider their organisational context, which refers to ergonomics, sociological or management approaches. Results of such analyses were first returned in a static way, as models (task, interaction and organisation models for instance in CommonKADS) (Schreiber

et al. 1999). These models were made operational using task languages and methods such as LISA, Task (Jacob-Delouis and Krivine 1995) or CML (Schreiber et al. 1994). The notion of trace of activities has then been widely explored to take into account activities in a more in-depth way. Traces are integrated to provide users with a precise and context sensitive help based on the knowledge of their behaviour. Therefore, Laflaquiére et al. (2008) define the notion of trace for software use or documentation system activities in order to be able to discover, represent, store traces and then exploit and reuse them.

3.3.2 Textual Documents as Knowledge Sources

Regarding textual documents, whether technical, linked to an activity or to an application domain, two problems arise when exploiting them as knowledge sources: their selection and their analysis. Document analysis is mainly based on the natural language in the text. Some approaches also exploit the text structure identified on the paper or screen layout and electronically manageable thanks to tags or annotations (Virbel and Luc 2001). The latter is generally referred as structured or semi-structured documents (XML documents). We first describe the strengths of textual document analysis, then the techniques and the tools used for that.

Strengths of Textual Document Analysis

Textual documents are rich knowledge sources. Text analysis has always been a part of KE but the way to address it changed drastically after 1990. We do not try anymore to recover automatically the understanding of a text by an individual (Aussenac-Gilles et al. 1995). The increasing importance of textual analysis is a consequence of the progress achieved by natural language processing (NLP), which has delivered robust specialised software programs to process written language. NLP maturity has been synchronous with ontology deployment. Designing ontologies and using them to semantically annotate documents became two applications of the analysis of written natural language. A strong assumption behind automatic text processing is that text provide stable, consensual and shared knowledge of an application domain (Bourigault and Slodzian 1999; Condamines 2002). However, this is not always the case, and two key points influence the quality of the extracted data: first, the creation of a relevant corpus early on in the process, then a regular contribution of domain experts or experts in modelling for interpreting the results. Text analysis is used to design ontologies and similar resources such as thesauri, indexes, glossaries or terminological knowledge bases.

Techniques and Tools for Textual Analysis

The aim of textual analysis in KE is to discover, in an automatic or cooperative way, linguistic elements and their interpretation and to help designing parts of conceptual models.

Linguistic approaches are based on wordings in the text to identify knowledge rich contexts (Barriere and Agbago 2006). Domain notions are expected to be mentionned using nominal or verbal phrases with a strong coherence. According to the way they

are used, these phrases can be considered as terms denoting domain concepts or relationships between domain concepts. Language may also provide clues with a lower reliability, linking more diffuse knowledge elements. Then analysts have to rebuild reference links in order to come up with knowledge-based elements, axioms or rules. Results established by lexical semantics, terminology and corpus linguistics research are set prior to the implementation of this kind of approach (Condamines 2002; Constant et al. 2008).

Statistical approaches process a text as a whole and take advantage of redundancies, regularities, co-occurrences in order to discover idioms and terms, but also words or sets of words (clusters) with a similar behaviour or linguistic context. Several such techniques are described in the book *Foundations of Statistical Natural Language Processing* from Manning and Schütze (1999).

In both cases, preliminary text analysis, as cutting a text into sentences and into token words or grammatical parsing of words, is needed. A description of this research work is given in chapter "Artificial Intelligence and Natural Language" of Volume 3. The more sophisticated the pre-processing is (as complete syntactic analysis of sentences), the easier it is to automatically define precise interpretation rules. Unfortunately, software performing sophisticated analyses are often less robust, and they are available in fewer languages, English being often favoured. Furthermore, resources are sometimes needed (such as glossaries or semantic dictionaries) and few of them are available in some languages.

When the structure of the documents is available as a result of explicit markers, linguistic approaches can be combined with the exploitation of the structure in order to benefit of their complementary semantics (Kamel and Aussenac-Gilles 2009). The underlying idea is that structural cutting process of documents contributes to the semantic characterisation of their content.

Regarding the design of ontologies, text analysis serves two purposes (Maedche 2002; Cimiano et al. 2010): the identification of concepts with their properties and relationships, or *ontology learning* process; and the identification of concept instances and relations holding between them, the *ontology population* process. Similar tools can be used in both cases: text corpora have to be parsed in order to discover linguistic *knowledge-rich* elements (Meyer 2000), linguistic clues that can be interpreted as knowledge fragments.

Vocabulary modelling motivated the design of dedicated software tools that provide higher level results than standard NLP tools. For instance, results such as terms and clusters of synonym terms can then be integrated in a model. Examples of such tools are term extractors – Terminoweb (Barriere and Agbago 2006), Syntex-Upery (Bourigault 2002), TermExtractor (Drouin 2003) or TermRaider in the GATE[13] framework -; pattern-based relation extractors - Caméléon (Aussenac-Gilles and Jacques 2008), RelExt (Schutz and Buitelaar 2005) or SPRAT (Maynard et al. 2009) that implements three types of lexico-syntactic patterns (Hearst's patterns, patterns derived from Ontology design patterns and contextual patterns) in

[13] http://gate.ac.uk/.

GATE; pattern-based languages like Jape in GATE, Nooj,[14] Unitex[15]; named-entity extractors (Poibeau and Kosseim 2000) that contribute to search for instances or relations between instances (as with the KIM platform[16]). To sum up, designing models from texts has strongly benefited from NLP frameworks (GATE, Linguastream,[17] UIMA[18]) that support the development of adapted processing chains. Finally, specific processing chains, as Text2Onto (Cimiano and Völker 2005), and the version integrated by NeOn,[19] have allowed an assessment of the strengths and limitations of this approach by increasing automation and exploiting machine learning techniques. Current research works combine text analysis, reuse of ontological components and human interpretation. Cimiano et al. (2010) gives a reasonably full picture of these works.

3.3.3 Modelling Frameworks

Modelling frameworks provide access to knowledge sources, or to their traces, to knowledge extraction techniques and software tools, as well as to modelling techniques and languages. They suggest a methodology that defines a processing chain and guides the modelling task step by step. In the following Sub-section, we first present the most significant results about problem-solving modelling in the early 1990s. Then we focus on methods and frameworks for ontology design which have been developed in the last ten years.

Methods for Problem-Solving Modelling

Methodological guidelines have been established to better design large knowledge-based system projects. Their principles are similar to those in software engineering because of the importance assigned to modelling. In both cases, development cycles have to be managed and one or several models of the system to be designed must be built. The design of an application is considered as a model transformation process with conceptual models defined in Sect. 2.1. This requires a set of epistemological primitives that characterises at a high level (knowledge level) inference capabilities of the system to be designed. These primitives define generic knowledge representation structures that can be further instantiated.

In the early 1980s and 1990s the notion of conceptual model evolved with an emphasis on problem-solving models, new related languages, inference and tasks notions articulated. From a methodological viewpoint, the research showed that modelling primitives provide a grid for collecting and interpreting knowledge; they guide modelling. The utility of having elements coming from generic models and

[14] http://www.nooj4nlp.net/.

[15] http://www-igm.univ-mlv.fr/~unitex/.

[16] http://www.ontotext.com/kim/.

[17] http://linguastream.org/.

[18] http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.index.html.

[19] http://www.neon-toolkit.org/.

of being able to reuse them by instantiation on a particular application has then emerged, in particular from results on Generic Tasks from Chandrasekaran (1983). Later, the CommonKADS methodology showed the interest of adaptable and modular elements. All these principles are general as they apply irrespective of the task, the domain and the problem-solving method performed. Modelling techniques and reusable components are integrated in frameworks including as well expertise extraction techniques.

Following the work on Generic Task and role-limited methods (Marcus and McDermott 1989), and the proposals made by L. Steels in the componentional COMMET approach and in the KREST framework (1990), several works distinguished explicitly the notions of tasks and methods. This distinction has the advantage to describe separately the goal to be reached from the way to reach it and it allows for the explicit definition of several ways to reach a same goal by associating several problem-solving methods to a same task. These works have been taken into account by the European project KADS (Schreiber and Wielinga 1992), a pioneer in KE, which has resulted in the most accomplished methodology and framework CommonKADS (Schreiber et al. 1999).

CommonKADS allows for the construction of several models related to each other and required to specify a KBS with an organisational model reflecting in-use knowledge. The expertise model of the system is now recognised as very different from a cognitive model of a human expert. It is described according to three viewpoints: tasks, domain models, methods. Each problem-solving method can be parametrised and its adaptation is defined using a questionnaire guiding for the choice of one of the solution methods corresponding to each main task of the reasoning process of a specific application. Tasks describe what must be performed by the KBS. Domain models describe the knowledge required for reasoning. Methods describe how the knowledge is used to solve a task. A method can decompose a task into sub-tasks or solve one or several task(s). The methodology suggests an iterative construction of an application model according to the three different viewpoints. These perspectives are all necessary and complementary. The choice of a domain model depends on the selection of a problem-solving method as problem-solving methods define the role of the knowledge to be filled. Specifically, methods largely define the nature of the controlled sub-tasks. The aim of the methodology is thus to identify and model all the relations between methods, tasks and domain models.

Methods and Frameworks for Designing Ontologies

The design process of ontologies took advantage of these methodologies. It started when the reuse of domain models put forward the interest in high quality consensual models designed according ≪ good ≫ principles facilitating reuse and adaptation. The specific challenges encountered during the ontology design process are the followings:

1. Define the ontology content and ensure its quality;
2. Exploit efficiently all available knowledge sources using, for instance, text analysis or ontology reuse processes;

3. Facilitate the knowledge engineer design by providing specific tools; and
4. Define a methodological setting and the relevant approach to perform the various tasks.

Ontology engineering frameworks are uniform and coherent environments supporting the ontology design. They help achieve the different tasks by providing various tools and supporting a methodology that guarantees that all tasks are run one after the other.

Various methods can be used to design ontologies.[20] In this paper, we present three methodologies that are paying close attention to the quality of the ontology content: OntoClean, ARCHONTE and OntoSpec.

The OntoClean methodology has been designed by Guarino and Welty (2004). The first ideas were presented in a series of articles published in 2000, the OntoClean name appeared in 2002. Inspired by the notion of formal ontology and by principles of analytical philosophy, OntoClean made a significant contribution as the first formal methodology in ontology engineering. It proposes to analyse ontologies and to justify ontological choices using metaproperties of formal classes independent of all application domains. These metaproperties were originally four (i.e. identity, unity, rigidity and dependence).

The ARCHONTE (ARCHitecture for ONTological Elaborating) methodology, designed by Bachimont et al. (2002), is a bottom-up methodology to design ontologies from domain texts in three steps. First, relevant domain terms are selected and then semantically normalised as concepts by indicating the similarities and differences between each concept, its siblings and its father (principle of *differential semantic*). The second step consists in knowledge formalisation (*ontological commitment*). The aim is to design a differential ontology by adding properties or annotations, by defining domains and ranges of relationships. Finally, the third step consists in ontology operationalisation using knowledge representation languages. This process results in a *computational ontology*.

OntoSpec (Kassel 2002) is a semi-informal ontology specification methodology. It finds its origins in the definitions that are associated in natural language with conceptual entities which allow users to collaborate with knowledge engineers in order to design ontologies. In addition, this methodology proposes a framework including a typology of properties that can be used in the definition of concepts, relationships or rules, in order to paraphrase properties using natural language. The framework serves as a guide to model and facilitate the design of formal ontologies.

The main component of the frameworks used for designing ontologies is usually an ontology editor. Therefore, Protégé[21] is an editor extensively used to create or modify RDFS or OWL ontologies, and can be available as a web service (Web-Protégé) which is particularly appropriate for cooperative ontology design. Swoop[22] has been designed for lightweight ontologies, whereas Hozo[23]'s original-

[20] For a survey of the main existing methodologies, see Fernández-López and Gómez-Pérez (2002).

[21] http://protege.stanford.edu/.

[22] http://code.google.com/p/swoop/.

[23] http://www.hozo.jp/ckc07demo/.

ity lies in the notion of role and the ability to distinguish concepts depending on particular contexts from basic concepts to ensure an easier ontology reuse. Besides this editing function, several other functionalities can be provided in ontology engineering frameworks, such as Schema XML translating functions, graph display of parts of the ontology, ontology modules management, ontology partition, translation of vocabularies, import functions of Web ontologies, access to ontology search engines, text processing modules (like Tree-Tagger[24] or Stanford Parsing tools), help for personalizing ontologies, generating documentation, managing ontology evolution, ontology evaluation, ontology alignment, reasoning and inference services, navigation assistance services, visualisation services, … As an illustration, most of these functionalities are available as plug-ins in the Neon[25] framework.

Some frameworks are designed to deal with a specific kind of data. Therefore, Text2Onto, successor of TextToOnto, and DaFOE4App are specially designed to use text documents and thesaurus as input knowledge sources. Text2Onto (Cimiano and Völker 2005) includes a text mining software and modules that generate structured information from weakly structured documents. Text2Onto is associated with KAON (Karlsruhe Ontology Management Infrastructure) framework (Oberle et al. 2004) in order to design ontologies. DaFOE4App (Differential and Formal Ontology Editor for Applications) (Szulman et al. 2009) focuses on the linguistic dimension while its design uses some of the ARCHONTE methodology principles (Bachimont et al. 2002). DaFOE4App covers all stages from corpora analysis (using a NLP framework) to the definition of a formal domain ontology. It guarantees persistence, traceability and the dimensioning of models (several millions of concepts). The TERMINAE framework (Aussenac-Gilles et al. 2008), designed before DaFOE4App, has evolved with the specifications of DaFOE4App. TERMINAE[26] was used and evaluated in many projects. To end this non-exhaustive list, PlibEditor is more specially tailored to databases. With PlibEditor, users can perform all the tasks required to design ontologies, import or export ontologies as well as data. PlibEditor is complementary to OntoDB, an ontology-based database system and it enables a database approach based on domain ontologies (Fankam et al. 2009).

3.4 Model Reuse

Just as software engineering aims to reuse software components, knowledge acquisition promotes the reuse of knowledge components. This reusability can be achieved in various ways.

Initially proposed in the settings of the KADS project, reuse of problem-solving models consists in taking up task models expressed in a domain-independent terminology and adapting them to specific tasks. This approach is attractive. However,

[24] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[25] http://www.neon-toolkit.org/wiki/Neon_Plugins.

[26] http://lipn.univ-paris13.fr/terminae/.

two specific problems are of paramount importance when adapting a problem-solving model to a specific domain. First, an application often performs several types of reasoning, with several models associated to each of them that have to be distinguished and combined. Second, the reuse and adaptation of predefined generic models to a specific application is difficult and highly time consuming. Indeed, both the task to be completed and the knowledge base of the system must be expressed in the terms of the same application domain, whereas reusable methods coming from libraries, are expressed using a generic vocabulary. Therefore, adapting problem-solving elements to an application is first and mainly a problem of term matching. Consequently, these challenges have led to more flexible approaches with reusable and adaptable elements of a finer granularity. Such approaches imply reusing parts of reasoning models instead of full generic problem-solving models.

Based on the KADS project's outcome, some frameworks support the combination of generic components. They include rich libraries of components as well as graphical editors dedicated to knowledge formalisation, task representation, and the selection and configuration of the methods allowing to solve the tasks (Musen et al. 1994). Solution to adapt generic models to a specific application are diverse, ranging from manual instantiation procedures (Beys et al. 1996) to automated processes including mechanisms that check the specification consistency (Fensel et al. 1996). The CommonKADS project settings led to the most successful results to design problem-solving models. The CommonKADS expertise model can be built by abstraction process or reusing components of problem-solving models. Its particular strength lies in the library of components with different granularities, and with a reuse and adaptation process guided by a questions grid which ensures the relevancy of designed model.

Ontology design is also shaped by the need to reuse existing models. The number of domain ontologies has grown significantly, their popularity being explained in part by the ability to reuse them from one information system to another. Specifically, ontology reuse aims at reducing the difficulties in ex-nihilo developments that constitute real obstacles to some applications. Issues raised by ontology reuse include: the selection of reusable and relevant ontologies, the specific support required to reuse large and complex ontologies that are hard to comprehend, and the integration of various reused ontologies in the under development ontology.

Ontology reuse has motivated the design of ontology search engines such as Watson,[27] Swoogle,[28] or OntoSearch.[29] Using key words, these engines provide a list of ontologies containing at least one concept, one relationship or another element labelled or identified by one of the key words. Then selecting the most relevant ontologies in this list requires that each ontology could be evaluated individually and that ontologies could be compared to eachother according to various criteria. Therefore, how to assess an ontology and to compare several ontologies is currently one of the main challenges in the field. Various questions should be addressed in order

[27] http://kmi-web05.open.ac.uk/WatsonWUI/.

[28] http://swoogle.umbc.edu/.

[29] http://asaha.com/ebook/wNjE3MzI-/OntoSearch--An-Ontology-Search-Engine.pdf.

to tackle this challenge: What criteria can be used? How to understand the modelling perspective adopted in an ontology? How to merge two ontologies? To what extend do two ontologies reflect the same conceptualisation of a given domain? Can we describe the differences in relation to level of detail, compatibility, key concepts and coverage? Are the differences artificial shifts (i.e. consequences of technical or terminological choices) or profound semantic differences that reflect diverging conceptualisations? A major area of research work focused on the development of algorithms and tools to identify and solve differences between ontologies (i.e. analysis of differences between terms, concepts, definitions). Moreover, some research studies bear on global ontologies comparison providing an overview on commonalities and differences. One interesting research direction is to best exploit ontology visualisation results. Visualisation software tools applied to large ontologies provide global views and some of them specifically enable the identification of the ontology main concepts.

The notion of knowledge pattern, directly based on the *design patterns* used in software engineering, aims at reducing the significant difficulties occurring when designing large ontologies or when adapting reusable ontologies. Knowledge pattern has been introduced in Ontology Engineering by Clark et al. (2000) and then in semantic web applications by Gangemi et al. (2004), Rector and Rogers (2004) and Svatek (2004). Knowledge patterns are recurrent and shared representations of knowledge, explicitly represented as generic models and validated through a cooperative process by the research community. Therefore, they are easily reusable after a further processing by symbolic relabelling required to obtain specific representations. Knowledge patterns provide "building blocks" that ensure faster ontology design.[30] Moreover, they lead to better results by solving, for instance, design problems and content-related issues independently of the conceptualisation (Gangemi 2005). Additionally, patterns can facilitate the application of good modelling practices (Pan et al. 2007). The "Semantic Web Best Practices and Deployment" W3C working group promotes the use of ODPs to design ontologies. A library of knowledge patterns is provided in the settings of the European NeOn project. It includes structural, correspondence, content, reasoning, presentation and lexico-syntactic patterns (Presutti et al. 2008). The eXtreme Design (XD) methodology provides guidelines for pattern-based ontology design (Daga et al. 2010).[31]

Reuse of knowledge models requires also to manage their integration within the system under development in order to allow for an easy communication between the reused model and the other models. Although ontologies aim at facilitating interoperability between applications they usually originate from different designers and refer to various modelling perspectives. Therefore, their use within a same application requires to solve specific issues associated with semantic heterogeneity. In practice, the same terms may be used to label different concepts in each reused ontology or ontology module; the same concepts may have different labels; and a particular concept can be characterised by different features in each model. Facing this het-

[30] Referred to as *Ontology Design Pattern* or ODP.

[31] http://ontologydesignpatterns.org/wiki/Main_Page.

erogeneity, significant progress has been made on *model reconciliation*. Models can be reconciled at two different levels. At the schema level, reconciliation consists in identifying correspondences or mappings between semantically-related entities of two ontologies. In the past years, considerable efforts have been made to build ontology alignment tools (Euzenat and Shvaiko 2013), many of which are available on the internet such as OnAGUI[32] or TAXOMAP (Hamdi et al. 2009). Each year since 2004, OAEI international campaigns aim at comparing ontology matching systems. At the data level, reconciliation consists in determining if two data descriptions refer to the same entity of the real world (e.g. the same person or the same hotel). This problem is referred to as *reference reconciliation* (Saïs et al. 2009) and it is close to coreference resolution in NLP.

3.5 Knowledge Representation in Models

Even though designing knowledge representation languages is not KE's main objective, researchers, when specifying knowledge and models, contribute to develop, evaluate and evolve these languages within normalisation groups, such as W3C. Knowledge representation languages as well as modelling languages were first dedicated to problem-solving and reasoning. Then, they related to ontologies (cf. Sects. 2, 2.1, 2.2); nowadays knowledge representation languages are back hand in hand with reasoning.

In the 1980s, ontology representation languages successfully took advantage of logic and conceptual graphs (Sowa 1984). Conceptual graphs could provide both a logic formalisation and a graphical symbolism when no powerful HMI was available to display semantic networks or trees, and to deploy or close them upon request. OWL was later developed as an evolution of DAML+OIL,[33] a language resulting from the merge of the DAML[34] and OIL project outcomes (Fensel et al. 2001). Drawn also on description logic (cf. Sect. I.5), and defined as a layer above XML, OWL became stable and included three languages OWL Lite, OWL-DL, OWL-full according to the W3C recommendations. Each of these three languages specificities results from the trade-off representativity *versus* calculability. In 2007, OWL was extended with new features. A new version, called OWL 2, was formally defined in 2012 with three sub-languages[35] (called *profiles*) offering distinct advantages, computational properties or implementation possibilities, in particular application scenarios: OWL 2 EL enables polynomial time algorithms for all standard reasoning tasks; OWL 2 QL enables conjunctive queries to be answered in LogSpace;

[32] https://github.com/lmazuel/onagui.

[33] http://www.w3.org/TR/daml+oil-reference.

[34] http://www.daml.org/.

[35] https://www.w3.org/TR/owl2-new-features/#F15:_OWL_2_EL.2C_OWL_2_QL.2C_OWL_2_RL.

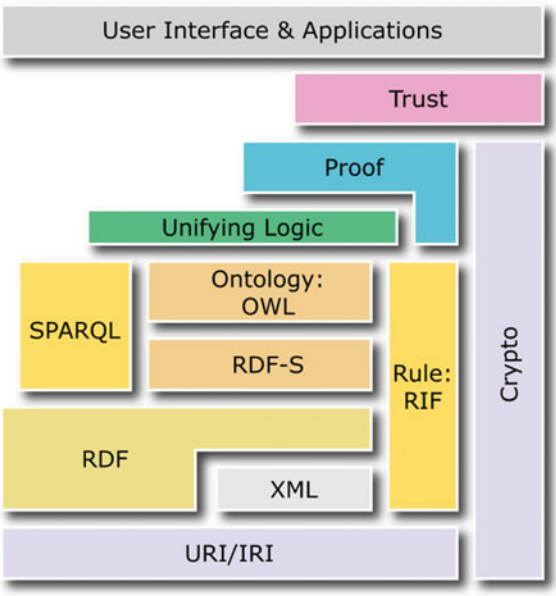


Fig. 2 The layer *cake* of the semantic Web as proposed in 2009

OWL 2 RL enables the implementation of polynomial time reasoning algorithms using rule-extended database technologies.

In the Semantic Web Stack proposed by Tim B. Lee (cf. Fig. 2), representing the stacking order of the Semantic Web languages, we can notice that RDF,[36] located in the bottom part, is the basic language of the Semantic Web. RDF is the common ground to all the languages of interest for KE (i.e. RDF, RDF-S, OWL, SPARQL and RIF). These languages allow applications to consistently use ontologies and associated rules. RDF is a simple language to express data models as a graph where nodes are web resources and edges properties. RDF Schema[37] is a semantic extension of RDF. It is written in RDF and provides mechanisms to structure data models, by describing groups of related resources and the relationships between these resources. OWL is another and more expressive extension allowing a better integration of ontologies and easier inferences. SPARQL[38] is an RDF semantic query language for databases, able to retrieve and manipulate data stored in RDF format. RIF[39] (Rule Interchange Format) is the rule layer in the Semantic Web Stack. RIF is not a rule language but rather a standard for exchanging rules among rule systems. Other rule languages may apply on ontologies, like SWRL,[40] or Description Logic Programs (DLP)[41] (Hitzler et al. 2005). None of them is proposed as a standard for

[36] https://www.w3.org/RDF/.
[37] https://www.w3.org/TR/rdf-schema/.
[38] https://www.w3.org/TR/rdf-sparql-query/.
[39] https://www.w3.org/TR/rif-overview/.
[40] http://www.w3.org/Submission/SWRL/.
[41] http://logic.aifb.uni-karlsruhe.de/wiki/DLP.

the semantic web, because the W3C assumes that a single language would not satisfy the needs of many popular paradigms for using rules in knowledge representation.

Another W3C recommendation defined as an application of RDF is SKOS[42] (for Simple Knowledge Organisation System). SKOS provides a model for expressing the basic structure and content of concept schemes such as thesauri, taxonomies, folksonomies, and other similar types of controlled vocabulary. In basic SKOS, conceptual resources (concepts) are related to each other in informal hierarchies but no logical inference is possible. Using SKOS, generalisation versus specialisation, (*broader-than* and *narrower-than* - – BT/NT) relations that are very often used in thesaurus can be represented without logical inferences associated to the subsumption relationship in OWL.

SKOS was even more necessary in that logical inferences based on the subsumption relationship are only valid if ontologies comply with the associated constraints (whereas such relationship is not valid on thesaurus). Furthermore, the applications using thesaurus and ontologies are increasingly efficient and the resources themselves – i.e. thesaurus and ontologies – are involved in the development processes using different knowledge representation languages at different steps in the development process and not always as intended by the language designers. For instance, a thesaurus and an ontology jointly used in an application can be modelled in OWL for that application. However, one could be originally developed in SKOS and the other one in OWL, and they could further be distributed in a format like CTS2.[43]

4 Methodological Issues and Today's Applications

The current KE challenges are both methodological and application oriented. A few founding principles tackle those issues and provide a general framework:

- The need for a multidisciplinary approach taking into account the recommendations of other disciplines such as cognitive psychology, ergonomics, management, linguistics, information retrieval, natural language processing or document management.
- The importance of a thorough modelling approach, bringing together different models whenever required during the system development process.
- The need to consider upstream the system ergonomic design, prior to any modelling stage; more specifically, the targeted uses of the system should be taken into account as well as its integration in the broader information processing architecture.

KE-related applications form a vast field of research, experimentation and transfer of AI technologies in which innovative methods must be developed. The articulation between methodology and applications guides the stakes described below.

[42] https://www.w3.org/TR/2009/REC-skos-reference-20090818/.

[43] http://www.3mtcs.com/resources/hl7cts.

4.1 Linking Language, Knowledge and Media

As said in Sect. 3.1, natural language is an ideal vector of knowledge, and written natural language is now a good support for knowledge extraction thanks to recent advances in NLP and machine learning techniques. To represent and manage knowledge from text, KE has to deal with various interdisciplinary methodological issues that appear in concordance with classes of applications related to various media.

4.1.1 Designing Problem-Solving Models and Ontologies from Natural Language in Textual Documents

In the 1990s, the first KE studies on knowledge acquisition for expert systems focused on text to identify heuristic knowledge and more or less explicitly explain human reasoning. At that time, text sources were either existing documents or documents elaborated for modelling purposes, such as transcriptions of interviews. Later, the focus on domain ontologies accentuated the sometimes provisional dissociation between the heuristic reasoning and the description of the concepts (and vocabulary) used by these heuristics. Subsequently, at the end of the 1990s, under the impetus of research studies like the one of the French TIA Group, textual corpora generated in relation with an activity were used to help design ontologies for support systems of this same activity. Thus textual corpora were considered as a complementary or alternative source of knowledge to experts and specialists in the field. Processing such corpora requires not only NLP tools but also platforms able to use the result of these tools to design ontologies, terminologies or any conceptual scheme. (cf. Sect. 3.3.2).

Moreover, in this perspective, the document as such is a valuable knowledge conveyer in its own right. The management of documents produced and used in the individual and collective activity, but also, as such, the management of documentary collections (images, sounds, videos) is of interest to KE. KE can then rely on document management technologies that support the sharing, dissemination, archiving, indexing, structuring or classification of documents or document flows. A major difficulty is to select the right documents in order to best meet the users' needs and to find the useful task supports (including knowledge). Because more and more KE projects integrate document management in a large variety of forms, researchers in the field cannot free themselves from an in-depth reflection on the notion of a document, particularly a digital document. To this end, several researchers contributed to the work of the multidisciplinary thematic network on the document (RTP-DOC) and its productions (Pédauque 2003, 2005).

4.1.2 Information Retrieval with Ontologies

Thanks to the Semantic Web, where ontologies provide metadata for indexing documents, ontologies are now at the heart of Information Retrieval (IR) applications. In

this context, they make it easier to access to relevant resources, because they can be used to link and integrate distributed and heterogeneous sources at both the schema and data level. Ontologies are also a means to query multiple sources using a unified vocabulary, to enrich queries with close concepts or synonym terms, to filter out and classify the query results. Given that thesauri are already in use in this field, this line of work obviously leads to compare the gains and limitations of ontologies with those of thesauri or terminologies and to evaluate their respective contributions to IR. These analyses contribute to specify which kind of ontology is more likely to support IR: those having a strong linguistic component, with at least many terms labeling the concepts. As a consequence, a new need emerged: the implementation of application environments where ontologies and thesaurus co-exist to serve the purpose of IR (Vandenbussche and Charlet 2009).

4.2 Coping with Data Explosion

For nearly 20 years, the amount of available data exploded. In a parallel movement, the Semantic Web turned out to be a web of Data in addition to a web a document. This means that the semantics should also be brought to data by labeling them with ontology concepts. Thus applications address increasingly numerous and diverse data that generate new needs in particular for their description and their integration. The so-called *Big Data* is frequently characterised by the four (or more) versus (4Vs): Volume, Velocity, Variety, Veracity. Velocity has to do with efficiency and calculability of knowledge representation, which is out of the scope of this chapter. In the following paragraphs, we explore the three others characteristics: Veracity, Variety, and, for the Volume problematic, we focus more specifically on the question of the size of designed models.

4.2.1 Volume

The description of these very numerous data requires the development of models in which the amount of information to be taken into account can be large enough to open new perspectives to statistical approaches and models. In order to maintain the use and management of symbolic models, the challenge is to be able to design models of very large size, for example by reducing the amount of information to be taken into account simultaneously. In this way, work on ontology modularity aims at designing very large ontologies needed for applications, and to consider these ontologies as sets of (more or less independent) modules. Modularity, in the general sense of the word, refers to the perception of a large knowledge repository (i.e. an ontology, a knowledge or data base) as a set of smaller repositories. Although the concept of modularity is widely used in computer science, it is a relatively new idea in KE. For example,

the Knowledge Web project[44] (2004–2007) provided guidelines to design modular ontologies (Stuckenschmidt et al. 2009). This project showed the diversity of views on modularity and pointed out the important research directions to be developed: guidelines to design modules (how to determine a coherent and meaningful set of concepts, relationships, axioms and instances), metadata to describe, to select and to use or re-use modules, specification of how they can be linked to one another, their composition and their reuse in different contexts. Managing a large mass of data in a distributed context can also lead to designing on a set of existing ontologies that need to be redesigned, aligned, transformed into modules or integrated with non-ontological resources such as databases, folksonomies or thesauri. The networked ontology construction method defined by the NeOn[45] project (2006–2010) includes a support for cooperative design and takes into account the dynamic and evolutionary features of ontologies (Gómez-Pérez and Suárez-Figueroa 2009), which are major issues for the development of large ontology-based applications.

4.2.2 Variety or Managing Knowledge Integration Through Ontologies

Both in the fields of databases and information retrieval, ontologies are experimented as a promising solution for data integration. When integrating data from multiple and heterogeneous sources, ontologies can help to understand and interpret data belonging to the same domain but represented in heterogeneous structures. Then ontologies are also a good support to relate them more easily (Assele Kama et al. 2010). In some domains, such as geography, few ontologies are practically available for data integration (Buccella et al. 2009) or they describe targeted domains, such as Towntology for planning and urbanism (Roussey et al. 2004) or FoDoMuSt in the field of image processing (Brisson et al. 2007). The challenge then consists in designing useful ontologies.

In other domains, like agriculture or medicine, ontologies exist but are very large and therefore difficult to exploit. In this case, the challenge is to enable the understanding of their content in order to help extract the relevant subset for an application. In the medical field, many classifications contain several tens of thousands of concepts and an ontology includes several hundred thousand concepts. Ontology reuse and management reaches an additional level of complexity: ontologies are developed to represent knowledge of a precise sub-domain, we speak of *Interface ontology*. Other large ontologies are developed to provide broad representations and to serve as references for future epidemiological studies, we speak of *Reference ontology* (Rosenbloom et al. 2006). In this context, the best known models are SNOMED-CT that covers the whole medical domain (Spackman 2005) and FMA for representing human anatomy in whole (Rosse and Mejino 2003). Between the two types of ontologies, we need alignment services and the possibility of extracting the relevant subsets for a target system. This is what a standard like CTS2 allows (cf. Sect. 3.5).

[44] http://cordis.europa.eu/ist/kct/knowledgeweb_synopsis.htm.

[45] http://www.neon-project.org/.

This context, reinforced by the need to exploit diversified knowledge or several partial models (or modules), requires to face the problem of heterogeneity between models/ontologies/knowledge, and motivates the current interest in semantic interoperability. Research work on semantic interoperability bears on automatic mapping tools that set links between elements of semantically heterogeneous concept schemes, ontologies or other knowledge sources. They define processes for schema matching, ontology alignment (cf. Sect. 3.4), or data reconciliation. For instance, recent medical studies have tried to integrate most of the knowledge needed to make a diagnosis – e.g. clinical, imaging, genomics knowledge – thanks to a pivotal ontology based on various available ontologies or models (Hochheiser et al. 2016; Sarntivijai et al. 2016).

4.2.3 Veracity

Veracity points out, with a step backwards, two things.

The quality of data is often a problem. For example, in medicine, the medical staff generally inputs data into information systems through poor interfaces, with little time, in difficult working conditions or with little involvement. As a consequence, the data quality is poor too. In a KE point of view, it is important to stress that quality ontologies, and quality Knowledge Organisation Systems in general, are necessary.

Secondly, it appears that medical data are coded (or tagged with concepts) with precise goals and strict coding rules. This process involves a reduction of the meaning, and raises difficulty when interpreting the data, which often requires to read again the original text or resource. Indeed, when reusing data in a new context or when trying to merge it with other data, we observe that the data is biased by the first context. It is then necessary to closely analyse the bias and to check that it can be taken into account or even compensated for in another way. Knowledge engineers must be aware of these limitations and anticipate them before data reuse.

4.3 Managing Distributed Data

The web and web standards have greatly changed the way data is distributed. In particular, new types of systems, web services, rely on a new communication protocol between machines. Thanks to web services, the Web became a distributed computing device where programs (services) can interact intelligently by being able to automatically discover other services, to negotiate among themselves and to compose themselves into more complex services. A considerable amount of knowledge is mandatory to get intelligible services from machines. When added a knowledge base, web services become *semantic web service*.

Semantic web services are the bricks to create a semantic Web of services whose properties, capabilities, interfaces and effects are described in an unambiguous way and can be exploited by machines. The semantics thus expressed must facilitate

the automatic management of services. Semantic web services are essential for the effective use of web services in industrial applications. However, they still raise a number of issues for the research community, including for the KE field because they use ontologies to explain which service they provide to other services or to end users. Semantic modelling contributes to evaluate the quality of a Web service and to take it into account in the process of discovery or composition of services. Peer-to-peer (P2P) systems have also grown significantly, and a substantial body of research work has recently sought to improve the search function in unstructured systems by replacing random routing with semantically guided routing. Several dimensions of the problem are analysed: Which semantics should be remembered? Which representation to adopt? How to design it? What is shared among peers? How to use semantics? How to disseminate it? These issues remained unresolved and have been brought into sharper focus by KE.

4.4 Leveraging New Knowledge Sources

Two knowledge sources currently raise major challenges: data from the Web 2.0 and data from the Web data-bases (web of data).

The Web 2.0 or social Web (OReilly 2007) devotes a considerable attention to users compared to the Web in its initial version, by allowing them to become active. Both authors and actors, Internet users can use the web 2.0 tools to store, implement and manage their own content and share it. These tools include blogs, social networks, collaborative sites, linking platforms, and on-line sharing services. These tools and services are increasingly used in organisations. However, the software tools managing these contents have their own data format and they are increasingly distributed and heterogeneous. These features raise important problems of information integration, reliable identification of the authors or history tracking to name but a few. Similarly, *tagging* or *labeling*[46] is a common practice to characterize and group similar contents and to facilitate data search. This process presents several limitations due to the ambiguity and heterogeneity of the labels, called *tags*. Enterprise 2.0 systems (McAfee 2006) recently tend to develop as a field of experimentation and promotion for KE techniques. It enables a kind of renewal within the KE domain by making new proposals for facilitating navigation, querying or retrieval. As proposed by Tim Berners-Lee, linked Web data refer to an RDF-based publication and interconnection of structured data on the Web, based on the RDF model. Tim Berners-Lee talks about a *Web of data*. It thus promotes a W3C project that goes in this direction, i.e. the Linking Open Data (LOD). The Web of Data, following the web of documents, intends to face the flood of information by connecting the data. Linked data has the advantage of providing a single, standardised access mechanism rather than using different interface and result formats. Data sources can be more easily searched

[46]I.e. content indexing with user's metadata. The sets of labels then form *folksonomies*.

by search engines, accessed using generic data browsers, and linked to different data sources.

The number of data published according to the principles of linked data is growing rapidly (we are talking about billions of RDF triplets available on the Internet). The site http://lov.okfn.org/dataset/lov/ gives a snapshot of existing vocabularies (more than 600) and highlights the numerous mutual reuse of terms between these vocabularies. Among this large number of data sources, DBPedia[47] structures the content of Wikipedia[48] into RDF triples so as to make the information of the encyclopedia reusable. DPpedia is a very powerful source as it is interconnected with other data sources, such as Geonames[49] and MusicBrainz[50]) and it has been linked to even larger data sets like YAGO[51] (Rebele et al. 2016) or BabelNet[52] (Navigli and Ponzetto 2012). These large generic knowledge bases are also used by search engines to display structured content in response to users' queries. Because of they propose unambiguous and linked vocabularies, these masses of data represent promising sources for KE.

4.5 Coping with Knowledge Evolution

The dynamic nature of the data on the Web gives rise to a multitude of problems related to the description and analysis of the evolution of such data. The existing models of knowledge representation are inadequately addressing the challenges of data evolution and, above all, they do not benefit from any adaptive mechanism that would allow them to rigorously follow the evolutions of a domain. Research work on ontology evolution underlines how much the Semantic Web and KE communities need to find appropriate solutions to this complex issue. Early studies defined the stages of an evolution process (Noy and Klein 2004; Stojanovic 2004), they specified a typology of changes (Plessers et al. 2007) and change descriptions. Other works proposed mechanisms, sometimes borrowed to belief revision (Flouris 2006) to keep the modified ontology consistent and logically sound (Haase and Stojanovic 2005) and defined how to propagate changes in distributed ontologies and in the applications that use them (Stuckenschmidt and Klein 2003). With similar purposes to ontology engineering, ontology evolution can be fed thanks to the knowledge identified in textual documents using NLP tools (Buitelaar and Cimiano 2008) and relying on document structure, like in (Nederstigt et al. 2014). More recently, when the ontology is used to generate semantic annotations of text, research studies deal

[47] http://wiki.dbpedia.org/.

[48] https://fr.wikipedia.org.

[49] http://www.geonames.org/.

[50] https://musicbrainz.org/.

[51] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.

[52] http://babelnet.org/.

with the evolution of these semantic annotations when the textual corpus or when the indexing vocabularies evolve (Tissaoui et al. 2011; Da Silveira et al. 2015; Cardoso et al. 2016).

Zablith et al. (2015) propose a recent overview of the major trend in this domain. Characterizing and representing domain data evolution raises issues both at the data level (Stefanidis et al. 2016) and at the model scheme level (Guelfi et al. 2010). Ontology evolution remains a hard issue, even at the era of machine learning, because a statistic processing of a massive amount of documents is relevant for building large knowledge bases like DBpedia, but produces poor results when trying to fix errors or to identify local changes in an existing model. Processing large amounts of data is much more appropriate to feed and update the data level in knowledge bases, which corresponds to instances of ontological classes.

4.6 Collective Versus Personal Knowledge

Most of the previous approaches place little emphasis on the social dimension of knowledge management. This dimension is strong enough in some professional communities to consider them as communities of interest or as communities of practices. Communities of practices designate social groups in which learning processes emerge through the sharing of networked knowledge. KE models need to capture these learning processes or to integrate them into their knowledge management process. To this end, Lewkowicz and Zacklad (2001) propose a new form of knowledge management based on the structuring of collective interactions. This approach aims at better using of the shared knowledge, at facilitating its reuse, the knowledge of an organisation being considered as above all a matter of collective competence.

The identification of communities of interest that emerged thanks to the development of Web 2.0 or the analysis of users' digital traces sharing similar thematic information implies the representation of individual knowledge about the fields of interest and activities of their members, together with the collective dimension of knowledge. This collective dimension is the focus of the Computer Supported Cooperative Work (CSCW) research community, that designs specific solutions to manage collective and in-use knowledge. For instance, M. Zacklad proposes a conceptual model mid-way between thesauri and formal ontologies, called *semiotic ontologies*, that should be more easily shared by a working community in an information retrieval framework (Zacklad 2007). Conversely, more and more software systems and Web interfaces are designed to be context sensitive or user customised. To do so, they adapt to the user profile, environment or interactions with the system, which requires the acquisition, the modelling and the processing of the interaction contexts (Garlatti and Prié 2004).

4.7 Model Quality Assessment

Finally, a fundamental question for KE concerns the quality assessment of the models used and the results produced. The use of poor quality knowledge may lead to errors, duplications and inconsistencies that must be avoided. Beyond its interest in research, the theme of quality has become critical with the deployment of systems in companies.

The quality of the models/ontologies can be guaranteed methodologically, when the ontology was designed following a rigorous method based on the theoretical and philosophical foundations of what an ontology is (such as the methods presented in Sect. 5). Other methodological works aim to move from manual and approximative approaches, the cost and duration of which are difficult to estimate, to more systematic, equipped and better controlled processes. Of course, they focus on reuse such as Methontology (Gómez-Pérez et al. 2007) and NEON in Suárez-Figueroa et al. (2012), on practical guidelines (Noy and Hafner 1997) or on systematic text analysis using NLP tools and modelling platforms such as Terminae (Aussenac-Gilles et al. 2000) or GATE and methods listed in Maedche (2002). In the case of Brank et al. (2005), a state of the art classifies the ontology evaluation techniques into four categories: (1) syntactic evaluations check whether the model complies the syntactic rules of a reference language (RDF, OWL, …) such as Maedche and Staab (2002), (2) in-use evaluations test the ontology when used by a targeted system, e.g. Porzel and Malaka (2004) (3) comparison with a reference source in the domain (either a gold model or a representative set of textual documents), such as Brewster et al. (2004) or, finally (4) human evaluation tests how well the ontology meets a set of predefined criteria, standards or needs, for example Lozano-Tello and Gomez-Perez (2004). Moreover, in Brank et al. (2005), validation approaches are organised into six levels: lexical level, level of taxonomic relations, level of other semantic relationships, application level (looking how the ontology impacts on the system that uses it), context level (how the ontology is reused by or reuses another ontology), syntactic level or, finally, the level of design principles. Practically, it may be easier to evaluate an ontology level by level because of its complexity.

5 Conclusion

KE has undergone successive changes of direction. This research field constantly evolves from the inside (experimenting new analyses, new perspectives, original ways of posing problems, new theoretical concepts) and from outside (targeting new types of applications, dealing with new types of data, in particular with the upheavals of the Web, integrating the contributions of other disciplines that come to bring new methods and concepts). Over the years, these developments gradually broadened the scope of KE. Each new proposed theoretical framework includes parts of the

previous work. Even if some changes of perspective correspond to actual breaks, the results of the domain complement each other over time and can be taken from a new angle when the context evolves.

For a long time, KE has been interested in producing knowledge models in a well-structured process under the control of knowledge engineers. The resulting models, generally complex, were used in specific applications. Today, applications in which knowledge is used as support for reasoning or activity have become much more diversified. Since 2000, they have been devoted to knowledge management in the broadest sense, including semantic information retrieval, navigation aids, decision support, and many semantic Web applications. This enlargement continues and new fields of application are still emerging, posing the problems of KE in new terms.

Thus, in the age of ubiquitous computing, it is the living room, the train, the automobile, the workshop, the classroom or meeting room, the smallest kitchen device that become "smart" tools. Within these tools, a dynamic process is required to continuously acquire context knowledge on the flow from a wide variety of sources (sensors, databases, the Internet, users with various profiles). In addition, these intelligent tools must have a pro-active behaviour that enables them to initiate communication or action based on their understanding of the current situation and on their goals. So, for example, phones know where we are at a given time and become capable of automating some operations, such as when taking pictures, labeling them with geographic and temporal metadata.

The last decade has seen a major transformation in the way individuals interact and exchange. Information is now co-produced, shared, filed and evaluated on the Web by thousands of people. These uses and the underlying technologies are known as Web 2.0. Web 3.0 is the latest evolution to date that combines the social web and the semantic technologies of the semantic Web. In the context of communities of interest or practices where spontaneous emergence and activity are allowed by these evolutions of the Web, KE and knowledge management are thus major stakes of the future decade.

Finally, KE must feed and evaluate all these new developments, compare them with previous models (reasoning models, rules bases), estimate the need to use ontologies and their alignment to type or organise data, to define new techniques and languages if necessary, to justify the use of metadata to enrich and reuse data, and so on. The speed of Web evolutions can be seen as a crazy accelerator of the research pace or as an alarm that invites us to step back and pose the problems at a higher abstraction level, necessarily interdisciplinary, in order to better qualify the essence of knowledge, their dissemination and their formalisation for digital processing.

References

Assele Kama A, Mels G, Choquet R, Charlet J, Jaulent M-C (2010) Une approche ontologique pour l'exploitation de données cliniques. In: Despres S (ed) Acte des 21èmes Journées Francophones d'Ingénierie des Connaissances, Nîmes, France. Ecole des Mines d'Alès, pp 183–194

Aussenac N (1989) Conception d'une méthodologie et d'un outil d'acquisition de connaissances expertes. Thése de doctorat, Université Paul Sabatier, Toulouse, France

Aussenac-Gilles N, Jacques M-P (2008) Designing and evaluating patterns for relation acquisition from texts with caméléon. Terminology 14(1):45–73 (special issue on Pattern-based approaches to semantic relations)

Aussenac-Gilles N, Krivine J, Sallantin J (1992) Editorial du numéro spécial Acquisition des connaissances. Revue d'Intelligence Artificielle 6(2):7–18

Aussenac-Gilles N, Bourigault D, Condamines A (1995) How can knowledge acquisition benefit from terminology? In: Proceedings of the 9th knowledge acquisition workshop, Banff, University of Calgary (CA)

Aussenac-Gilles N, Laublet P, Reynaud C (eds) (1996) Acquisition et ingénierie des connaissances: tendances actuelles. Cepadues Editions, Toulouse

Aussenac-Gilles N, Biébow B, Szulman S (2000) Revisiting ontology design: a method based on corpus analysis. In: 12th international conference on knowledge engineering and knowledge management, Juans-Les-Pins, 03/10/2000–06/10/2000. Springer, Heidelberg, pp 172–188

Aussenac-Gilles N, Després S, Szulman S (2008) The TERMINAE method and platform for ontology engineering from texts. Ontology learning and population: bridging the gap between text and knowledge, pp 199–223

Bachimont B, Isaac A, Troncy R (2002) Semantic commitment for designing ontologies: a proposal. In: EKAW, pp 114–121

Barriere C, Agbago A (2006) Terminoweb: a software environment for term study in rich contexts. In: International conference on terminology, standardization and technology transfer, Beijing, pp 103–113

Beys B, Benjamins V, Van Heijst G (1996) Remedying the reusability - usability trade-off for problem-solving methods. In: Gaines B, Musen M (eds) Proceedings of the 10th knowledge acquisition workshop (KAW), Banff, Canada, pp 2–1/2-20

Bourigault D (2002) Upery: un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In: Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy, France, pp 75–84

Bourigault D, Slodzian M (1999) Pour une terminologie textuelle. Terminol Nouv 19:29–32

Bourigault D, Aussenac-Gilles N, Charlet J (2004) Construction de ressources terminologiques ou ontologiques à partir de textes: un cadre unificateur pour trois études de cas. Revue d'Intelligence Artificielle 18(1/2004):87–110

Brank J, Grobelnik M, Mladenic D (2005) A survey of ontology evaluation techniques. In: Data mining and data warehouses conference (SIKDD), Lubiana, Slovénie

Brewster C, Alani H, Dasmahapatra S, Wilks Y (2004) Data driven ontology evaluation. In: LREC

Brisson R, Boussaid O, Gançarski P, Puissant A, Durand N (2007) Navigation et appariement d'objets géographiques dans une ontologie. In: EGC, pp 391–396

Buccella A, Cechich A, Fillottrani P (2009) Ontology-driven geographic information integration: a survey of current approaches. Comput Geosci 35(4):710–723 (Geoscience knowledge representation in cyberinfrastructure)

Buitelaar P, Cimiano P (eds) (2008) Proceedings of the 2008 conference on ontology learning and population: bridging the gap between text and knowledge, Amsterdam, The Netherlands. IOS Press

Camilleri G, Soubie J-L, Zaraté P (2008) A replanning support for critical decision making situations: a modelling approach. In: Intelligent decision making: an AI-based approach, pp 173–192

Cardoso SD, Pruski C, Da Silveira M, Ying-Chi L, Anika G, Erhard R, Reynaud-Delaître C (2016) Leveraging the impact of ontology evolution on semantic annotations. In: Knowledge engineering and knowledge management - 20th international conference, EKAW, Bologna, Italy
Chandrasekaran B (1983) Towards a taxonomy of problem solving types. AI Mag 4(1):9–17
Charlet J (1991) ACTE: a strategic knowledge acquisition method, pp 85–93
Charlet J, Zacklad M, Kassel G, Bourigault D (eds) (2000) Ingénierie des connaissances: Evolutions récentes et nouveaux défis. Eyrolles, Paris
Cimiano P, Völker J (2005) Text2onto. In: NLDB, pp 227–238
Cimiano P, Buitelaar P, Völker J (2010) Ontology construction. In: Indurkhya N, Damerau, FJ (eds) Handbook of natural language processing, 2nd edn. CRC Press, Taylor and Francis Group, Boca Raton. ISBN 978-1420085921
Clark P, Thompson JA, Porter BW (2000) Knowledge patterns. In: KR, pp 591–600
Condamines A (2002) Corpus analysis and conceptual relation patterns. Terminol. Int J Theor Appl Issues Spec Commun 8(1):141–162
Constant M, Dister A, Ermikanian L, Piron S (2008) Description linguistique pour le traitement automatique du français. Cahier du CENTAL
Cordier M-O, Reynaud C (1991) Knowledge acquisition techniques and second-generation expert systems. Appl Artif Intell 5(3):209–226
Da Silveira M, Dos Reis J, Pruski C (2015) Management of dynamic biomedical terminologies: current status and future challenges. Yearb Med Inform 24:125–133
Daga E, Blomqvist E, Gangemi A, Montiel E, Nikitina N, Presutti V, Villazon-Terrazas B (2010) NeOn project: NeOn D2.5.2. Pattern-based ontology design: methodology and software report. Rapport de contrat
Darses F, Montmollin M (eds) (2006) L'ergonomie. La Découverte - Col. Repères, Paris
Dieng-Kuntz R, Corby O, Gandon F, Gibouin A, Golebiowska JNM, Ribière M (eds) (2005) Knowledge management: Méthodes et outils pour la gestion des connaissances. Dunod
Drouin P (2003) Term extraction using non-technical corpora as a point of leverage. Terminology 9:99–117
Euzenat J, Shvaiko P (2013) Ontology matching, 2nd edn. Springer, Heidelberg
Fankam C, Bellatreche L, Hondjack D, Ameur YA, Pierra G (2009) Sisro, conception de bases de données à partir d'ontologies de domaine. Technique et Science Informatiques 28(10):1233–1261
Fensel D, Schnanegge R, Wielinga B (1996) Specification and verification of knowledge-based systems. In: Proceedings of the 10th knowledge acquisition workshop (KAW), Banff (Can). University of Calgary (Can)
Fensel D, van Harmelen F, Horrocks I, McGuinness DL, Patel-Schneider PF (2001) Oil: an ontology infrastructure for the semantic web. IEEE Intell Syst 16(2):38–45
Fernández-López M, Gómez-Pérez A (2002) Overview and analysis of methodologies for building ontologies. Knowl Eng Rev 17(2):129–156
Flouris G (2006) On belief change in ontology evolution. AI Commun 19(4):395–397
Gangemi A (2005) Ontology design patterns for semantic web content. In: International semantic web conference, pp 262–276
Gangemi A, Catanacci C, Battaglia M (2004) Inflammation ontology design pattern: an exercise in building a core biomedical ontology with descriptions and situations. In: Maria PD (ed) Ontologies in medecine. IOS Press, Amsterdam
Garlatti S, Prié Y (2004) Adaptation et personnalisation dans le web sémantique. Revue I3 - Numéro hors série Web Sémantique
Gómez-Pérez A, Suárez-Figueroa M-C (2009) Scenarios for building ontology networks within the neon methodology. In: K-CAP 2009, pp 183–184
Gómez-Pérez A, Fernández-López M, Corcho O (2007) Ontological engineering: with examples from the areas of knowledge management, e-commerce and the semantic web. (Advanced information and knowledge processing). Springer, New York
Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquis 5:199–220

Guarino N, Welty CA (2004) An overview of ontoclean. Handbook on ontologies, pp 151–172
Guelfi N, Pruski C, Reynaud C (2010) Experimental assessment of the target adaptive ontology-based web search framework. In: NOTERE, pp 297–302
Haase P, Stojanovic L (2005) Consistent evolution of owl ontologies. In: ESWC, pp 182–197
Hamdi F, Safar B, Niraula N, Reynaud C (2009) TaxoMap in the OAEI 2009 alignment contest. In: The fourth international workshop on ontology matching, Chantilly, Washington DC, États-Unis
Hendler JA, Tate A, Drummond M (1990) AI planning: systems and techniques. AI Mag 11(2):61–77
Hitzler P, Sure Y, Studer R (2005) Description logic programs: a practical choice for the modelling of ontologies. In: Principles and practices of semantic web reasoning
Hochheiser H, Castine M, Harris D, Savova G, Jacobson RS (2016) An information model for computable cancer phenotypes. BMC Med Inform Decis Mak 16(1), 121
Jacob-Delouis I, Krivine J (1995) Lisa: un langage réflexif pour opérationnaliser les modèles d'expertise. revue d'Intelligence. Artificielle 9(1):53–88
Kamel M, Aussenac-Gilles N (2009) Utiliser la Structure du Document dans le Processus de Construction d' Ontologies (regular paper). In: L'Homme M-C, Szulman S (eds) Conférence Internationale sur la Terminologie et l'Intelligence Artificielle (TIA), Toulouse (France), 18–20/11/2009, page (on line). http://www.irit.fr/ (IRIT)
Kassel G (2002) Ontospec: une méthode de spécification semi-informelle d'ontologies. In: Actes d'IC, pp 75–87
Klinker G, Bhola G, Dallemagne G, Marquès D, Dermott M (1991) Usable and reusable programming constructs. Knowl Acquis 3:117–136
Laflaquière J, Prié Y, Mille A (2008) Ingénierie des traces numériques d'interaction comme inscriptions de connaissances. In: Actes d'IC, pp 183–195
Lewkowicz M, Zacklad M (2001) Une nouvelle forme de gestion des connaissances basée sur la structuration des interactions collectives. In: Grundstein M, Zacklad M (eds) Ingénierie et Capitalisation des connaissances. Hermes Sciences Europe LTD, pp 49–64
Lozano-Tello A, Gomez-Perez A (2004) ONTOMETRIC: a method to choose the appropriate ontology. J Database Manag 15(2):1–18
Luong PH (2007) Gestion de l'évolution d'un web sémantique d'entreprise. Thèse de doctorat, Ecole des Mines de Paris, Paris, France
Maedche A (2002) Ontology learning for the semantic web. Kluwer Academic Publisher, Boston
Maedche A, Staab S (2002) Measuring similarity between ontologies. In: EKAW, pp 251–263
Manning C, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge
Marcus S, McDermott J (1989) SALT: a knowledge acquisition language for propose and revise systems. Artif Intell 39(1):1–38
Maynard D, Funk A, Peters W (2009) SPRAT: a tool for automatic semantic pattern-based ontology population. In: International conference for digital libraries and the semantic web
McAfee A (2006) Enterprise 2.0: the dawn of emergent collaboration. MIT Sloan Manag Rev 47(3):21–28
Meyer I (2000) Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework. In: Bourigault D, L'Homme M-C, Jacquemin C (eds) Recent advances in computational terminology
Musen MA, Eriksson H, Gennari JH, Tu SW, Puert AR (1994) PROTEGE-II: a suite of tools for development of intelligent systems for reusable components. In: Proceedings of the annual symposium on computer application in medical care
Navigli R, Ponzetto SP (2012) BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif Intell 193:217–250
Neches R, Fikes R, Finin TW, Gruber TR, Patil RS, Senator TE, Swartout WR (1991) Enabling technology for knowledge sharing. AI Mag 12(3):36–56
Nederstigt LJ, Aanen SS, Vandic D, Frasincar F (2014) FLOPPIES: a framework for large-scale ontology population of product information from tabular data in e-commerce stores. Decis Support Syst 59:296–311
Newell A (1982) The knowledge level. Artif Intell 18(1):87–127

Noy NF, Hafner CD (1997) The state of the art in ontology design: a survey and comparative review. AI Mag 18(3):53–74
Noy NF, Klein MCA (2004) Ontology evolution: not the same as schema evolution. Knowl Inf Syst 6(4):428–440
Oberle D, Volz R, Staab S, Motik B (2004) An extensible ontology software environment. Handbook on ontologies, pp 299–320
OReilly T (2007). What is web 2.0: design patterns and business models for the next generation of software. Commun. Strat (1):17
Pan J, Lancieri L, Maynard D, Gandon F, Cuel R, Leger A (2007) Success stories and best practices. Knowledge web deliverable d.1.4.2.v2
Pédauque RT (ed) (2003) Le document: forme, signe et medium les re-formulations du numérique. STIC-CNRS
Pédauque RT (ed) (2005) Le texte en jeu, permanence et transformations du document. STIC-SHS-CNRS
Plessers P, Troyer OD, Casteleyn S (2007) Understanding ontology evolution: a change detection approach. J Web Semant 5(1):39–49
Poibeau T, Kosseim L (2000) Proper name extraction from non-journalistic texts. In: CLIN, pp 144–157
Porzel R, Malaka R (2004) A task-based approach for ontology evaluation. In: ECAI - workshop on ontology, learning and population
Presutti V, Gangemi A, David S, De Cea GA, Surez-Figueroa MC (2008) NeOn project: NeOn D2.5.1. a library of ontology design patterns: reusable solutions for collaborative design of networked ontologies - NeOn project. Rapport de contrat
Puerta A, Egar JW, Tu SW, Musen M (1992) Method knowledge-acquisition shell for the automatic generation of knowledge-acquisition tools. Knowl Acquis 4(2):171–196
Rastier F (2009) Sémantique interprétative. PUF
Rebele T, Suchanek FM, Hoffart J, Biega J, Kuzey E, Weikum G (2016) YAGO: a multilingual knowledge base from wikipedia, wordnet, and geonames. In: The semantic web - ISWC 2016 - 15th international semantic web conference, Kobe, Japan, 17–21 October 2016, proceedings, part II, pp 177–185
Rector A, Rogers J (2004) Patterns, properties and minimizing commitment: reconstruction of the GALEN upper ontology in OWL. In: EKAW
Reymonet A, Thomas J, Aussenac-Gilles N (2007) Modélisation de ressources termino-ontologiques en owl. In: Actes d'IC, pp 169–181
Reynaud C, Aussenac-Gilles N, Tchounikine P, Trichet F (1997) The notion of role in conceptual modeling. In: EKAW, pp 221–236
Rosenbloom S, Miller RA, Johnson KB (2006) Interface terminologies: facilitating direct entry data into electronic health record systems. J Am Med Inform 13(3):277–288
Rosse C, Mejino JLV (2003) J Biomed Inform 36(6):478–500
Roussey C, Laurini R, Beaulieu C, Tardy Y, Zimmermann M (2004) Le projet towntology: un retour d'expérience pour la construction d'une ontologie urbaine. Revue Internationale de Géomatique 14(2):217–237
Saïs F, Pernelle N, Rousset M-C (2009) Combining a logical and a numerical method for data reconciliation. J Data Semant 12:66–94
Sarntivijai S, Vasant D, Jupp S, Saunders G, Bento AP, Gonzalez D, Betts J, Hasan S, Koscielny G, Dunham I, Parkinson H, Malone J (2016) Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. J Biomed Semant 7:8
Schreiber G, Wielinga B (eds) (1992) KADS: a principled approach to knowledge-based system development. Academic, London
Schreiber G, Wielinga BJ, Akkermans H, de Velde WV, Anjewierden A (1994) CML: the commonKADS conceptual modelling language. In: EKAW, pp 1–25

Schreiber G, Akkermans A, Anjewierden A, DeHoog R, Shadbolt N, Van de Velde W, Wielinga B (eds) (1999) Knowledge engineering and management: the CommonKADS methodology. MIT Press, Cambridge

Schutz A, Buitelaar P (2005) RelExt: a tool for relation extraction from text in ontology extension. In: International semantic web conference, pp 593–606

Shadbolt N, O'Hara K, Crow L (1999) The experimental evaluation of knowledge acquisition techniques and methods: history, problems and new directions. Int J Hum-Comput Study 51(4):729–755

Sowa JF (1984) Conceptual structures: information processing in mind and machine. Addison-Wesley, London

Spackman KA (2005) Rates of change in a large clinical terminology: three years experience with SNOMED clinical terms. In: AMIA annual symposium proceedings, pp 714–718

Steels L (1990) Components of expertise. AI Mag 11(2):28–49

Stefanidis K, Flouris G, Chrysakis I, Roussakis Y (2016) D2V - understanding the dynamics of evolving data: a case study in the life sciences. ERCIM News 2016(105)

Stefik M (1995) Introduction to knowledge systems. Morgan Kaufmann, San Francisco

Stojanovic L (2004) Methods and tools for ontology evolution. PhD thesis

Stuckenschmidt H, Klein MCA (2003) Integrity and change in modular ontologies. In: IJCAI, pp 900–908

Stuckenschmidt H, Parent C, Spaccapietra S (eds) (2009) Modular ontologies: concepts, theories and techniques for knowledge modularization, vol 5445. Lecture notes in computer science. Springer, Berlin

Studer R, Benjamins VR, Fensel D (1998) Knowledge engineering: principles and methods. Data Knowl Eng 25(1–2):161–197

Suárez-Figueroa M-C, Gómez-Pérez A, Motta E, Gangemi A (eds) (2012) Ontology engineering in a networked world. Springer, Berlin

Svatek V (2004) Design patterns for semantic web ontologies: motivation and discussion. In: Conference on business information systems

Szulman S, Charlet J, Aussenac-Gilles N, Nazarenko A, Sardet E, Teguiak V (2009) DAFOE: an ontology building platform from texts or thesauri. In: Dietz J (ed) Proceedings of the international joint conference on knowledge discovery, knowledge engineering and ontology development (KEOD 2009), Madeira (Portugal). Poster, pp 1–4

Tissaoui A, Aussenac-Gilles N, Hernandez N, Laublet P (2011) EVONTO - joint evolution of ontologies and semantic annotations. In: KEOD 2011 - proceedings of the international conference on knowledge engineering and ontology development, Paris, France, 26–29 October 2011, pp 226–231

Tu SW, Eriksson H, Gennari JH, Shahar Y, Musen MA (1995) Ontology-based configuration of problem-solving methods and generation of knowledge-acquisition tools: application of PROTÉGÉ-II to protocol-based decision support. Artif Intell Med 7:257–289

Vandenbussche P-Y, Charlet J (2009) Méta-modèle général de description de ressources terminologiques et ontologiques. In: Actes d'IC, pp 193–204

Virbel J, Luc C (2001) Le modèle d'architecture textuelle: fondements et expérimenation XXII I(1):103–123

Zablith F, Antoniou G, d'Aquin M, Flouris G, Kondylakis H, Motta E, Plexousakis D, Sabou M (2015) Ontology evolution: a process-centric survey. Knowl Eng Rev 30(1):45–75

Zacklad M (2007) Classification, thesaurus, ontologies, folksonomies: comparaisons du point de vue de la recherche ouverte d'information (roi). In: Conférence CAIS/ACSI

Afterword – From Formal Reasoning to Trust

Luis Fariñas del Cerro

To formalize the different types of reasoning is a main task for logic. Initially, the focus was on reasoning in all its generality, then, in the first half of the last century, it shifted towards mathematical reasoning, promoting its unprecedented development. In that period, contributions that were dedicated to the foundation of mathematics gave a precise meaning to the concept of algorithm that plays a central role in our discipline, computer science. Generally speaking, works from e.g. Alonzo Church, Jacques Herbrand, Alain Turing and Rudolf Carnap laid the foundations of many domains of computer science.

The problematic of logic was renewed by computer science, in particular since the advent of artificial intelligence, highlighting that reasoning is not only mathematical.

We are used to reason and take decisions in our daily life even if we do not have an accurate knowledge about our environment and more specifically about actions that other human beings we interact with may perform. We only have partial knowledge of the laws and conventions of communication with other humans. The modeling of such interactions was, for a long time, a very important subject of philosophy and philosophical logic, and has become an important subject of artificial intelligence, motivated by the need of modeling interactions between human and artificial agents.

The incomplete, uncertain, or partially contradictory character of such information manipulated by artificial agents makes use of very diverse types of reasoning. These include among others, hypothetical reasoning, reasoning with uncertainty, and temporal reasoning.

The formalisation of such types of reasoning is one of the essential challenges that artificial intelligence poses to logic. In that perspective, we have recently seen the development of new logical systems that try to capture certain aspects of reasoning, for instance:

L. Fariñas del Cerro (✉)
IRIT-CNRS, University of Toulouse, Toulouse, France
e-mail: luis.farinas@irit.fr

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7

- Substructural logics
- Linear logic
- Logics of uncertainty
- Modal logics
- Non-monotonic logics

Most prominently, non-monotonic logics that allow us to treat laws with exceptions is a privileged domain of artificial intelligence. Thanks to numerous equivalence results these formalisms converge towards a unique framework that is based on the notion of preference. An important aspect of all these logical systems is the definition of automated deduction mechanisms. Important results have been obtained recently that can be traced back to both a clearer understanding of the logics.

The formalisation of reasoning is a problem by itself that is at the heart of both the foundations of artificial intelligence as well as of its applications. Naturally, it has interfaces with many other domains, in particular artificial intelligence and formal logic. In recent years, important efforts have been devoted to the introduction of humans as a fundamental element in the modeling and implementation of computing systems. This allows us to imagine other ways to design these computing systems that require the support of multiple disciplines of different nature, such as ethics and economy. Let us take for example the notion of trust, which is a fundamental concept in actual computing systems.

Trust has been studied in a large number of disciplines, such as philosophy, psychology, economy, and computer science.

Here are some everyday life examples that show its importance regarding our use of the internet:

- Is the information shown on that website correct?
- I have got two contradicting pieces of information on two different websites, which one should I believe?
- Is this email safe?
- Can this web service use my information without my consent?
- Does this web service offer what I am asking for and not something different?

If we look at the different models of decentralised open systems, such as the web, we can see that the concept of trust is at the top of the different elements taken into consideration by such systems. See for example, in Fig. A.1, the diagram of the semantic web proposed by Tim Berners-Lee.[1]

We can also note that the notion of trust raises many other questions, for example:

- How is it justified or motivated?
- What is the ability to produce actions from intentions?
- Are the norms met?
- What is the mutual interest of two agents?

[1] Image from https://www.w3.org/2002/Talks/04-sweb/slide12-0.html

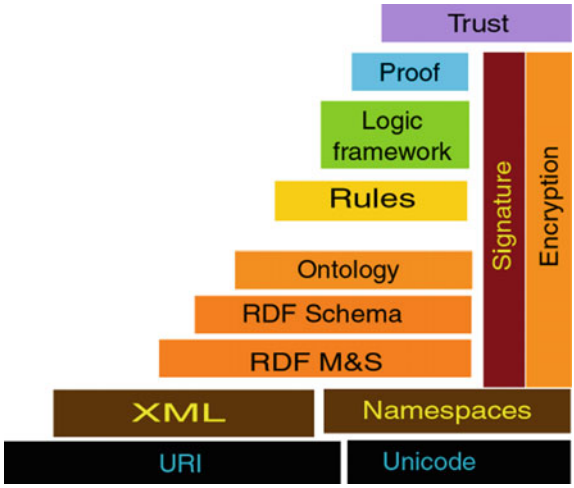


Fig. A.1 Diagram of the semantic web proposed by Tim Berners-Lee

In the relation between agents, trust plays an important role, in particular when computer agents play an intermediary role and act on behalf of human agents or institutions, which is the case e.g. for the relations between a bank and its clients for example.

When information management is entrusted to a computer agent, trust issues arise in the preservation of:

- Privacy;
- Integrity;
- Information availability and access.

Think of an action such as an online hotel reservation: it is immediately clear that agents have to quickly decide if they can or cannot trust the transaction, and to which extent such trust is justified.

The justification of this type of beliefs can be of different kinds:

- Empirical, if it comes from observation or reputation.
- Analytical, if the validity of an information source can be deduced from (beliefs about) its sincerity and competence. For example, the validity of an information source can be
- deduced from its sincerity and its competence.

When we say that we trust a computer system, we actually assert that we trust the method by which the system was designed and implemented, since we believe that design and implementation followed a series of protocols and standards.

This is the case for all artefacts constructed by man that we are using in our every day lives. Thus, when we push the brake pedal we believe that the car will slow down. In other words, we trust the fact that best practice design and production rules were followed during the manufacturing process.

The fundamental question is to know what explicit, measurable, and communicable criteria ensure that the software that we are using follows standards, or in other words, that it possesses a good behavior.

These kind of procedure was followed in a very efficient and thorough way in so-called critical systems, such as systems tied to transportation or to energy production. It should also be adopted by all these new computer systems that we use in our everyday life and that allow human or institutional agents to connect to each other. When studying the notion of trust, artificial intelligence researchers should study ethical problems, in the sense that they should better understand and explicit the impact of the actions done by the computer systems that are at the origin of these new interactions between human beings.

These thoughts, around our example, show that artificial intelligence researchers need to develop new formal frameworks integrating concepts that allow to model and mechanise increasingly rich and complex reasoning modes. They should also allow the formalisation of high-level cognitive capabilities, such as those related to trust or emotion. All this should be taken into consideration if we want to improve the cooperation between human and computer agents.

Index

P. Marquis et al. (eds.), *A Guided Tour of Artificial Intelligence Research*,
https://doi.org/10.1007/978-3-030-06164-7

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y