UMEÅ UNIVERSITY

Principle-based Non-Monotonic Reasoning - From Humans to Machines

Timotheus Kampik

Doctoral Thesis, April 2022
Department of Computing Science
Umeå University
Sweden

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

tkampik@cs.umu.se

The following papers have been published by Springer International Publishing:
Paper I, Timotheus Kampik, Dov Gabbay.
Explainable Reasoning in Face of Contradictions: From Humans to Machines.
In *Explainable and Transparent AI and Multi-Agent Systems*, 2021.
Paper IV, Timotheus Kampik, Dov Gabbay.
The Degrees of Monotony-Dilemma in Abstract Argumentation.
In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 2021.
Paper V, Timotheus Kampik, Kristijonas Čyras.
Explanations of Non-monotonic Inference in Admissibility-Based Abstract Argumentation.
In *Logic and Argumentation*, 2021.
Paper VI, Timotheus Kampik, Dov Gabbay, Giovanni Sartor.
The Burden of Persuasion in Abstract Argumentation. In *Logic and Argumentation*, 2021.

The following paper has been published by Oxford Academic Publishing:
Paper II, Timotheus Kampik, Juan Carlos Nieves.
Abstract Argumentation and the Rational Man. In *Journal of Logic and Computation*, 2021.

The following paper has been published by Elsevier:
Paper III, Timotheus Kampik, Juan Carlos Nieves, Dov Gabbay.
Ensuring reference independence and cautious monotony in abstract argumentation.
In *International Journal of Approximate Reasoning*, 2022.

The following paper has been published by the Institute of Electrical and Electronics Engineers:
Paper VII, Helena Lindgren, Timotheus Kampik, Esteban Guerrero Rosero, Madeleine Blusi, Juan Carlos Nieves. Argumentation-Based Health Information Systems: A Design Methodology.
In *IEEE Intelligent Systems*, 2021.

ISBN 978-91-7855-757-8
ISSN 0348-0542
UMINF 22.02

Printed by City Print i NorrAB, Umeå, 2022.
Title image by Mike Kononov/from Unsplash.com

Abstract

A key challenge when developing intelligent agents is to instill behavior into computing systems that can be considered as intelligent from a "common-sense" perspective. Such behavior requires agents to diverge from typical decision-making algorithms that strive to maximize simple and often one-dimensional metrics. A striking parallel to this research problem can be found in the design of formal models of human decision-making in microeconomic theory. Traditionally, mathematical models of human decision-making also reflect the ambition to maximize expected utility or a preference function, which economists refer to as the *rational man* paradigm. However, evidence suggests that these models are flawed, not only because human decision-making is subject to systematic fallacies, but also because the models depend on assumptions that do not hold in reality. Consequently, the research domain of formally modeling *bounded rationality* emerged, which attempts to account for these shortcomings by systematically relaxing the mathematical constraints of the formal model of economic rationality. Similarly, in the field of symbolic reasoning, approaches have emerged to systematically relax the notion of *monotony of entailment*, which stipulates (colloquially speaking) that when inferring a set of statements from a knowledge base, the addition of new knowledge to the knowledge base must not lead to the rejection of any of the previously inferred statements.

By drawing from these developments in microeconomic theory and symbolic reasoning, this thesis explores different principle-based approaches to decision-making and non-monotonic reasoning. Thereby, abstract argumentation is used as a fundamental method for reasoning in face of conflicting knowledge (or: *beliefs*) that reduces non-monotonic reasoning to the problem of drawing conclusions (*extensions*) from a directed graph, and hence provides a neat abstraction for theoretical exploration. In particular, the works collected in this thesis *i)* introduce the *consistent preferences* property of microeconomic theory, as well as some relaxed forms of monotony of entailment as mathematical principles to abstract argumentation-based inference; *ii)* show how to enforce some of these principles in dynamic environments; *iii)* devise a formal approach to maximize monotony of entailment, given the constraints imposed by an inference function; *iv)* extend and apply the aforementioned approaches to the domains of machine reasoning explainability and legal reasoning.

Sammanfattning

Principbaserat icke-monotont resonemang - Från människor till maskiner

En central utmaning i utvecklingen av autonoma mjukvaruagenter är att bibringa datorsystem beteende som kan anses vara intelligent från ett "sunt förnuft"-perspektiv. Metoder för att generera sådant beteende krävs som alternativ till algoritmer som optimerar enkla, och ofta en-dimensionella metriker för beslutfattande, som i komplexa situationer inte är optimala. En motsvarighet till det här problemet finns i formella modeller av mänskligt beslutfattande i mikroekonomisk teori. Sådana matematiska modeller baseras traditionellt på maximering av en individs egen nytta eller preferenser, ett paradigm som ekonomer refererar till som "rational man". Emellertid, empiriska studier visar att sådana modeller är bristfälliga, för det är vanligt att människor begår systematiska fel i sådana slutledningar, men också för att modellerna gör antaganden som inte håller i verkligheten. Ekonomer har därför utformat nya modeller för *avgränsad* rationalitet som syftar till att adressera dessa begränsningar.

Liknande metoder utvecklas inom det vetenskapliga området *formellt, symbol-baserat resonerande och beslutsfattande* inom artificiell intelligens (symbolisk AI) för att tillåta att ny kunskap läggs till i kunskapsbasen så att när man tidigare har dragit en slutsats från ett kunskapsbas, måste man inte förkasta en ny slutsats, även om den är i konflikt med den tidigare slutsatsen (s.k. icke-monontont resonemang, till skillnad från *monotony of entailment*).

I den här avhandlingen tas ekonomiska modeller och modeller baserade på symbolisk AI som en utgångspunkt för att utforska olika metoder för att bibringa datorsystem "sunt förnuft"-baserat beteende. Särskilt studeras *abstrakt argumentation*, som modellerar icke-monotont resonemang som slutledning från en riktad graf, vilket också används för att illustrera resultaten från den teoretiska forskning som presenteras. Sammanfattningsvis bidrar avhandlingen med följande: *i) konsistenta preferenser* introduceras, en egenskap från mikroekonomisk teori, och även olika former av mildrad *monotony of entailment* som matematiska principer för abstrakt argumentation-baserat resonerande; *ii)* det visas hur dessa principer kan garanteras i dynamiska miljöer; *iii)* en formell metod har utformats för att maximera *monotony of entailment*, med beaktande av de begränsningar som en inferensfunktion medför; samt *iv)* dessa principer och metoder har byggts ut och används för att generera förklarbart automatiskt beslutsfattande och juridiskt resonemang.

Acknowledgements

I thank the following persons:

- My family, in particular Sam, Lisa, and Katya;
- My supervisors Helena Lindgren and Juan Carlos Nieves;
- My reference person, Thomas Hellström;
- Our *Interactive and Intelligent Systems* group, as well as all other researchers at Umeå University's Department of Computing Science, and in particular my fellow PhD students;
- The excellent support and administrative staff at the Department, as well as Frank and Erik (doctoral education/department management);
- My WASP study peers, Chris and Tobias, as well as everyone who keeps WASP running;
- The scientific community and my collaboration partners around the world, in particular (this list is non-exhaustive):
  - Amro Najjar, Dov Gabbay, and everyone in Leendert van der Torre's group at Luxembourg University, who have welcomed me so nicely;
  - Cleber Jorge Amaral, Jomi Hübner, and Cesar Tacla for inviting me to Brazil; covid killed the plans, but I will visit eventually!
  - Guillermo Simari for the nice words and valuable feedback during my Licentiate seminar;
  - Kristijonas Čyras for the neat collaboration on explainability principles;
  - Giovanni Sartor for an excellent mentoring session and the nice research results that emerged from it;
  - Andrei Ciortea for inviting me to his Dagstuhl seminar, and for his great efforts to bring agent-oriented software engineering closer to practice;
  - The *Governance of Agents on the Web*-crowd: Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian Padget, Terry Payne, Munindar Singh, Valentina Tamma, and Antonoine Zimmermann;
  - The many reviewers who gave me valuable feedback.
- Ron and the Product Team at Signavio/SAP.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

List of papers

This thesis is based on the following papers.

Paper I Timotheus Kampik, Dov Gabbay. Explainable Reasoning in Face of Contradictions: From Humans to Machines. In *Explainable and Transparent AI and Multi-Agent Systems*, Springer International Publishing, 2021.

Paper II Timotheus Kampik, Juan Carlos Nieves. Abstract Argumentation and the Rational Man. In *Journal of Logic and Computation*, Oxford Academic Publishing, 2021.

Paper III Timotheus Kampik, Juan Carlos Nieves, Dov Gabbay. Ensuring Reference Independence and Cautious Monotony in Abstract Argumentation. In *International Journal of Approximate Reasoning*, Elsevier, 2022.

Paper IV Timotheus Kampik, Dov Gabbay. The Degrees of Monotony-Dilemma in Abstract Argumentation. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Springer International Publishing, 2021.

Paper V Timotheus Kampik, Kristijonas Čyras. Explanations of Non-monotonic Inference in Admissibility-Based Abstract Argumentation. *Logic and Argumentation*, Springer International Publishing, 2021.

Paper VI Timotheus Kampik, Dov Gabbay, Giovanni Sartor. The Burden of Persuasion in Abstract Argumentation. *Logic and Argumentation*, Springer International Publishing, 2021.

Paper VII Helena Lindgren, Timotheus Kampik, Esteban Guerrero Rosero, Madeleine Blusi, Juan Carlos Nieves. Argumentation-Based Health Information Systems: A Design Methodology. *IEEE Intelligent Systems*, Institute of Electrical and Electronics Engineers, 2021.

The following papers have been formally published but are not included in the thesis.

Journal Papers

Paper VIII Timotheus Kampik, Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian Padget, Terry R. Payne, Munindar P. Singh, Valentina Tamma, Antoine Zimmermann. Governance of Autonomous Agents on the Web: Challenges and Opportunities. In *ACM Transactions on Internet Technology*, Association for Computing Machinery. 2022.

Paper IX Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, Christophe Nicolle. The Quest of Parsimonious XAI: a Human-Agent Architecture for Explanation Formulation. In *Artificial Intelligence*, Elsevier, 2022.

Paper X Timotheus Kampik, Amro Najjar. Simulating, Off-Chain and On-Chain: Agent-Based Simulations in Cross-Organizational Business Processes. In *Information*, MDPI, 2020.

Conference Papers and Workshop Papers in Formal Proceedings

Paper XI Timotheus Kampik, Cleber Jorge Amaral, Jomi Fred Hübner. Developer Operations and Engineering Multi-Agent Systems. In *Engineering Multi-Agent Systems*, Springer International Publishing, 2022.

Paper XII Timotheus Kampik, Juan Carlos Nieves. JS-son - A Lean, Extensible JavaScript Agent Programming Library. In *Engineering Multi-Agent Systems*, Springer International Publishing, 2020.

Paper XIII Timotheus Kampik, Juan Carlos Nieves, Helena Lindgren. Empathic Autonomous Agents. In *Engineering Multi-Agent Systems*, Springer International Publishing, 2019.

Paper XIV Timotheus Kampik, Juan Carlos Nieves, Helena Lindgren. Implementing Argumentation-enabled Empathic Agents. In *Multi-Agent Systems*, Springer International Publishing, 2019.

Paper XV Timotheus Kampik, Juan Carlos Nieves, Helena Lindgren. Explaining Sympathetic Actions of Rational Agents. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer International Publishing, 2019.

Paper XVI Timotheus Kampik, Amro Najjar. Technology-Facilitated Societal Consensus. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, 2019.

Paper XVII Timotheus Kampik, Amro Najjar. Integrating Multi-agent Simulations into Enterprise Application Landscapes. In *Integrating Multi-agent Simulations into Enterprise Application Landscapes*, Springer International Publishing, 2019.

Paper XVIII Timotheus Kampik, Avleen Malhi, Kary Främling. Agent-Based Business Process Orchestration for IoT. In *IEEE/WIC/ACM International Conference on Web Intelligence*, Association for Computing Machinery, 2019.

Paper XIX Timotheus Kampik, Juan Carlos Nieves, Helena Lindgren. Coercion and Deception in Persuasive Technologies. In *Proceedings of the 20th International Trust Workshop*, CEUR-WS.org, 2018.

Paper XX Timotheus Kampik, Amro Najjar, Davide Calvaresi. MAS-Aided Approval for Bypassing Decentralized Processes: An Architecture. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, 2018.

Demonstration Papers

Paper XXI Timotheus Kampik, Andres Gomez, Andrei Ciortea, Simon Mayer. Autonomous Agents on the Edge of Things. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, International Foundation for Autonomous Agents and Multiagent Systems, 2021.

Paper XXII Cleber Jorge Amaral, Timotheus Kampik, Stephen Cranefield. A Framework for Collaborative and Interactive Agent-oriented Developer Operations. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, International Foundation for Autonomous Agents and Multiagent Systems, 2020.

Paper XXIII Yazan Mualla, Timotheus Kampik, Igor Tchappi, Amro Najjar, Stéphane Galland, Christophe Nicolle. Explainable Agents as Static Web Pages: UAV Simulation Example. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer International Publishing, 2020.

Doctoral Consortium

Paper XXIV Timotheus Kampik. Empathic Agents: A Hybrid Normative/ Consequentialistic Approach. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Reports and Report Chapters

Paper XXV Timotheus Kampik, Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian Padget, Terry R. Payne, Munindar P. Singh, Valentina Tamma, Antoine Zimmermann. Norms and Policies for Agents on the Web. In *Autonomous Agents on the Web (Dagstuhl Seminar 21072)*, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl Publishing, 2021.

Paper XXVI Viviana Mascardi et al. Engineering Multi-Agent Systems: State of Affairs and the Road Ahead. In *ACM SIGSOFT Software Engineering Notes 44.1*, Association for Computing Machinery, 2019.

Contributions

Paper VII reports on a long-running line of research by the Department's *Interactive and Intelligent Systems* group and hence lists Helena Lindgren as the first author. Here, Timotheus' role was to coordinate and execute the elicitation of the design methodology, to draft and mature the paper, and to provide practice-oriented perspectives. In all other papers, the work of drafting the initial research outline, conducting background research, producing the core parts of the theoretical and empirical results, implementing engineering artifacts, and writing the final paper were mainly conducted by the first author. However, the co-authoring supervisors (or other senior contributors) were continuously involved in most of the work and provided feedback and corrections as well as detailed training and instructions regarding methods and strategies at the different stages of research.

Contents

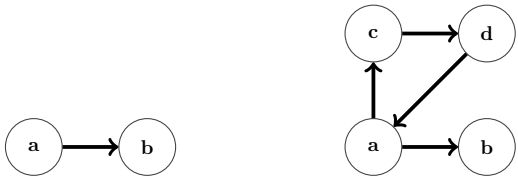
Chapter 1
Introduction

Systems that incorporate Artificially Intelligence (AI) research results are omnipresent in our society, increasing both efficiency and effectiveness of organizations. This facilitates economic growth and potentially also societal progress. However, an issue with the rise of autonomous systems is that decision-making processes are more frequently delegated from humans to machines, which can be problematic when the way an autonomous system makes decisions is not properly understood. For example, when a high-stakes financial decision is made, it is important that the decision outcome is both well-reasoned and transparent, which is currently not always the case. To address this issue, it is worth exploring whether decision-making algorithms and systems comply with *economic rationality* – optimal decision-making paradigms as defined in economic theory – *to the extend that this can be considered reasonable.* Enabling autonomous systems to make this trade-off between *economically rational* and *common sense* behavior can be considered a highly relevant research challenge. From a formal reasoning perspective, a similar line of thought has motivated research on principle-based non-monotonic reasoning approaches that relax the notion of *monotony of entailment*, which stipulates (roughly speaking) that when drawing a conclusion from a knowledge base, the addition of new knowledge to the knowledge base must not cause us to reject any part of our previous conclusion. Similarly to the formal models of bounded rationality that have emerged over the past decades, many methods of non-monotonic "common sense" reasoning have been devised, and many principles that systematically relax monotony of entailment under some conditions have found their way into the literature.

To further advance the latter line of research, and to integrate it with the *bounded rationality* perspective, this thesis sets out to explore the intersection of the design of formal models of bounded rationality, autonomous agents, and non-monotonic reasoning methods. In particular, the seven papers presented in this thesis work towards answering the following research questions:

1. How can one *conceptually* integrate models of (bounded) economic rationality with automated reasoning approaches? (Papers I and II)
2. What is the *formal* relation between economic rationality-based and (relaxed) monotony-based principles of reasoning, in particular in the context of abstract argumentation and how can these principles be enforced and extended in the context of dynamic reasoning scenarios? (Papers II, III, and IV)
3. How can our formal models and principles be extended and applied to different domains? (Papers V, VI, and VII)

In this thesis, we use formal argumentation, and in particular *abstract argumentation* as our formal theory of choice for investigating these questions. Abstract argumentation is a foundational approach to non-monotonic reasoning, in which conflicts between a set of abstract *arguments*, which may – for example – stand for logical statements, are modeled as a binary relation on the argument set. The resulting directed graphs are called *argumentation frameworks*. For instance, Figure 1.1a depicts the argumentation framework AF, consisting of the arguments a and b and a directed edge (attack) from a to b: we say that "a attacks b". Figure 1.1b depicts another argumentation framework AF', which we call a *normal expansion* of AF, because it contains all arguments and attacks of AF, but no new attacks between arguments that already existed in AF have been added. Arguments are abstract: they are merely *elements* whose internal structure is not, but can be, formally modeled. For example, in AF, a and b may stand for so-called business rules in a knowledge-based system, and the attack may model that in case both rules are active, rule a attacks rule b and hence, only a should fire. Or an argumentation framework may model potentially contradicting witness statements in a legal reasoning scenario. These can be modeled using mutually attacking arguments or using a unidirectional attack in case one witness statement is considered superior to another; the latter could then give rise to the argumentation frameworks AF and AF'. From argumentation frameworks, we infer *extensions* – sets of arguments that satisfy certain properties as specified by an inference function, a so-called argumentation semantics. Intuitively, we would expect that any "reasonable" semantics infers $\{a\}$ (and only $\{a\}$) from AF, because there is a conflict between a and b, and a attacks b. In contrast, the argumentation framework AF', as depicted by Figure 1.1b may yield several extensions: assuming the three-argument cycle that includes a, c and d allows us to infer "either a or c or d", a semantics may yield the extensions $\{a\}$, $\{c, b\}$, and $\{d, b\}$. Because of its simplicity, abstract argumentation, as well as methods that extend and generalize it, enjoy tremendous success in the symbolic AI community, in particular as a theoretical tool for investigating fundamental questions of non-monotonic reasoning. In this thesis, the fundamental questions we ask are related to the *consistency* of inferences we draw from argumentation frameworks that are then (normally) expanded, so that we again draw inferences from them and so on; *i.e.*, we study the *dynamics* of formal argumentation. For example, we may introduce a principle for argumentation semantics stipulating that when we infer an extension from an argumentation framework and normally expand the framework, we must be able to infer an extension from the expanded argumentation framework that entails the previously inferred extension, *unless* we have added a new argument that attacks our initial extension (see Paper II). Or, we may stipulate that given we have inferred $\{a\}$ from AF, we want to select an extension a semantics infers from AF' that entails as many arguments of our initial inference $\{a\}$ as possible, which would be again $\{a\}$ in our example (see Papers IV and V). In contrast to many previous works that investigate dynamics in formal argumentation, the work in this thesis does not merely study mathematical properties that appear interesting from an intuitive perspective, but is instead motivated by a range of different interdisciplinary perspectives, ranging from economic decision theory via legal reasoning scenarios to explainable artificial intelligence. In this way, this thesis strengthens the study of formal argumentation not only from formal, but also from philosophical and – broadly, if not immediately – practical perspectives.



(a) AF. (b) AF'.

Figure 1.1: An argumentation framework AF and its normal expansion AF'.

The rest of this thesis is structured as follows. Chapter 2 provides an overview of the background of this thesis at the intersection of autonomous rational agents, economically rational decision-making, and abstract argumentation as a theory of non-monotonic reasoning. Then, Chapter 3 discusses the contribution of the papers that are presented in this thesis to the state of the art. Subsequently, Chapter 4 provides a high-level, conceptual overview of future research directions, as well as concluding remarks, followed by a compilation of the included papers at the end of this thesis.

Chapter 2
Background

This chapter first presents a conceptual contextualization of the the thesis and its results, considering the history and the state of the art of the field. Then, it provides a brief overview of the study of autonomous rational agents, formal models of bounded rationality, and non-monotonic reasoning and abstract argumentation, introducing the theoretical foundations that the thesis rests upon.

2.1 The Thesis in the Context of the History of the Field

This section places the work that this thesis presents in the context of the history (and the state of the art) of the field at the intersection of logic-based reasoning and formal approaches to decision-making. For pragmatic reasons, the overview provided in this section is limited in its temporal scope and does by no means aim at providing a complete survey; instead, it focuses on works and research trends that have had a direct impact on the thesis, and hence is subjective.

In the first half of the twentieth century, mathematical models of reasoning ("Given the knowledge we have, what can one infer to be true?") and decision-making ("How should one act in a given situation?") rose to prominence within scientific circles and beyond. Ground-breaking contributions to the former question were, for example, made by the so-called Vienna circle – an informal collective of scientists and philosophers, active during the 1920s and 1930s – and associated logicians such as Gödel and Tarski[1], resulting in modern classical logic. The latter question was most notably pursued by Von Neumann and Morgenstern in their seminal work *Theory of Games and Economic Behavior* [35] (but Von Neumann also made important contributions to mathematical logic relatively early in his career [23]), resulting in microeconomic decision theory (as well as game theory for strategic decision-making, which is of less relevance for this thesis).

Both classical logic and microeconomic decision theory make somewhat strong assumptions about the world and satisfy formal principles that reflect these assumptions. In classical logic, a crucial assumption is that what has been proven to be true remains true, which in turn leads to the principle of *monotony of entailment*[2]: when

[1] For example, both Gödel and Tarski provided formal results about the undefinability of truth in formal systems that are *sufficiently strong*, *i.e.* given some constraints [22], and Tarski made seminal contributions to classical logic with his semantic theory of truth [31].

[2] Other assumptions are made in classical logic, and are relaxed in different *non-classical* logics. For example, intuitionistic logics reject the *law of the excluded middle* that stipulates that given a proposition, either the proposition or its negation holds true (but not

we have inferred that a statement is true, given our knowledge about the world, we must never infer that this statement is false, even if we obtain more knowledge. The philosophy behind this principle is reflected by the idea of *logical positivism* – *i.e.*, the creation of new knowledge by verification – that is associated with the Vienna circle [32]. In microeconomic theory, a crucial assumption is that decision-makers are *rational*. Hereby, an agent's choices are defined as rational if and only if the agent's choices are *clear* (a preference order is established on the set of choice items, such that one item is preferred over all others) and *consistent* (adding or removing elements from a set of choice items does not lead to choices that establish preferences orders that are inconsistent with previously established orders), *ceteris paribus*, *i.e.* given all else remains the same, a constraint that is assumed, but not explicitly modeled [24] (also see Section 2.3).

From the perspective of practical reasoning, the assumptions that are made by classical logic and microeconomic decision theory may be questioned. When reasoning, one may not rely on knowledge that is and always will be true, but instead on *beliefs* that one considers true at the time one draws an inference and that may later be rejected based on newly obtained beliefs. To account for this, the research direction of non-monotonic logics that reject monotony of entailment emerged and principle-based belief revision has received substantial attention from the mid-1980s onward, when Alchourrón, Gärdenfors, and Makinson introduced their rationality postulates for *belief revision* [1] – which have, however, no formal relation to rationality in the microeconomic sense of the word. Around the same time, Gabbay introduced what presumably is the first principle that *systematically relaxes* monotony of entailment, which allows for the revision of beliefs given some constraints [14]; the corresponding definition was further developed in an influential paper by Kraus, Lehman, and Magidor [20], and is now widely known as *cautious monotony*. Over time, different approaches to non-monotonic reasoning have emerged, in particular logic programming with negation as failure (colloquially: if we fail to infer a statement, this implies that we infer the statement's negation) [8]. Research on non-monotonic reasoning was further boosted by Dung's seminal paper that introduced abstract argumentation [13]. In abstract argumentation, non-monotonic reasoning is reduced to drawing inferences from argumentation frameworks, directed graphs that model an *attack relation* on a set of elements, which may – but do not necessarily – represent logical statements. So-called *argumentation semantics* are then used as inference functions that determine an argumentation framework's *extensions*, *i.e.* sets of arguments with desirable properties (see Section 2.4). Different argumentation semantics exists and the design of argumentation semantics that satisfy certain mathematical principles is a vividly active research direction [3, 33]. Prior to the works that are included in this thesis, principles that are systematic relaxations of monotony of entailment have not been introduced to abstract argumentation, although some related principles exist for approaches that can be considered extensions or generalizations thereof [12, 5].

From the perspective of microeconomic decision-making, the limits of economic rationality were prominently highlighted as early as the 1950s [29]. Over the decades, critics of economic rationality have highlighted, most notably, the following two points: *i)* empirical studies of human decision-making indicate that humans are not economically rational; *ii)* classical models of rational decision-making are simplis-

both) [21]. However, in the context of this work, monotony of entailment is the most relevant principle.

tic, as they do not model the decision-maker's knowledge about the environment, which typically changes over time: in other words, the *ceteris paribus* assumption is typically violated, which limits the models' applicability. Obviously, both points are intertwined; for example, Tversky and Kahneman famously conducted empirical studies that gathered evidence of how human decision-making systematically violates rationality assumptions to then create formal models of *bounded rationality* to account for the empirical evidence [17]. Over time, formally modeling boundedly rational behavior emerged as its own line of research. Thereby, one important research topic is the modeling of knowledge [27] – or, from our perspective: of *beliefs*, which we may later revise. This can be achieved by integrating logic-based non-monotonic reasoning with microeconomic theory.

Although both microeconomic decision theory and (argumentation-based) non-monotonic reasoning are active and influential research directions in which formal principles play a crucial role, the integration of the two fields is weak from a principle-based perspective. A formal relation between game theory and abstract argumentation has already been established in Dung's seminal paper that introduces the latter, and decision-oriented perspectives on argumentation, as well as argumentation-based perspectives on decision and game theory are somewhat frequently studied [2, 16]. However, no principle-based integration of the two research fields existed. The works in this thesis provide contributions toward closing this gap between the fields. In particular, Paper I gives a conceptual overview of research potential at the intersection of (boundedly) rational decision-making and formal argumentation as a collection of methods for non-monotonic reasoning. Then, Papers II and III introduce economic rationality-based principles to abstract argumentation and argumentation-based decision-making. Notably, most argumentation semantics violate the *(weak) reference independence* principle that we introduce based on the notion of *consistent preferences* in economically rational decision-making. Subsequently, Papers IV, V and VI formally explore the relevance of different ways to relax monotony of entailment for argumentation-based decision-making, and hence continue a byline of the contributions of Papers II and III, in which relaxed forms of monotony are dealt with alongside the economic rationality-based principles. However, in contrast to the previous papers, monotony of entailment is not seen as a principle that needs to be weakened, but is instead maximized, either given a cardinality-based *degree of monotony*, or with respect to the inclusion of previously inferred statements. In these papers, decision-oriented perspectives allow us to define novel monotony-maximizing notions that reflect the intuition that in face of doubt one should attempt to be *as consistent as possible*, given some practical constraints that are formally imposed by an inference function. Colloquially speaking, instead of asking "Can we, under the given circumstances, violate monotony?", we say that "Given the options that are available, we are as monotonic – and hence as consistent – as possible". Finally, Paper VII places formal argumentation more broadly into a software engineering context, outlining its application potential.

2.2 Autonomous Rational Agents

This thesis is concerned with the study of the reasoning capabilities of autonomous agents. An autonomous agent is an entity that perceives its environment through sensors, processes its perception in some way(s) (*deliberates*), and interacts with its environment through actuators [28]. Consequently, an agent can be any sensing and acting instance, for example a human or a software system that powers an Internet of Things (IoT) device. The primary objective of the research domain of Artificial Intelligence (AI) is to find new and "better" ways to instill *intelligence* and *autonomy* into artificial agents. In this context, it is traditionally considered desirable that an agent behaves *rationally*, *i.e.* at "each possible percept sequence, [it] should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has" [28, p. 37]. However, a question that arises is to what extent an agent that is fixated on maximizing a particular measure – somewhat reflecting the notion of a *single-minded agent* [26] – is truly intelligent. In a human organizational context this concern is reflected by the colloquialization of Goodhart's Law by Marilyn Strathern as "when a measure becomes a target, it ceases to be a good measure" [30]. Surely, an intelligent agent should be able to systematically revise its beliefs about its environment, and to adjust its goals and actions accordingly. For instance, when an autonomous vehicle that is en route from locations A to B encounters a dangerous situation, it must be able to revise its belief that the goal of reaching B is immediately desirable in order to prioritize the goal of bringing the vehicle to a safe stop. In the spirit of this assumption, this thesis explores formal reasoning methods for instilling behavior into autonomous agents that is aligned with the notion of rationality without suffering from the problem of single-mindedness. Hereby, the key question is *how exactly* notions of rationality or single-mindedness should be relaxed, and which constraints should still apply.

2.3 Formal Models of Bounded Rationality

In order to specify how humans make – or ought to make – decisions, economists have introduced formal models of rational economic choice, the most influential of which is the model of *Rational (Economic) Man*[3]. When choosing from a set of items, which can either represent physical items or abstract decision options, Rational Man's choice establishes *preferences* over the items he chooses from; the preference relation he establishes by choosing have to fulfill certain properties.

Let us first formally define a *partial preference order*, before we introduce the formal model of Rational Man.

Definition 1 (Partial Preference Order)
A preference order $\succeq$ on a set S is a partial preference order iff for all $x, y, z \in S$ it holds true that:

- $x \succeq x$ *(reflexivity);*
- $x \succeq y$ *and* $y \succeq x$ *implies* $x = y$ *(antisymmetry);*
- $x \succeq y$ *and* $y \succeq z$ *implies* $x \succeq z$ *(transitivity).*

For $a \succeq b$ we say that "a is preferred over b". When Rational Man chooses an element a^* of S, this choice establishes his preference of a^* over all all other elements in the set.

Definition 2 (Rational Economic Man [27])
Let S be a set of choice items. Rational Man's choice $a^* \in S$ establishes the following preferences:

$$\forall a \in S, a^* \succeq a$$

The definition of Rational Economic Man implies that the preferences established by choices from different, potentially intersecting sets should be consistent, a concept that is called *reference independence*.[4]

Definition 3 (Reference Independence [27])
Given two sets of choice items S and S', such that $S \subseteq S'$,the following holds true for Rational Man's choices $a^* \in S$ and $a'^* \in S'$:

$$a'^* \notin S \vee a'^* = a^*$$

Note that Rational Economic Man is defined for *choose one from a set*-scenarios. To better relate the underlying properties to non-monotonic reasoning approaches (see: Section 2.4), we define economic rationality for *choose a subset of a set*-scenarios.

[3] We use the terms *Rational Economic Man*, *Rational Man*, and *rational decision-maker* interchangeably. Let us highlight that the formal model of rational economic man does not describe the behavior of actual men (or of humans in general), and we deliberately keep the rather old-fashioned name, including the questionable gendering, to reflect the dated assumptions about human decision-making that the model is based on (see, notably, Tversky and Kahneman [17], as well as Collier's popular science notion of *rational social woman* [9]).

[4] The proof that reference independence emerges from rational choice is, for example, provided by Rubinstein [27].

Definition 4 (Rational Economic Man - Choose Subset of Set)
Let S be a set of choice items. Rational Man's choice $A^* \subseteq S$ establishes the following preferences:

$$\forall A \in 2^S, A^* \succeq A$$

We can adjust the definition of reference independence analogously.

Definition 5 (Reference Independence - Choose Subset of Set)
Given two sets of choice items S and S', such that $S \subseteq S'$, the following holds true for Rational Man's choices $A^* \subseteq S$ and $A'^* \subseteq S'$:

$$A'^* \not\subseteq S \vee A'^* = A^*$$

The definition above is the starting point of the integration of models of economic rationality and abstract argumentation as a non-monotonic reasoning method (see: Section 2.4) we present in Paper II. Let us illustrate the notion of economic rationality with the help of an example.

Example 1
We visit a café and want to place our order. The café has merely two items on its menu: tea and coffee. We choose to order coffee, i.e. *given $S = \{tea, coffee\}$, our choice $A^* = \{coffee\}$ establishes the preferences $\{coffee\} \succeq \{tea, coffee\}$, $\{coffee\} \succeq \{tea\}$, and $\{coffee\} \succeq \{\}$*[5]*. The next day, we enter the café again and find an additional item on the menu: a* cookie. *This time, we choose to order a tea and a cookie. Interestingly, this choice is still* rational *when considering our previous choice, because it merely establishes that $\forall A \in 2^{\{tea, coffee, cookie\}}$, it holds true that $\{tea, cookie\} \succeq A$; while we prefer $\{tea, cookie\}$ over $\{coffee\}$, our previous preferences of $\{coffee\}$ over all elements in $2^{\{tea, coffee\}}$ still hold. In contrast, if we were to order only tea this time, the resulting preference $\{tea\} \succeq \{coffee\}$ would be inconsistent with the preference $\{coffee\} \succeq \{tea\}$ that our previous choice has established.*

An aspect that severely limits the applicability of these basic models of rational decision-making is the lack of consideration of context knowledge, *i.e.* Rational Man's preferences must be consistent over time, *ceteris paribus* (all else unchanged). In the example above, we can, for instance, imagine that we want to order tea (without a cookie) only during weekends, but cannot specify the fact that it is (or is not) a weekend day in our formal model. Approaches that model knowledge in addition to choice options and preferences can account for this limitation to some extent.

Initially, economists hoped that these formal models of rational decision-making cannot only be used as an approximate and simplistic guideline of how to make good decisions, but also to predict decision-making in strategic situations, for example to analyze how to act optimally in the face of an adversary. By now, it is clear that the dependency of these formal models on a fully observable and static environment, as well as on humans whose decision-making is perfectly aligned with formal models of rationality, is not realistic. Indeed, the concept of *bounded rationality* and its premise that humans are not complying with the *rational man* paradigm of microeconomic theory, has already been observed by Herbert Simon in the 1950s [29], although it took several decades for this observation to receive mainstream attention,

[5] Let us ignore the fact that we can potentially order more than one of each item.

most prominently through the Nobel Prize-winning work of Tversky and Kahneman [17]. In this line of work, the researchers conducted experiments that provide empirical evidence that human decision-making is not well-aligned with the rational man paradigm. Some of the experimental evidence was then used to construct formal models of bounded rationality, most notably *prospect theory* [18]. A range of other works that introduce formal microeconomic models of bounded rationality exist, and an overview of such models is given in Rubinstein's book *Modeling Bounded Rationality* [27].

2.4 Non-monotonic Reasoning and Abstract Argumentation

Soon after the advent of the research field of artificial intelligence with the 1956 Dartmouth workshop, AI research split into different sub-domains. Two prominent sub-domains of basic AI research are *machine learning*, of which Marvin Minsky was a proponent at Dartmouth, and logic-based (machine) *reasoning*, whose foundations were laid by John McCarthy and others after McCarthy's attendance of Dartmouth. In contrast to machine learning, which is primarily concerned with the correlation of – often unstructured – data, machine reasoning devises methods and principles for inferring *conclusions* from a knowledge base (structured data), based on the relationships between propositions in this knowledge base.

An important formal property of reasoning methods is *monotony* (or lack thereof, depending on the method) [14]. In monotonic reasoning, the addition of information does not change the inferences that have been drawn from existing information. Let us consider any sets of propositions[6] A, B, and C and let us denote "we infer B from A" by $A \mathrel{|\!\sim} B$. Monotony is satisfied iff it holds true that if $A \mathrel{|\!\sim} B$, then $(A \cup C) \mathrel{|\!\sim} B$. Conversely, any inference function for which if $A \mathrel{|\!\sim} B$, then $(A \cup C) \mathrel{|\!\sim} B$ does not hold true, given some sets of propositions A, B, and C, is *non-monotonic*.

Let us introduce an example that demonstrates the implications of monotony.

Example 2
From our set A that contains the propositions "all birds can fly" and "penguins are birds", we infer the conclusion B "penguins can fly". When adding the new observation C that "penguins cannot fly" to A, we cannot account for the inconsistency of this proposition with the already inferred conclusion B.

In contrast, non-monotonic reasoning methods allow for the revision of conclusions based on new propositions that are potentially contradicting existing propositions and the conclusions that have been drawn from them. Restrictions as to *under which circumstances* the addition of new propositions to a knowledge base is allowed to affect previous conclusions have been introduced, most prominently with the definition of *cautious monotony* (also called *restricted monotony*) [14]. Let us again consider any sets of propositions A, B, and C. Cautious monotony is satisfied iff it holds true that if ($A \mathrel{|\!\sim} B$ and $A \mathrel{|\!\sim} C$) then $(A \cup C) \mathrel{|\!\sim} B$.

Let us explain cautious monotony by example.

Example 3
When applying cautiously monotonous inference rules to the propositions presented in Example 2, we can derive the conclusions we would expect from a common-sense perspective:

1. *A: "All birds can fly" and "penguins are birds".*
2. *B: From A, we infer that "penguins can fly".*
3. *C: "Penguins cannot fly".*
4. *Because $\neg(A \mathrel{|\!\sim} C)$, C can defeat B and we can infer C from $(A \cup C)$ without violating cautious monotony (whereas monotony dictates that we infer B from $(A \cup C)$).*

[6] For the sake of generality and conciseness, we do not define the structure of the propositions; the reader may consider them simply "elements".

A collection of methods for non-monotonic reasoning that has received considerable attention in the artificial intelligence community is *formal argumentation* [25]. The foundations of formal argumentation were laid by Dung in his seminal paper *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games* that formalizes the *abstract argumentation* method. Here, the central concept is the notion of an *argumentation framework*.

Definition 6 (Abstract Argumentation [13])
*An argumentation framework AF is a tuple $AF = (AT, AT)$, such that AR is a set of elements (*arguments*) and $AT \subseteq AR \times AR$ (*attacks*).*

In our work we assume that the set of arguments in an argumentation framework is finite. Let $AF = (AR, AT)$ be an argumentation framework and let $a, b \in AR$. We say that a attacks b iff $(a, b) \in AT$. For $S \subseteq AR$ we say that a attacks S iff $\exists c \in S$, such that $(a, c) \in AT$ and that S attacks a iff $\exists d \in S$, such that $(d, a) \in AT$; for $P \subseteq AR$, we say that S attacks P iff an argument in S attacks an argument in P; we denote $\{e | e \in AR, S \text{ attacks } e\}$ by S^+.

The focus of this thesis is on dynamic aspects of abstract argumentation; the concept of *argumentation framework expansions* plays an important role.

Definition 7 (Argumentation Framework Expansions [6])
Let $AF = (AR, AT)$ and $AF' = (AR', AT')$ be argumentation frameworks.

- *AF' is an* expansion *of AF, denoted by $AF \preceq AF'$, iff $AR \subseteq AR'$ and $AT \subseteq AT'$.*
- *AF' is a* normal expansion *of AF, denoted by $AF \preceq_N AF'$, iff $AF \preceq AF'$ and $(AT' \setminus AT) \cap (AR \times AR) = \{\}$.*

Let us introduce an example that illustrates the notions of expansions and normal expansions.

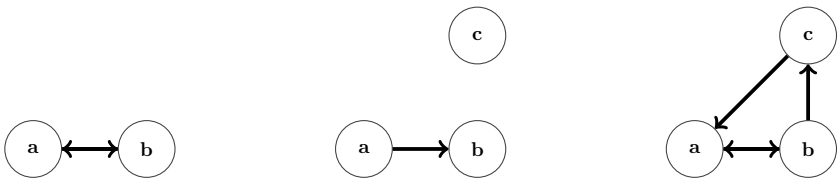
Example 4
Consider the following argumentation frameworks (see Figure 2.1):

- *$AF = (AR, AT) = (\{a, b\}, \{(a, b), (b, a)\})$.*
- *$AF' = (AR', AT') = (\{a, b, c\}, \{(a, b)\})$. AF' is not an expansion of AF' (and hence not a normal expansion of AF') because $AT \not\subseteq AT'$: $AF \not\preceq AF'$ and $AF \not\preceq_N AF'$.*
- *$AF'' = (AR'', AT'') = (\{a, b, c\}, \{(a, b), (b, a), (b, c), (c, a)\})$. AF'' is an expansion of AF ($AF \preceq AF''$) because $AR \subseteq AR''$ and $AT \subseteq AT''$ and also a normal expansion of AF ($AF \preceq_N AF''$) because $AF \preceq AF''$ and $(AT'' \setminus AT) \cap (AR \times AR) = \{\}$. AF'' is an expansion of AF' ($AF' \preceq AF''$) because $AR' \subseteq AR''$ and $AT' \subseteq AT''$ but not a normal expansion of AF' ($AF' \not\preceq_N AF''$) because $(AT'' \setminus AT') \cap (AR' \times AR') = \{(b, a)\}$.*

Important properties of a set of arguments in an argumentation framework are conflict-freeness, defense, and admissibility.

Definition 8 (Conflict-freeness, Defense and Admissibility [13])
Let $AF = (AR, AT)$ be an argumentation framework. A set $S \subseteq AR$:

- *is conflict-free (in AF) iff $\nexists a, b \in S$ such that a attacks b;*



(a) AF. (b) AF'. (c) AF''.

Figure 2.1: Example: argumentation framework expansions.

- *defends an argument $a \in AR$ (in AF) iff $\forall b \in AR$ such that b attacks a, $\exists c \in S$ such that c attacks b;*
- *is admissible (in AF) iff it is conflict-free and $\forall a \in S$, S defends a;*

Given an argumentation framework, an argumentation *semantics* infers a set of *extensions*, *i.e.* conclusions that entail a subset of the argumentation framework's arguments. The semantics that Dung defines in the seminal paper that introduces abstract argumentation are all admissible set-based, *i.e.* every extension must defend all of its arguments.

Definition 9 (Admissible Set-based Argumentation Semantics [13])
Let $AF = (AR, AT)$ be an argumentation framework. A set $S \subseteq AR$ is:

- *a stable extension of AF iff S is admissible and S attacks each argument that does not belong to S;*
- *a complete extension of AF iff S is admissible and every argument that is defended by S belongs to S;*
- *a preferred extension of AF iff S is a maximal admissible subset (w.r.t. set inclusion) of AR;*
- *a grounded extension of AF iff S is a minimal (w.r.t. set inclusion) complete extension.*

We denote the stable, complete, preferred, and grounded extensions of an argumentation framework AF by $\sigma_{stable}(AF)$, $\sigma_{complete}(AF)$, $\sigma_{preferred}(AF)$, and $\sigma_{grounded}(AF)$, respectively.

Another semantics *family* merely requires conflict-freeness and not admissibility.

Definition 10 (Naive and Stage Semantics [34])
Let $AF = (AR, AT)$ be an argumentation framework and let $S \subseteq AR$.

- *S is a naive extension of AF iff S is a maximal conflict-free subset of AR w.r.t. set inclusion. $\sigma_{naive}(AF)$ denotes all naive extensions of AF.*
- *S is a stage extension of AF iff S is conflict-free and $S \cup S^+$ is maximal w.r.t. set inclusion,* i.e. *$\nexists S' \subseteq AR$, such that S' is a conflict-free set and $S \cup S^+ \subset S' \cup S'^+$. $\sigma_{stage}(AF)$ denotes the stage extensions of AF.*

One way to define argumentation semantics (which may be admissible or naive set-based) is by recursing the strongly connected components of the argumentation framework (which can be seen as a directed graph). As a preliminary, let us introduce some basic graph theoretical definitions.

Definition 11 (Reachability and Strongly Connected Components)
Let $AF = (AR, AT)$ be an argumentation framework.

- *A path from an argument $a_0 \in AR$ to another argument $a_n \in AR$ is a sequence of arguments $P_{a_0,a_n} = \langle a_0, ..., a_n \rangle$, such that for $0 \leq i < n$, a_i attacks a_{i+1}.*
- *Given two arguments $a, b \in AR$, b is reachable from a iff there exists a path $P_{a,b}$ or $a = b$.*
- *$S \subseteq AR$ is a Strongly Connected Component (SCC) of AF iff $\forall a, b \in S$, a is reachable from b and b is reachable from a and $\nexists c \in AR \setminus S$, such that a is reachable from c and c is reachable from a. Let us denote the strongly connected components of AF by $SCCS(AF)$.*

In addition, we need to define the helper function UP (roughly: given an argumentation framework and two subset of its arguments, we remove all arguments from the first set that are attacked by an argument that is in the second set but not in the first set).

Definition 12 (UP [4])
Let $AF = (AR, AT)$ be an argumentation framework. Let $E \subseteq AR$ and let S be a strongly connected component of AF ($S \in SCCS(AF)$). We define $UP_{AF}(S, E) = \{a | a \in S, \nexists b \in E \setminus S$ such that $(b, a) \in AT\}$.

Now, we can introduce the SCC-recursive CF2 semantics, which makes use of naive semantics on SCC-level.

Definition 13 (CF2 Semantics [4])
Let $AF = (AR, AT)$ be an argumentation framework and let $E \subseteq AR$. E is a CF2 extension iff:

- *E is a naive extension of AF if $|SCCS(AF)| = 1$;*
- *$\forall S \in SCCS(AF)$, $(E \cap S)$ is a CF2 extension of $AF \downarrow_{UP_{AF}(S,E)}$, otherwise.*

$\sigma_{CF2}(AF)$ denotes all CF2 extensions of AF.

Intuitively, CF2 semantics recurses the acyclic directed graph of an argumentation framework's strongly connected components, starting by determining the naive extensions of the *top level* SCCs that are not attacked by any argument that is not part of the corresponding SCC. For each naive extension of these SCCs, the arguments in the extension determine which arguments on the next SCC level can be inferred, *i.e.* arguments that are attacked by the upper SCC's extension are removed, as they obviously have to be rejected.

In addition to the admissible set-based and naive set-based semantics families, *weakly admissible* semantics have been defined in a recent publication [7]. We do not formally work with this new semantics family (although we typically discuss it using simple examples). Hence, we abstain from introducing it as a formal preliminary here. Still, considering its significance for the abstract argumentation community, we give an intuition of its philosophy in comparison to the other two semantics families by introducing an example.

Example 5
Consider the argumentation framework $AF = (\{a, b, c, d\}, \{(b, c), (c, d), (d, a), (d, b)\})$. *In this argumentation framework, only the empty set defends itself against all its attackers, and consequently, all admissible set-based semantics can only yield the empty set as an extension. Intuitively, one could say that because in the cycle "b attacks c attacks d attacks b" (the bcd-cycle), each argument is indirectly contradicting itself, we do not know how to resolve the cycle, and hence discard all arguments from the cycle, as well as all arguments that are "downstream" of the cycle. Weakly admissible set-based semantics approach this problem with a different philosophy: colloquially speaking, given these semantics, the bcd-cycle is considered a self-contradiction, all of its arguments are rejected, and hence, we must infer* $\{a\}$ *as our only extension. In contrast, naive set-based semantics do not consider the bcd-cycle an indirect self-contradiction of each of its arguments, and hence (typically) infer that* $\{b, a\}$, $\{c, a\}$ *and* $\{d\}$ *are possible extensions.*

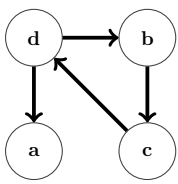


Figure 2.2: Example: differences between semantics families.

Let us introduce a trivial example that highlights the non-monotonic nature of abstract argumentation.

Example 6
Consider the argumentation frameworks $AF = (AR, AT) = (\{a\}, \{\})$ *and* $AF' = (AR', AT') = (\{a, b\}, \{(b, a)\})$. *Note that* AF' *is a normal expansion of* AF ($AF \preceq_N AF'$). *Given any of the argumentation semantics whose definitions we provide in this section – with the notable exception of naive semantics – the only extension we can infer from* AF *is* $\{a\}$ *and the only extension we can infer from* AF' *is* $\{b\}$. *Because* $\{b\}$ *is not a subset of (or equal to)* $\{a\}$, *we say that the behavior of the semantics is non-monotonic in this particular case.*

Formally, we can define different notions of monotony as properties of (or: *principles for*) abstract argumentation semantics as follows.

Definition 14
Let σ *be an argumentation semantics.* σ *satisfies weak monotony iff for every two argumentation frameworks* $AF = (AR, AT)$, $AF' = (AR', AT')$, *such that* $AF \preceq_N AF'$, *the following statement holds true:*

$$\forall E \in \sigma(AF), \exists E' \in \sigma(AF'), \text{ such that } E \subseteq E'$$

σ *satisfies strong monotony iff for every two argumentation frameworks* $AF = (AR, AT)$, $AF' = (AR', AT')$, *such that* $AF \preceq_N AF'$, *it holds that* $|\sigma(AF')| \geq 1$ *and the following statement holds true:*

$$\forall E \in \sigma(AF), \forall E' \in \sigma(AF'), E \subseteq E'$$

As Example 6 illustrates, most argumentation semantics satisfy neither weak nor strong monotony and hence are *non-monotonic*.

Finally, let us illustrate how we can build a bridge between microeconomic theory and abstract argumentation with the help of yet another example.

Example 7

We observe a decision-maker who considers to launch a new product. Initially, the decision-maker has two potential products she can launch: a' *and* b'*. The products are very similar; launching both of them is not feasible. Market research suggests that the target consumer group would typically rather buy* a' *than* b'*. Our decision-maker denotes this by modeling the arguments* a *and* b *(*launch a' *and* launch b'*), as well as an attack from* a *to* b*. Consequently, she ends up with the argumentation framework* $AF = (\{a, b\}, \{(a, b)\})$*. In this scenario, our decision-maker has established clear preferences over the product launch options (and also over the* sets of options *in* $2^{\{a,b\}}$*). Let us assume she resolves the argumentation framework using preferred semantics:* $\sigma_{preferred}(AF) = \{\{a\}\}$ *and decides to launch* a'*. However, she needs to wait for the company's board of directors to approve the decision. While she is waiting for approval, the R&D department is evaluating another product prototype (*c'*), and the board requests that the decision-maker consider this one, too. Market research suggests that the target consumer group would typically rather purchase* b' *instead of* c'*, but* c' *instead of* a'*;* i.e.*, our decision-maker constructs the argumentation framework* $AF' = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\})$*, which she resolves as* $\sigma_{preferred}(AF') = \{\{\}\}$*, which means she does not intend to launch any product. Now, it is obvious that if she relays this decision to the board, her reasoning abilities will probably be questioned. How can it be that the presence of an additional launch option triggers the decision-maker to change her mind and decide to not choose* any *asset?*[7] *Indeed, this common-sense fallacy is reflected in the formal model of economic rational man: choosing* $\{b\}$ *from* $2^{\{a,b\}}$ *implies the preference* $\{b\} \succeq \{\}$*, which is inconsistent with the preference* $\{\} \succeq \{b\}$ *as implied by choosing* $\{\}$ *from* $2^{\{a,b,c\}}$*.*

In Paper II we formally analyze this apparent, yet previously unexplored intersection of abstract argumentation and formal models of economic rationality. Let us note that argumentation semantics exist that determine the extensions $\{a\}$, $\{b\}$, and $\{c\}$ for the argumentation framework AF' in the example. However, as we show in Paper II, these semantics do not necessarily satisfy principles of rational decision-making in other scenarios.

[7] A more reasonable course of action would be to delay the decision; however, in this scenario we assume the decision-maker should be able to make rational decision at any point.

Chapter 3

Discussion of the Included Papers in the Context of the Thesis

This section provides a discussion that embeds the included papers into the holistic context of the thesis.

3.1 Paper I

The main objective of this thesis is to explore and advance the research frontier of principle-based approaches to reasoning and decision-making at the intersection of human and machine reasoning. Paper I contributes to this objective from a bird's eye view, by introducing four levels of intelligent and explainable reasoning in face of contradictions that are motivated from a microeconomics and behavioral economics perspective and then illustrated using abstract argumentation-based examples. These levels can be summarized as follows:

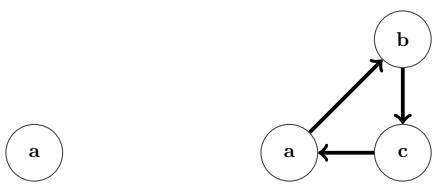
1. Decisiveness in face of contradictions (*clear preferences* in microeconomic theory);
2. Consistent inference in face of contradictions (*consistent preferences* in microeconomic theory);
3. Explanation of inference given principles that happen to be satisfied (*reasoning backwards* in behavioral economics);
4. Principle-based and evidence-based reasoning, including the ability of a reasoner to revise the principles according to which it reasons (no microeconomics/behavioral economics equivalent).

Along these four levels, a research roadmap is established, which Papers II-VI then work towards[1]. Because this paper is a position paper, it does not present novel formal results.

[1] Because Paper I was published after some of the other papers, it already relates to some of the results that are presented later in this thesis.

3.2 Paper II

As a first step towards the vision outlined by Paper I, Paper II builds a bridge between the *rational man* paradigm of economic theory and abstract argumentation, but also introduces relaxed notions of monotony of entailment to abstract argumentation. To give a better intuition of the paper's results, let us illustrate the key principle – reference independence – using an example. Consider the argumentation frameworks $AF = (\{a\})$ and $AF' = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\})$ (see Figure 3.1). Note that AF' is a normal expansion of AF, *i.e.* it contains additional arguments and attacks, but the attacks between existing arguments remain the same. Typically, an argumentation semantics yields the extension $\{a\}$, given AF. We assume that this inference establishes the preference "$\{a\}$ is preferred over all other sets we could have potentially inferred from AF", *i.e.* $\{a\}$ is preferred over $\{\}$. Given AF', an argumentation semantics typically yields either the empty set ($\{\}$) or the three extensions $\{a\}$, $\{b\}$, and $\{c\}$. To satisfy reference independence, we must be able to find – for every argumentation framework, every of its extensions, and every normal expansion – an extension of the normal expansion that establishes preferences that are *consistent* with the preferences established by the inference of the extension from initial argumentation framework. In our example, semantics that infer $\{\}$ from AF' cannot establish consistent preferences: "$\{\}$ is preferred over $\{a\}$" is inconsistent with "$\{a\}$ is preferred over $\{\}$". In contrast, if we infer any of $\{a\}$, $\{b\}$, $\{c\}$ from AF', the preferences these inferences establish do not contradict the previously established preference.



(a) AF. (b) AF'.

Figure 3.1: Example: reference independence.

We show by formal analysis that most argumentation semantics, with the notable exception of CF2 semantics, do not satisfy the weak reference independence principle.

Cautious monotony – another key principle that this paper introduces to abstract argumentation – stipulates that an extension that has been inferred from an argumentation framework may only be rejected after a normal expansion if the expansion adds a new argument that attacks the previously inferred extension; an analysis of argumentation semantics w.r.t. their satisfaction of this principle is provided by Paper III. Beyond this, Paper II lays the foundations for follow-up works that are presented by Papers III-VI[2] at the intersection of non-monotonic reasoning, formal argumentation, and formal models of bounded rationality.

[2] While Paper VI does not directly depend on the results presented by Paper II, it relies on a variant of the notion of *maximally monotonic extensions* as introduced by Paper IV, which in turn relies on Paper II.

3.3 Paper III

Paper III builds on the results of Paper II and further examines how the most relevant introduced consistency principles (reference independence and cautious monotony) can be ensured in dynamic environments, in which inferences are drawn from an initial argumentation framework, which is normally expanded to then draw inferences from the expansion and so on. To ensure one of the consistency principles is practically satisfied[3], several formal approaches are introduced:

Reductionist approach. Given an argumentation framework, one of its extensions, and one of its normal expansions, we remove arguments (as few as possible) from the expansion until we can infer exactly one extension and this extension satisfies the consistency principle w.r.t. the initial argumentation framework, the extension that has been inferred from it, and the normal expansion.

Expansionist approach. Given an argumentation framework, one of its extensions, and one of its normal expansions, we add *annihilator arguments* and attacks that originate from these arguments (as few as possible) to the expansion until we can infer exactly one extension and this extension satisfies – not considering the annihilator arguments – the consistency principle w.r.t. the initial argumentation framework, the extension that has been inferred from it, and the normal expansion.

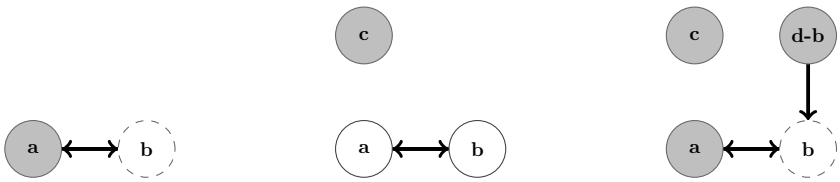
Extension-selecting approach. If we apply an argumentation semantics that guarantees – given an argumentation framework, one of its extensions, and one of its normal expansions – that we can select at least one extension of the normal expansion that satisfies the consistency principle w.r.t. the initial argumentation framework and the extension we have inferred from it, we can use the expansionist approach to select this (or: *one of these*) extensions, given some other straightforward formal argumentation principles are satisfied as well.

The reductionist and expansionist approaches have the advantage that they can guarantee reference independence or cautious monotony for almost any argumentation semantics, given merely very simple constraints on semantics behavior that are typically satisfied. However, these approaches may change (reduce or expand) the argumentation frameworks in ways that, roughly speaking, lead to counter-intuitive semantics behavior that is not aligned with other principles the applied semantics would otherwise satisfy. This is not the case when using the extension-selecting approach, which requires, however, the satisfaction of principles that many semantics do not satisfy.

Let us introduce an example (see Figure 3.2) that illustrates the extension-selecting approach to ensure cautious monotony (a principle that Paper II introduces to abstract argumentation). Let us start with the argumentation framework $AF = (\{a, b\}, \{(a, b), (b, a)\})$; assume we have already selected the preferred extension $\{a\}$ from it. After normally expanding AF to $AF' = (\{a, b, c\}, \{(a, b), (b, a)\})$, we want to select an extensions that entails $\{a\}$ (because no new argument attacks $\{a\}$ and we want to satisfy the cautious monotony condition, which stipulates that monotony of entailment must not be violated unless a normal expansion adds new attacks to a previously inferred extension). Because there are two preferred extensions

[3] Let us note that cautious monotony and reference independence do not complement each other, but can be considered alternatives. None of the "well-established" argumentation semantics, with the exception of the simplistic and impractical naive semantics, satisfies both.

of AF' ($\{a, c\}$ and $\{b, c\}$), we add the annihilator argument `d-b` that attacks b to AF', *i.e.* we normally expand AF' to $AF'' = (\{a, b, c, \texttt{d-b}\}, \{(a, b), (b, a), (\texttt{d-b}, b)\})$, from which we then infer $\{a, c, \texttt{d-b}\} \setminus \{\texttt{d-b}\} = \{a, c\}$.

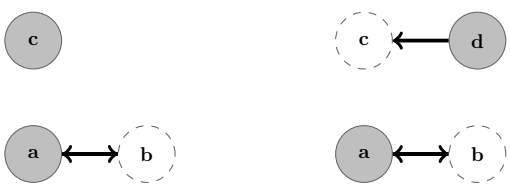


(a) AF. (b) AF'. (c) AF''.

Figure 3.2: Example: ensuring weak cautious monotony. Here and henceforth, arguments with a gray background have necessarily been inferred (are in all extensions a specific semantics yields for a given argumentation framework); arguments with a dashed border have been rejected (are in none of the extensions) and the remaining arguments (solid border, white background) are in at least one, but not in all extensions.

3.4 Paper IV

In Papers II and III we have introduced new abstract argumentation principles that are based on the notions of *consistent preferences* in microeconomic theory and *cautious monotony* in non-monotonic logics. However, these principles are Boolean conditions (*i.e.* they are either satisfied or violated) and can be complemented by dynamic approaches. In this paper, we introduce such an approach, which allows us to be *as monotonic as possible* – given the constraints of an argumentation semantics – when iteratively drawing inferences from an argumentation framework and normally expanding it. In particular, we introduce the notion of *maximal monotonic extensions* (w.r.t. cardinality), which we can motivate and explain using an example. Consider the argumentation framework $AF = (AR, AT) = (\{a, b, c\}, \{(a, b), (b, a)\})$. Assume we have inferred $\{c, a\}$ from it, using preferred semantics (for instance; we assume we have selected one of the preferred extensions of AF). We (normally) expand AF to $AF' = (AR', AT') = (\{a, b, c, d\}, \{(a, b), (b, a), (d, c)\})$. This allows us to infer the following preferred extensions: $\{a, d\}$ and $\{b, d\}$. Intuitively, if we want to remain somewhat consistent (colloquially speaking), we want to infer $\{a, d\}$ and not $\{b, d\}$ (Figure 3.3). However, inferring $\{b, d\}$ is still aligned with the notions of reference independence and cautious monotony: because we infer the new argument d as part of our extension, the preferences we establish on $2^{AR'}$ are consistent with the preferences we have established (on 2^{AR}) when inferring $\{a, c\}$ from AF; also, d as a new attacker of our extension $\{a, c\}$ justifies the violation of monotony (from a cautious monotony perspective). To actually infer $\{a, d\}$, Paper IV introduces the notion of *maximally monotonic extensions*: given $\{a, c\}$, we select the extensions from the set $\{\{a, d\}, \{b, d\}\}$ that are maximally monotonic (w.r.t. cardinality), *i.e.* that include as many of the previously inferred arguments as possible. In our example, this approach makes us select $\{a, d\}$.



(a) AF. (b) AF'.

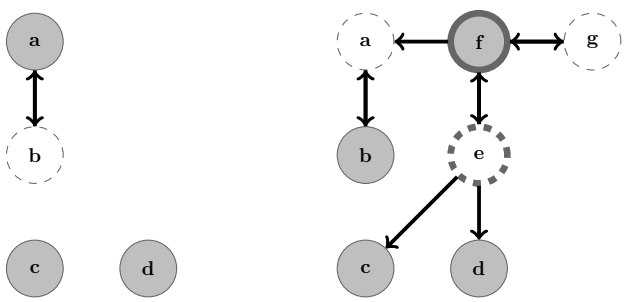
Figure 3.3: Example: maximizing monotony w.r.t. a previous inference, considering the constraints of a given argumentation semantics.

However, in a sequence of normally expanding argumentation frameworks, selecting a maximally monotonic extension w.r.t. an extension inferred from the immediate predecessor may force us to select an extension that is not maximally monotonic w.r.t. an "earlier" predecessor (we can say that monotony maximization is not transitive). In the paper, we formalize this observation as the *degrees of monotony*-dilemma, and provide a mitigation approach.

3.5 Paper V

In Paper V, we re-use and extend some of the formal concepts that we have introduced in Papers II and IV, in particular *weak cautious monotony* and *maximally monotonic extensions*, and apply them to the problem of machine reasoning explainability. In particular, we introduce a formal approach that allows us, given an argumentation framework and an extension that has been inferred from it, to generate an explanation that tells us why a subsequent inference from a normal expansion of this argumentation framework – which we draw by selecting an extension that is maximally monotonic w.r.t. the initially inferred extension – violates monotony (if it does). Such an explanation is a subset of the arguments that have been added by the normal expansion and that satisfies specific properties (which we explain below, using an example). We can guarantee for complete, preferred, and grounded semantics that if monotony is violated, such an explanation always exists (the set of arguments that satisfy our explanation criteria is not empty), and that if monotony is not violated, no explanation exists (the explanation set is empty).

Let us introduce an example that gives an intuition of the approach (see Figure 3.4). We start with the argumentation framework $AF = (\{a, b, c, d\}, \{(a, b), (b, a)\})$ and infer $\{a, c, d\}$ from it, for example by selecting one of the preferred extensions. Then, we normally expand AF to $AF' = (\{a, b, c, d, e, f, g\}, \{(a, b), (b, a), (e, c), (e, d), (e, f), (f, a), (f, e), (f, g), (g, f)\})$. Again assuming preferred semantics, we can infer either $\{a, e, g\}$, or $\{b, e, g\}$, or $\{b, c, d, f\}$. Pragmatically, we want to reject as few of the previously inferred arguments as possible, and hence opt to infer $\{b, c, d, f\}$. To explain why our inference result no longer entails a, we can use the arguments e (e attacks the previously inferred arguments c and d and hence needs to be defeated) and f (f successfully attacks e, but also a, *i.e.* we are only able to keep inferring c and d if we allow for the attack from f to a to succeed).



(a) AF. (b) AF'.

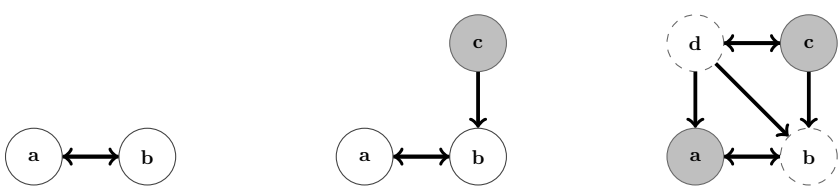
Figure 3.4: Example: explaining the violation of monotony. A bold border indicates that an argument is part of an explanation set.

3.6 Paper VI

In addition to Paper V, Paper VI illustrates how some of the formal results that this thesis introduces help solve problems at the intersection of formal theory and a specific application domain. In this paper, the application domain is legal reasoning, and the specific issue the paper addresses is the notion of the *burden of persuasion*, which is a prominent concept in legal reasoning, but also more broadly in defeasible reasoning in general. In some legal cases (but also in other domains, such as political persuasion scenarios), the burden of persuasion rests on certain arguments, for example on the claim that a defendant is guilty[4]; hence, if in doubt – and only then – arguments that carry the burden of persuasion are to be rejected (*e.g.* following the legal notion of *in dubio pro reo*).

The paper provides a generally applicable, argumentation-based model of burdens of persuasion of arbitrarily many levels, and addresses some open questions that a recently published paper on the topic has raised. In particular, we introduce the notions of an argumentation-based *burden of persuasion*-framework and of burden of persuasion-semantics (which extend the notions of an abstract argumentation framework and an abstract argumentation semantics, respectively). As a prerequisite, we take the notion of maximal monotonic extensions (w.r.t. cardinality) as introduced by Paper IV and adjust it to get pareto-optimal $\subseteq$-maximal monotonic extensions.

Let us introduce an abstract example that illustrates the application of this notion to the burden of persuasion. Consider the argumentation framework $AF_2 = (\{a, b, c, d\}, \{(a, b), (b, a), (c, b), (c, d), (d, a), (d, b), (d, c)\})$ and assume that a and b are unburdened, c carries a "light-weight" burden, and d carries a "heavy" burden (see Figure 3.5). Given, for instance, preferred semantics, we can infer the following extensions from AF_2: $\{a, c\}$ or $\{d\}$. When considering the two unburdened arguments – and the inferences we can draw from the restriction of AF_2 to these arguments (See Figure 3.5 (a) – it is clear that we have to reject b, but we are able to accept a: *i.e.*, the only pareto-optimal $\subseteq$-maximal monotonic extension of AF_2 w.r.t. $\{a\}$ and $\{b\}$ is $\{a, c\}$. Note that pareto-optimality is important, because it allows us to discard the extension $\{d\}$, which is $\subseteq$-maximal monotonic w.r.t. $\{b\}$, but not pareto-optimal, because $\{a, d\}$ is "just as good" for $\{b\}$ but "better" for $\{a\}$. Next, we could potentially consider the "light-weight"-burdened argument c (see Figure 3.5 (b)), but as we have already narrowed down the extensions we can infer from AF_2 to one, this step is not necessary in this particular scenario.



(a) AF_0. (b) AF_1. (c) AF_2.

Figure 3.5: Example: the burden of persuasion in abstract argumentation.

[4] Note that the model can be extended to support multiple *levels of burden*.

3.7 Paper VII

Papers I-VI focus on formal aspects of argumentation (and in particular, of abstract argumentation). Even Papers V and VI, which highlight the application potential of the general research direction (and of some of the results) in this thesis, are, first and foremost, formally oriented in that they provide theoretical results that are of relevance to the application domains of machine reasoning explainabilitiy and legal reasoning, respectively. Purely practical aspects of applying formal argumentation as a method of non-monotonic reasoning to real-world systems are merely hinted at.

In contrast, Paper VII focuses on practical software design and engineering aspects, by introducing a design methodology for argumentation-based health information systems. From a process perspective, the design methodology is well-aligned with common industry practices for the iterative development of software systems; for each of a range of common iteration steps, design choices and procedures that are particular to the development of argumentation-based systems are discussed.

Let us highlight that while the methodology addresses some issues that are of particular relevance to the health domain (like patient empowerment), most aspects are applicable to the design and development of argumentation-based systems in general. With this practical perspective, Paper VII wraps up the thesis by providing an engineering-oriented framework of how to make use of the research results of the formal argumentation community in general, and of this thesis in particular. In this context, it is worth noting that the formal abstractions that this thesis introduces have been made available as a software library [19], and hence can be conveniently re-used, for example when implementing prototypes for applied information systems research.

Chapter 4

Future Work and Concluding Remarks

Each of the papers that are part of this thesis points to opportunities for future research. Typically, these research directions concern detailed aspects that remain open questions in the specific context of the results that the corresponding paper presents. In this chapter, we approach future work from a conceptual, less technical, and arguably *bolder* perspective, which is somewhat similar (but not identical) to the positions we present in Paper I. Below, we present three broad research directions, before we conclude with a broader outlook.

4.1 Formal Argumentation as a Fundamentally Dynamic Process

In ongoing research [15], we work towards a formal approach to (abstract) argumentation that treats dynamics as a first-class citizen. We assume that we start with an empty argumentation framework, which we then expand one argument at a time. We also start with an initially empty set of conclusions (an empty extension). After each expansion, we check how we need to adjust our set of conclusions:

- Can we simply add the new argument to our extension?
- Or can we simply reject the argument?
- Should we now reject arguments we have inferred before?
- Should we now infer arguments we have rejected before?

Answering these questions does not necessarily require an argumentation semantics. However, if we *can* answer these questions, we can generalize the approach so that we gain an argumentation semantics: we simply relax the constraints that we have a total order on the set of arguments (which we have when we normally expand an argumentation framework argument-by-argument), and consider all possible total orders (permutation sequences) we can establish on a given set of arguments. We speculate that this new way to approach the domain of formal argumentation may lead to intriguing theoretical insights.

4.2 Empirically Integrating Principle-based Automated Reasoning and Human Reasoning

Historically, one can consider the line of research on the systematic relaxation of monotony in automated reasoning as one that runs roughly parallel to the line of research in microeconomic theory and behavioral economics that is concerned with the design of formal models of bounded rationality, *i.e.* of models that systematically relax the principles of economic rationality. However, while models of bounded rationality – for example Tversky's and Kahneman's prospect theory [18] – are often based on experimental evidence, when designing methods and principles of automated "common sense" reasoning, researchers typically rely on their own intuitions of what can be considered reasonable and what cannot. In the field of formal argumentation, some recent works have started to rely on empirical studies of human reasoning, in particular to assess whether semantics behavior is aligned with what human study subjects consider reasonable inference [11, 10]. Still, this line of research covers merely a small fraction of the assumption that formal argumentation approaches (and, more broadly, methods of non-monotonic reasoning) make. Also, in the context of this line of research, computer science-based approaches that are often employed in practice to match human (user) and computing system behavior seem to be underutilized. For example, one could apply reinforcement learning and recommender systems approaches to dynamically determine which consistent subset of a set of principles of an automated reasoning approach is considered the most intuitive or useful – either in general or considering a specific context or application domain.

4.3 Learning Principles of Reasoning and Decision-Making

As a next step after implementing the vision of dynamic principle selection that can potentially be enabled by reinforcement learning and recommender systems approaches (as discussed in the previous subsection), one could (semi)-automatically design entirely new principles. For example, given an approximate specification of what a principle ought to achieve, one could use machine learning approaches to search for principles that seem to roughly satisfy the constraints, and then automatically or semi-automatically formally verify the satisfaction or violation of these constraints. For the (partial) automation of the verification procedure tools like SAT solvers and interactive theorem provers can be used. Ideally, the approximate specifications can be evolved automatically as well, for example by using reinforcement learning approaches that tie specification parameters to the Key Performance Indicators (KPIs) of software systems that operate in a specific organizational (business) context.

4.4 Concluding Remarks

The works presented in this thesis have introduced a new perspective on drawing and analyzing non-monotonic inferences, integrating formal theories of economic decision-making (traditionally: "for humans", considering the title of the thesis) and automated, argumentation-based reasoning ("for machines"). The focus of the presented formal approaches is to remain, in some sense, *consistent* when repeatedly drawing inferences from an *expanding* – *i.e.*, growing – knowledge or belief base. The results formalize and formally analyze the intuition that inferences we draw and decisions we make at a given point are typically informed, and hence constrained, by previous inferences and decisions, and that in practice, it is more pragmatic to stay course than to "zick-zack around", if no compelling evidence to revise previous inferences is presented. This intuition can be explored from a range of other viewpoints, both formally and practically, considering the future research potential that has been outlined in this chapter. Still, the underlying philosophy of the presented formal approaches can potentially be applied beyond the outlined research directions to any field or sub-field of any discipline in which the study of reasoning and decision-making is of conceptual, formal, or practical relevance.

References

[1] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] L. Amgoud and H. Prade. Using arguments for making and explaining decisions. *Artificial Intelligence*, 173(3):413–436, 2009.

[3] P. Baroni, M. Caminada, and M. Giacomin. Abstract argumentation frameworks and their semantics. In P. Baroni, D. Gabbay, G. Massimiliano, and L. van der Torre, editors, *Handbook of Formal Argumentation*, chapter 4, pages 159–236. College Publications, 2018.

[4] P. Baroni, M. Giacomin, and G. Guida. Scc-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, 168(1):162 – 210, 2005.

[5] P. Baroni, A. Rago, and F. Toni. From Fine-Grained Properties to Broad Principles for Gradual Argumentation: A Principled Spectrum. *International Journal of Approximate Reasoning*, 105:252–286, feb 2019.

[6] R. Baumann and G. Brewka. Expanding argumentation frameworks: Enforcing and monotonicity results. *COMMA*, 10:75–86, 2010.

[7] R. Baumann, G. Brewka, and M. Ulbricht. Revisiting the foundations of abstract argumentation-semantics based on weak admissibility and weak defense. In *AAAI*, pages 2742–2749, 2020.

[8] K. L. Clark. *Negation as Failure*, pages 293–322. Springer US, Boston, MA, 1978.

[9] P. Collier. *The future of capitalism: Facing the new anxieties*. Harper New York, 2018.

[10] M. Cramer and M. Guillaume. Empirical cognitive study on abstract argumentation semantics. *Frontiers in Artificial Intelligence and Applications*, 2018.

[11] M. Cramer and M. Guillaume. Empirical study on human evaluation of complex argumentation frameworks. In F. Calimeri, N. Leone, and M. Manna, editors, *Logics in Artificial Intelligence*, pages 102–115, Cham, 2019. Springer International Publishing.

[12] K. Čyras and F. Toni. Non-monotonic inference properties for assumption-based argumentation. In E. Black, S. Modgil, and N. Oren, editors, *Theory and Applications of Formal Argumentation*, pages 92–111, Cham, 2015. Springer International Publishing.

[13] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[14] D. M. Gabbay. Theoretical foundations for non-monotonic reasoning in expert systems. In K. R. Apt, editor, *Logics and Models of Concurrent Systems*, pages 439–457, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg.

[15] Gabbay, Dov and Kampik, Timotheus. A brief introduction to the shkop approach to conflict resolution in formal argumentation. In B. Liao, J. Luo, and L. Van der Torre, editors, *Logics for New-Generation AI 2021*, pages 46–62, London, 2021. College Publications.

[16] E. Hadoux, A. Hunter, and S. Polberg. Biparty decision theory for dialogical argumentation. In *COMMA*, pages 233–240, 2018.

[17] D. Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475, 2003.

[18] D. Kahneman and A. Tversky. *Prospect Theory: An Analysis of Decision Under Risk*, chapter Chapter 6, pages 99–127. WSPC, 2013.

[19] T. Kampik and D. Gabbay. Towards diarg: An argumentation-based dialogue reasoning engine. In *SAFA@ COMMA*, pages 14–21, 2020.

[20] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1):167–207, 1990.

[21] J. Moschovakis. Intuitionistic Logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[22] R. Murawski. Undefinability of truth. the problem of priority: Tarski vs Gödel. *History and Philosophy of Logic*, 19(3):153–160, 1998.

[23] R. Murawski and A. Mickiewicz. John von Neumann and Hilbert's school of foundations of mathematics. *Studies in Logic, Grammar and Rhetoric*, 7(20):37–55, 2004.

[24] M. J. Osborne and A. Rubinstein. *Models in Microeconomic Theory*. Open Book Publishers, 2020.

[25] I. Rahwan and G. R. Simari. *Argumentation in Artificial Intelligence*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[26] A. S. Rao and M. P. Georgeff. Modeling rational agents within a BDI-architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, KR'91, page 473–484, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[27] A. Rubinstein. *Modeling bounded rationality*. MIT press, 1998.

[28] S. J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[29] H. A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 02 1955.

[30] M. Strathern. 'improving ratings': audit in the british university system. *European Review*, 5(3):305–321, 1997.

[31] A. Tarski. The concept of truth in formalized languages. In A. Tarski, editor, *Logic, Semantics, Metamathematics*, pages 152–278. Oxford University Press, 1936.

[32] T. Uebel. Vienna Circle. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[33] L. van der Torre and S. Vesic. The principle-based approach to abstract argumentation semantics. *IfCoLog Journal of Logics and Their Applications*, 4(8), October 2017.

[34] B. Verheij. Two approaches to dialectical argumentation: admissible sets and argumentation stages. *Proc. NAIC*, 96:357–368, 1996.

[35] J. Von Neumann, O. Morgenstern, and H. W. Kuhn. *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007.