This is a section of doi:10.7551/mitpress/11252.001.0001

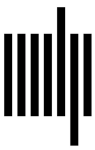
The Handbook of Rationality

Edited by: Markus Knauff, Wolfgang Spohn

Citation:

The Handbook of Rationality
Edited by: Markus Knauff, Wolfgang Spohn
DOI: 10.7551/mitpress/11252.001.0001
ISBN (electronic): 9780262366175
Publisher: The MIT Press
Published: 2021

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

2.6 Logico-Computational Aspects of Rationality

Johan van Benthem, Fenrong Liu, and Sonja Smets

Summary

Taking a broad historical line, we discuss major aspects of rationality that can be analyzed in a logical and computational perspective. Topics include classical notions from the foundations of computability, insights from the development of computer science and AI, and the richer picture of rationality emerging in current logical studies of agency. We also discuss two challenges: distributed computing and evolutionary games replace "high rationality" by "low rationality" in behavior; machine learning defies classical views of representation and inference. Both trigger new logical themes in the study of rationality.

1. Logic and Rationality

Rational behavior is a rich phenomenon, not captured in a single formula, but mapped out in this entire handbook. Let us take the commonsense view. We all believe, or try to believe, that our behavior is driven by reasons and reasoning and that we are susceptible to reason, changing our minds when confronted with new facts or considerations. This is how we see ourselves, how we justify our actions to others, and how academic organizations present themselves to a general public. Further important aspects of rationality, such as preferences and goals, will only be touched upon lightly in this chapter.

Since Antiquity, reasoning has been at the heart of logic. In fact, logic is often seen as rationality in its purest, perhaps its most intimidating, form. We will chart this match, without claiming logic is all there is to rationality.

Example: Valid and invalid consequence. Classical logic tells us things like this: an inference to $\neg B$ from $A \to B$ and $\neg A$ is invalid (B might hold for other reasons than A)—but the inference from $A \to B$ and $\neg B$ to $\neg A$ is valid and in fact the engine of refutation. Valid inferences, put together, form complex proofs that can yield surprising new insights. Over time, studying all this has produced a rich semantic and syntactic discipline of logical systems. ⊣

Before we start, here is a distinction. Logical proof can be seen as a practical engine of rational behavior, and different logical systems can model rational reasoning practices. But there is also a theoretical foundational use of logic, as a study of the structure of rationality: its laws and its limitations. In this second sense, logical analysis can target any practice that rational agents engage in: not just reasoning but also observing, taking decisions, or debating. Both uses, practical and theoretical, will occur in what follows.

But here is a further step. Logical systems for analyzing reasoning are cultural artifacts that interact with human practices. In particular, they feed into computational devices. For instance, the valid inference from $A \to B$ and $\neg B$ to $\neg A$ is also a basic law of binary arithmetic, at work in one's computer. This association has proved fruitful in the foundations of mathematics, and practically, it has triggered the development of computers, information technology, and artificial intelligence, which are transforming our world. Some even fear that our logical tools have started overtaking us.

A full picture of reasoning requires that "us." Many themes in this chapter connect naturally to empirical cognitive psychology. This interface is beyond our scope, and we refer the reader to the empirical entries in this handbook. Instead, we proceed to the logico-computational perspective on rationality.

2. Mini-History of Reasoning and Computation

To understand the links between logic and rationality, a historical perspective is helpful (cf. Kneale & Kneale, 1962). Over time, many forms of reasoning were captured in logical systems by philosophers, mathematicians, and others—a process that is still continuing. Once discovered, these systems became intellectual tools that enhanced rational thinking. Then, in the early modern age, Llull and Leibniz realized that reasoning is close to

computation. From there runs a straight road to the logic machines of Babbage and Lovelace and onward to modern computers and artificial intelligence (AI) systems. On the way, the notion of computation acquired sharper contours. A computing device need not be tied to one specific task; it can be programmable, the way a loom can weave different textiles depending on its book. Thus, computing means finding algorithms performing tasks, and since algorithms work on code, it also means finding data structures that represent information in appropriate ways.

All this is similar to the reality of logic itself. Textbooks say that logic is the study of inference or reasoning, but much more is involved. Reasoning presupposes a vehicle, often a language, representing the notions one reasons with. Thus, as noted in the perceptive essay (Beth 1971), logic has always been about a tandem of proof and definition or, if you wish, proof theory and model theory. Moreover, Beth added as a crucial third historical constant of logical thought the notion of algorithm, which combines the former two.

The history of computing is remarkable in that major principles were discovered before practical success, unlike in many other fields. Gödel (1931/1986) analyzed the limits of what logical proof systems can achieve, finding that systems whose expressive power suffices for encoding basic arithmetic are either inconsistent or incomplete: unable to prove all intuitive mathematical truths about their domain. Gödel's proof involved a deep analysis of computable ("recursive") functions, and subsequently, Turing (1936) defined machines that can compute all recursive functions. There is even a universal Turing machine that, given any program code and input, computes the effects of running that program on that input. In this setting, Church (1936) showed that standard reasoning systems such as first-order logic, although axiomatizable, are undecidable: no computing method can decide for arbitrary first-order consequence problems whether they are valid. Thus, a major trade-off came into view: increased expressivity of a language and complexity of its decision problem for validity are at odds.

This history highlights several points that seem crucial to understanding the nature and scope of rationality even today. The first is a practical issue of the modus operandi: if rationality has a computational engine, how should we understand its interplay of reasoning, information, and concept formation? The second point is theoretical: are there principled limitations to logical rationality—say, are some natural tasks beyond the scope of rational inquiry? Gödel's theorems keep generating discussion (Wang, 1996), and a common moral is that there is more to rational thinking than what is captured in logical systems. Even so, whenever this "more" is explained systematically, the limitation theorems apply again. Finally, results on what proof systems and computing devices cannot do were in fact immensely helpful in the further development of systems of inference and computation that can do a lot. Likewise, the modern "challenges to rationality" discussed in this chapter may actually yield new insights into what rationality can achieve. Having said that, much current literature in AI or cognitive psychology is of the "can do" type: one seldom reads about exciting discoveries of deep new limitations.

The foundational era of Gödel and Turing showed what is provable or computable in principle. While this high abstraction level remains a valid perspective on rationality, subsequent history tells us many further things.

3. Computer Science and Artificial Intelligence

3.1 Computer Science

The development of computing in the 20th century has generated major practical achievements but also an ever-growing insight into fine structure. There are different models for computation, from Turing machines to more recent formats for computing, and crucially, these models come in hierarchies. Some tasks are solved by simple finite automata; others require memory management to varying degrees. Likewise, there is a wide diversity of, poorer or richer, languages for specifying data structures and writing programs (Harel, 1987). Next, significantly, around 1980, computing architecture moved away from single Turing machines to distributed networks, the reality of computing today (Andrews, 2000). These developments have given rise to new fields such as automata theory (Chakraborty, Saxena, & Katti, 2011), complexity theory (Papadimitriou, 1994), and process algebra (Bergstra, Ponse, & Smolka, 2001), which chart the varieties of computation in different ways, many of them connected to logic. This process is still ongoing, and the foundations of computation remain under debate. For instance, there is no consensus on a definitive notion of algorithm, a more intensional notion than an extensional input–output record of Turing machines (Bonizzoni, Brattka, & Löwe, 2013; Haugeland, 1997). Computation today is rather a way of producing behavior.

All this comes with notions and insights that are relevant to rationality. The fine structure of computation gives a precise sense to the earlier double-edged modus operandi: reasoning engine and representational apparatus. And the variety in real computation suggests that

the norm in rational performance is not one idealized super-device but "boundedness" of resources and powers. Matching this practical concern is a fundamental issue. In the complexity theory of space and time resources needed for computational tasks, we are really talking about the information in the world and how to process it. But this forces us to reflect on what is the information available to rational agents (Adriaans & van Benthem, 2008). And there is yet more to be learnt from the world of computing. If we think of the behavior of rational agents as performing many tasks at once, just as networks of computers do, then there is a fundamental issue of architecture. How do the different components of the overall system pass information and cooperate (Gabbay, 1999)?

3.2 Artificial Intelligence

Moving closer to humans, computer science blends seamlessly into AI. From the start, computers have been seen as a powerful model for human intelligence. In an interesting departure from the detailed internal analysis of computing by Turing machines, the famous Turing Test approaches intelligence in the tradition of measuring theoretical notions by external observable behavior (Turing, 1950). It proposes that a computer achieves intelligence if an observer using natural language cannot tell that computer apart from a human by asking questions and engaging in conversation. Over the years, computers started to pass variants of the Turing Test, or other types of intelligent behavior. Actually, none of these are usually considered conclusive, as the criteria are a moving target (van Harmelen, Lifschitz, & Porter, 2008). Passing the test is dismissed as not a display of "real intelligence," and then the demands are shifted a little further. But behavioral tests are crucial to judging human rationality as well. We seldom look inside people's heads to monitor their considerations but rather observe their words and actions.

A final intriguing feature of the Turing Test is its hybrid scenario where different types of agents, humans and machines, interact, presaging the reality of human–machine interactions in modern society. This scenario goes beyond classical emulation or competition concerns. How can societies of mixed agents, with different strengths and weaknesses, interact successfully (Wooldridge, 2002/2009)? The resulting diversity in agency is only beginning to be acknowledged more widely. Most paradigms in logic or philosophy assume that agents have similar abilities for reasoning, observing, and communicating, although their information and preferences may differ. Such uniformity assumptions underlie generic notions like "human" or "rational actor," and they may even seem to embody moral imperatives, like treating everyone equally *qua* rights and duties. If one accepts diversity, however, notions and theories concerning rationality must be rethought.

The above trends in AI and computer science considerably extend the logical agenda for studying rationality. A major view of computation from the 1980s onward is that of behavior by agents (van Benthem, 2018), studied by merging ideas from computational and philosophical logic (Gabbay & Guenthner, 1983–; Gabbay, Hogger, & Robinson, 1993–1998). After explaining what is involved in such agency in section 4, concrete examples will be given in sections 5 and 6, showing logic at work in this modern setting. However, the logic-oriented agency approach has not gone unchallenged. In section 7, we will discuss the "high" versus "low rationality" competition in understanding social agency and in section 8 the rise of nonrepresentational machine-learning techniques. Both come with a greater emphasis on probability, the other main formal paradigm for studies of rationality. We will end with an assessment of the current landscape of logico-computational approaches to rationality.

4. From Machines to Rich Rational Agents

4.1 A Conceptual Catalogue

Let us first think of what agents can do in general. Human agents have a much wider range of rational activities than just reasoning from given data, that is, elucidating what was already implicitly there. A rich dynamic information flow guides action. Agents constantly pick up new information from their environment by means of observation and communication, and they search their memory for information, too. Rationality is about picking up relevant information from whatever source is available, as much as reasoning, and that both in daily life and in science.

However, even rich processing of correct information is just one dimension. Information can be less or more reliable, and agents do not just accumulate knowledge but also form beliefs that can be shown wrong by new information. Thus, robustly rational agents are not those who are always correct but those who learn from errors and have a talent for correcting themselves (Kelly, 1996; Popper, 1963). Many facets of belief revision and learning are dealt with in chapter 5.2 by Rott; chapter 5.3 by Kern-Isberner, Skovgaard-Olsen, and Spohn; and chapter 5.4 by Gazzo Castañeda and Knauff (in this handbook). Logic is a major supplier of models here (van Benthem & Smets, 2015).

And rational agency does not stop here. Truly rational agents maintain a harmony between their information and beliefs with, on a par, their preferences, goals, or intentions. One can discuss which harmony is essential, whether maximizing expected utility (see chapter 8.2 by Peterson, this handbook) or some other option: logic cannot, and should not, decide. Real agents may differ widely in their distance from classical decision-theoretic views (cf. chapter 8.3 by Glöckner, this handbook), but the crucial point of rationality remains maintaining a workable balance between information, goals, and actions.

Summing up, a rational agent can gather information in a variety of ways, integrating observation, inference, and communication (van Benthem, 2011). In this process, the agent can form a rich variety of attitudes, from knowledge and belief to rejection or doubt. Also, the agent can function in environments where beliefs turn out wrong and learns from errors. And all this maintains a purpose, a balance between the agent's goals and its information or actions. And finally, even this is not yet a full picture: rational agents display their skills in social interaction, a topic that will return.

4.2 Multiagent Systems

Bits and pieces of this richer notion of rational agency have long been studied by philosophers and logicians: witness the *Handbook of Philosophical Logic* (Gabbay & Guenthner, 1983–). In the 1980s, a congenial picture of agency emerged in computer science and AI, namely in the study of multiagent systems (Fagin, Halpern, Moses, & Vardi, 1995; Shoham & Leyton-Brown, 2008; Wooldridge, 2010). This reflected a shift in thinking about computing, from machines to agents with a behavior analyzed in terms of features normally ascribed to humans. This extends to autonomous systems. Robots investigate their environment with sensors, decide, and act in performing their tasks (Cardon & Itmi, 2016). Again, tools from philosophical logic make sense (cf. Brafman, Latombe, Moses, & Shoham, 1997, on epistemic specifications for real robots, whose sensors have a margin of error). But conversely, notions from multiagent systems can be found in modern epistemology (Arló-Costa, Hendricks, & van Benthem, 2016). For instance, robots acting on evidence of varying quality have inspired new models for evidence-based belief (van Benthem & Pacuit, 2011).

4.3 Games

There is a natural confluence here with one more discipline. Agents that acquire information, choose actions, and pursue goals are like players in games. Indeed, computer science has drawn closer to game theory (Nisan, Roughgarden, Tardos, & Vazirani, 2007), and logic, with its connections to games of argumentation (cf. chapter 5.5 by Hahn & Collins and chapter 5.6 by Woods, both in this handbook) and information seeking (Hintikka, 1973), is a natural partner. Indeed, epistemic game theory (see chapter 9.2 by Perea, this handbook) can be seen as a venture created by these contacts.

Even with all this, no canonical view has crystallized yet of what a rational agent is and does that would be similar in elegance and fertility to that of a computing machine, let alone of a "universal rational agent" comparable in its sweep to the universal Turing machine. In fact, the earlier recognition of diversity suggests a focus on different kinds of rational agents rather than uniqueness, changing standard assessments of performance. Is a rational agent someone who wields vast cognitive powers or someone doing their best with limited powers? Are rational agents those who perform well against other rational agents or those who cope with a large bandwidth of types of agents in their environment?

5. Logical Models of Rational Agency

In recent decades, many features of rational agents have been studied by logicians. Information update and knowledge change occur in temporal logics of agency (Belnap, Perloff, & Xu, 2001; Fagin et al., 1995; Parikh & Ramanujam, 2003). Another paradigm is dynamic epistemic logic (Baltag, Moss, & Solecki, 1998; van Ditmarsch, van der Hoek, & Kooi, 2007), which models processes whereby agents form and modify representations of the information at their disposal. Such updates are not inferences, but they can be described just as precisely in logical terms.

Example: Dynamics of information flow. In a simple two-party dialogue, agent 1, who is uncertain about the truth or falsity of p, asks "p?" A second agent then truthfully and publicly replies, "Yes." Analyzing the information flow in the dialogue, the first agent's question conveys that she doesn't know the answer but also that she thinks the second agent, a fully reliable source, does know. The second agent's answer conveys that 1's assumption about 2's knowledge was correct, and also, after 2's answer, p is common knowledge in the group of these two agents. ⊣

This mixture of knowledge of facts and knowledge about others is typical for communication. More complex scenarios, such as the famous Muddy Children puzzle (Fagin et al., 1995), illustrate how even truthful public announcements of ignorance following consecutive questions can gradually lead agents to knowledge.

A symbolic language capturing all of this has formulas $[!\varphi]K\psi$ saying that the agent will know ψ after a public event carrying the information that φ is true. A key law in this logic of public announcements is the equivalence

$$[!\varphi]K\psi \leftrightarrow (\varphi \rightarrow K[!\varphi]\psi)$$

This relates new knowledge after the event φ happened to its "preencoding" before the event: the agent had conditional knowledge that the event $!\varphi$ would result in the truth of ψ. Such logical laws interchange dynamic operators for events and epistemic operators for attitudes of agents, a crucial ingredient in understanding information update and knowledge change.

Logics of belief change under hard and soft information have been developed in the same style (cf. van Benthem & Smets, 2015), and other formal approaches, such as "AGM," occur in chapter 5.2 by Rott (this handbook). Learning fits well with belief revision: connections between dynamic logics of knowledge and belief, on the one hand, and formal learning theory, on the other, are found in Baltag, Gierasimczuk, and Smets (2011).

Example: Belief change. In the above two-party dialogue, now assume that agent 2 is *not* a fully reliable information source. Starting from the initial situation, in which agent 1 believes neither p nor $\neg p$, the answer of 2 to her question can trigger 1 to change her mind and, possibly, to adopt a wrong belief. Yet how exactly she changes her mind depends on the trust that 1 has in 2 as an information source about p. If 2's answer "Yes" is considered to be reliable but not infallible, the "belief upgrade" that it triggers can be more radical, inducing a strong belief in p, or more conservative, inducing a weak belief in p. ⊣

Again, this process obeys logical laws. A formal language now has constructs $[\uparrow\varphi]$ for effects of conservative upgrades and $[\Uparrow\varphi]$ for radical upgrades. This brings to light many principles of belief change. For instance, $\neg K\neg\varphi \rightarrow [\uparrow\varphi]B\varphi$ says, for factual statements φ, that the agent comes to believe that φ is the case, unless she already knew before the announcement that φ was false. In logical studies of learning, one studies iterations of such upgrades and analyzes how well they perform as a learning method.

The study of key features of rationality in this style continues. Further aspects of purposeful rational behavior brought into the scope of logic include the management of current "issues" that guide inquiry for tasks at hand (cf. the "inquisitive logics" of Ciardelli, Groenendijk, & Roelofsen, 2019).

Next, moving from informational tasks to agents' preferences and goals, which determine how they evaluate situations, Liu (2011) studies logics of preference change, which is connected to goal dynamics. Preference dynamics shows similarities with deontic logics, describing what is obligatory and permitted for agents in environments where new commands change the moral ordering of situations and actions (Yamada, 2008).

Finally, rational agents balance their information with preferences and goals. Logics combining all these features occur in influential frameworks for multiagent systems such as BDI (Rao & Georgeff, 1991), inspired by Bratman (1987), describing how agency is driven by a balance of beliefs, desires, and intentions. But perhaps the most active research area where all these features of rational agents meet is in the logical study of strategic behavior and equilibria in games (cf. chapter 9.1 by Albert & Kliemt, this handbook).

Example: Reasoning about extensive games. Consider a finite game with two players, A and E, and outcome-values written in the order A-value—E-value (see figure 2.6.1).

Intuitively, the outcome (99, 99) seems best for both players—but that is not what the standard game solution algorithm of backward induction yields. Looking from the bottom to the top, if E is to play, she will choose left, and so, if A believes that E will make this choice, she herself will play left in the first round and end the game, with outcome (1, 0). ⊣

Analyzing why this game might play out in this non-Pareto-optimal manner involves many notions, including players' actions, beliefs, preferences, and plans. All of these have been studied by logicians, in different settings or as separate topics. For precise definitions of equilibria in games and a survey of logical game analysis, see van Benthem and Klein (2019).

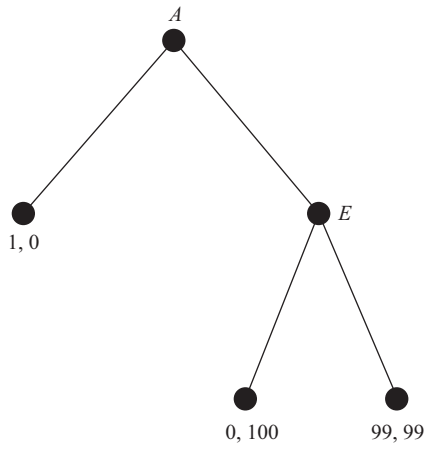


Figure 2.6.1
An extensive game.

All dynamic logics discussed here exemplify the earlier-mentioned tandem of algorithm and data in computation. The events that produce new information or new desires operate on well-chosen static models that support attitudes of knowledge, belief, preference, and the like.

Digression: Nonclassical logics. There are also approaches folding all of the above activities under varieties of inference, emphasizing departures from classical consequence to nonclassical, nonmonotonic logics (Horty, 2012) and resource-conscious substructural logics (Restall, 2000). For a comparison of the two methodologies, see van Benthem (2019).

Example: Nonmonotonic reasoning. Consequence in classical logic is monotonic; new premises do not invalidate earlier conclusions: if $\Gamma \models \varphi$ and $\Gamma \subseteq \Delta$, then $\Delta \models \varphi$. In contrast, default inferences are defeasible (cf. chapter 5.2 by Rott, this handbook): if I know that Tweety is a bird, I can conclude that Tweety can fly, yet with a further premise that Tweety is a penguin, it no longer follows that Tweety can fly. ⊣

The field of nonmonotonic logic studies properties of default reasoning. In contrast, dynamic logics of belief revision capture default phenomena on a classical base, locating the nonmonotonicity in belief change rather than in changing the inference rules: I believed that Tweety can fly, but after an event $!Penguin(Tweety)$, I have lost that belief.

Discussion. The logical study of ever more aspects of agency aims at a not exclusively behavioral view of rationality by identifying key internal features and mechanisms. But combining logical and computational agendas does not make logical systems realistic software agents or human agents. Far more is needed for algorithms to work, and implementation requires further syntax. Recent studies mediate between semantic models and syntactic representations for computing agents (Halpern & Rêgo, 2013; Lorini, 2018).

Also, the development of ever richer models raises questions. Where is the boundary of agency, as more and more topics are taken on board, and what is "rational" about the activities so described? Are we describing what agents actually do, or are these logical systems normative? A common view holds that logics of agency describe idealized laws that may or may not be followed by actual agents. This tension may be just what is needed. We cannot say, for instance, that belief revision leads to "correction" of earlier beliefs unless we have a norm for what is correct in the given circumstances. These are big issues that we cannot settle here but that permeate much of this handbook.

6. Rationality in Interactive Social Settings

The modern study of agency reflects the fact that the paradigm of computing today are distributed systems, not single machines. Likewise, multiagent systems put interacting individual agents at center stage and, at a next level, view groups themselves as actors, up to crowds or societies. This shifts the location of rationality from single agents to the quality of their interactions. And it broadens the focus from individual desires and actions to include emergent properties of the social system.

The interactive perspective is not alien to logic. Ever since Antiquity, dialogue, argumentation, and debate have been paradigmatic scenarios, and the rich interface of logic and games has been noted already (Hodges & Väänänen, 2001/2019; van Benthem, 2014). A core topic in epistemic logic is the rational ability to reason about others, with iterated forms such as "Agent *i* knows that agent *j* knows that . . ." and analogues for belief and other attitudes (cf. chapter 5.5 by Hahn & Collins and chapter 5.6 by Woods, both in this handbook). This recursion to higher levels is widespread: we can even be afraid of fear, of fear of fear, and so on. The extent to which human agents truly display these abilities is studied in cognitive psychology under the heading of Theory of Mind (Isaac, Szymanik, & Verbrugge, 2014; Premack & Woodruff, 1978). Iterated knowledge is used in computer science in analyzing correctness and security of communication protocols (Fagin et al., 1995).

But there is much more to social interaction than epistemic reflection. Strategic action involves dependencies of one agent's behavior on that of others or, better, on her expectations about the behavior of others (Aumann, 1995). Here is a simple illustration, a computational task in an interactive setting.

Example: Sabotage game. A Traveler in a graph moves along edges to reach some specified goal region. This graph-reachability problem is solvable in polynomial time. But now, there is a malevolent Demon who cancels an edge after each move Traveler makes. After that, Traveler goes along some still existing edge, and so on. ⊣

This "sabotage game" models search tasks under adverse circumstances and other social-informational scenarios. The solution complexity of the sabotage game jumps from polynomial time to complete for polynomial space. Logic helps determine who has a winning strategy in a given sabotage game by defining the basic challenge–response pattern, and it helps reason about general properties of such games. This is one case where logic meets "gamifications" of agency scenarios, and logics can even be used to devise concrete new practical games, thus

becoming a tool of design as much as of analysis (van Benthem, 2014).

With preference added, different notions of rationality have been investigated by logical means, such as those in game solution methods like backward induction, iterated removal of strictly dominated strategies (see chapter 9.1 by Albert & Kliemt and chapter 9.2 by Perea, both in this handbook), or iterated regret minimization (Halpern & Pass, 2009/2012). The structure of strategies in themselves is studied extensively at the interface of game theory, logic, and computer science (Brandenburger, 2014; van Benthem, Ghosh, & Verbrugge, 2015). Also, games influence computational logic: witness the "Boolean games" of Harrenstein, van der Hoek, Meyer, and Witteveen (2001), where players can manipulate truth values of propositions toward achieving their goals.

7. High and Low Rationality

At this point, a challenge arises to the preceding analyses of individual and social rationality. Classical game theory has agents that deliberate and design complex strategies, and rich epistemic and dynamic logics of agency reflect this. However, in evolutionary game theory (see chapter 9.3 by Alexander, this handbook), poor agents, perhaps hardwired biological types (Maynard Smith, 1982), do just as well.

Example: Evolutionary games. In a "Hawk–Dove" game, two individuals compete for a resource and can adopt either a Hawk or a Dove strategy. Hawks fight aggressively against other Hawks, in order to obtain the resource, until injury occurs and one retreats or, when facing a Dove, just take the resource, while a Dove retreats when facing a Hawk or shares the resource when facing a Dove. ⊣

Game theory computes equilibria here, which are typically in mixed strategies. With repeated Hawk–Dove games, the appropriate notion is that of an "evolutionarily stable" strategy, and it can be shown that certain mixtures of Hawk–Dove populations are stable, when the value of obtaining the resource is greater than the cost associated with possible injury in a fight. One can think of these mixed strategies as complex behavior for individual reasoning agents but also as just percentages of a population consisting of two types of agents, each just doing what it does, perhaps for biological reasons.

In the terminology of Skyrms (2010), the realm of complex reasoning players is that of "high rationality" and the realm of hardwired simple agents that of "low rationality." Often the latter do as well as the former. For instance, a classical game-theoretic analysis might say that through some sophisticated Kantian or Rawlsian argument involving thinking about others, we all arrive at the conclusion that we should live by the principles of morality, with the exception perhaps of a few free riders. By contrast, a game-theoretic evolutionary stability argument may tell us that in the long run, a population with a certain proportion of simple law-abiders (the prey) and lawbreakers (the predators), who both cannot help being what they are, is stable. No reasoning need be involved at all; the morality is emergent system behavior. This influx from evolutionary game theory reinforces the message of distributed computing: a society of many simple agents can produce highly complex behavior.

Here is one more scenario of emergent long-term complex behavior.

Example: Limit behavior in social networks. The following network has an update rule that agents s (the nodes) adopt a belief p if p is held by all their neighbors (that is, all nodes with an arrow pointing to them from s). Applied iteratively, the following evolutions may occur with different initial situations (see figure 2.6.2).

Case 1: Initial $p = \{1\}$. The second stage has $p = \varnothing$ (no one believes p), and this remains the outcome ever after. Case 2: Initial $p = \{2\}$. The next successive stages are $\{3\}$, $\{4\}$, $\{2\}$; from this stage onward, the network activation states loop. Case 3: Initial $p = \{1, 2\}$. The next stage is $\{3\}$, and we get an oscillation as before in case 2. Case 4: Initial $p = \{1, 2, 3\}$. We get $\{1, 3, 4\}$, $\{2, 4\}$, $\{2, 3\}$, $\{1, 3, 4\}$, and an oscillation starts here. ⊣

Thus, network update dynamics can stabilize in a single state (case 1) or oscillate in loops. Sometimes, successive models in a loop are isomorphic (cases 2 and 3), and sometimes the loop runs through different nonisomorphic network configurations (case 4). In infinite networks, an even further option is divergence toward ever different configurations. In all these scenarios, no logic seems involved in belief formation: behavior arises from agent types (the update rule) and the global structure of the network.

To put the challenge starkly, perhaps complex, logic-based rationality is not necessary to understand the behavior of human and artificial agents? But things are more complicated. In daily life, we think carefully in the "high" style about certain issues, but given our limited resources, we just follow, "low" style, our neighbors on perhaps the

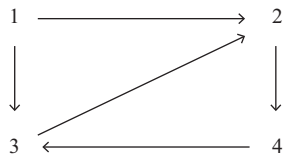


Figure 2.6.2
A social influence network.

majority of issues. This mixture calls for explanation, and current investigations are charting its details.

Combined high–low scenarios. Liu, Seligman, and Girard (2014) study agents in social networks that follow their neighbors' preferences, beliefs, or behavior via rules like following the majority or some other threshold, reflecting different agent types. The resulting diffusion process models the spread of fashions and new ideas. But agents still have epistemic states, and these can change dynamically as before, as described in an "epistemic friendship logic." In this setting, among other things, a logical characterization can be given of conditions for stabilization of agent's beliefs, thus predicting long-term system behavior. This framework combines ideas from sociology (Friedkin, 1998) with epistemic logic, adding an essential element to the earlier logical models of agency: the structure of the social network (see also chapter 10.1 by Dietrich & Spiekermann, this handbook). For a congenial study in another logical framework, see Xue (2017).

Other combined scenarios that have been studied include groups of individually rational agents who reason toward a common decision. Two things can happen: individual agents can enhance each other's reasoning power and bring about a higher level of group rationality, surpassing that of each individual agent. But groups may also get locked in irrational behavior. Whether the one will happen or the other is investigated in Baltag, Boddy, and Smets (2018), in terms of differences in interests and abilities between agents. One striking conclusion is that irrational group behavior is often not caused by irrational behavior of individual agents but by misalignments of their interests.

Other social phenomena that have been studied by logical techniques are informational cascades (Bikhchandani, Hirshleifer, & Welch, 1992), where a sequence of individual agents follows the decisions of their predecessors while ignoring their own private evidence. Baltag, Christoff, Hansen, and Smets (2013) ask whether individual rational agents, who use all their higher-order reasoning power, can stop a cascade from happening. The answer, surprisingly, is "no," and this fact can be proved by logical techniques that track information updates. However, the protocol regulating the agents' strategies matters: when agents have total communication and sharing of evidence, cascades can be stopped.

The preceding examples show how prima facie ideological differences between high and low rationality turn into a deeper study of how the two interface. This is logic at work at the ground level of agent activity. However, there is also a second, more methodological contact between the two sides: logic can analyze the structure of the dynamical systems theory underlying most low-rationality approaches and find patterns there, usually amounting to high-level qualitative descriptions of system behavior. Explorations in this direction include Kremer and Mints (2005) and Klein and Rendsvig (2017).

8. Machine Learning and Probability

In addition to the preceding tensions, the contemporary world of computing and AI offers a new challenge to logic-based views of rationality.

8.1 Machine Learning

Machine learning (Kelleher, Mac Namee, & D'Arcy, 2015) works well on large data sets, outperforming symbolic approaches, which tend to have problems of scalability. For instance, in supervised learning, a neural network is constructed, consisting of nodes with adjustable thresholds and links of adjustable strengths between nodes. Each setting for all of these produces an activity in the output layer given an input to the initial layer of the network. A cost function measures the distance of the current outputs to the desired ones on the training inputs. The network can then, by well-known techniques such as gradient descent (Russell & Norvig, 1994), adjust its weights and thresholds in the direction of lowering the cost function. In the end, stable optima in network activation are reached that work very well in new cases outside of the training set, in many computational and cognitive tasks. These networks, related to spin glass models in physics (Nishimori, 2001), use general statistical methods rather than specifically human agent features.

Neural networks in machine learning do not have any features immediately corresponding to classical logical models. There is no language and no representation, and the dynamic operations of the network do not reflect logical operations in any obvious manner. Also, very different stable states of the network resulting from training sessions can perform the same tasks, and invariants are hard to detect. Thus, whereas low-rationality methods raised the question whether logical analysis was necessary, machine-learning methods raise the question whether logical analysis is even possible.

It is far too early to adjudicate this debate. But here, too, there are some promising developments toward cooperation rather than antagonism. Integrating inference as used in neural networks and learning systems with symbolic reasoning is an active area of research (Baggio, van Lambalgen, & Hagoort, 2015; Balkenius & Gärdenfors,

2016; Leitgeb, 2004). "Explainable AI" seeks humanly intelligible qualitative patterns behind machine-learning systems, with topics such as causal reasoning (Halpern, 2016; Pearl, 2000; van Rooij & Schulz, 2019; see also chapter 7.1 by Pearl, this handbook), conditional logics as a way of classifying types of machine learning (Ibeling & Icard, 2018), or extracting logical inference modules from actual neural networks (Geiger, Icard, Lu, & Potts, 2021), while there is also a trend toward finding joint perspectives on learning in itself.

But also, recall a distinction made at the start of this chapter. If logic is only seen as a direct model for activities of reasoning or information update, other frameworks look like competitors. To some, the only question under debate is then whether logic can enhance such frameworks in terms of representation or computation. But in the more foundational sense of logic as an analysis of the structure of theories of computation and agency, even machine learning works on spaces with logical structure that can be described in logical terms (Leitgeb, 2017), and a meeting of the minds seems entirely feasible.

8.2 Probability

Continuing with methodological issues, here is one final contrast. A conspicuous feature of most studies of agency is the extensive use of probabilistic methods, a quantitative paradigm often seen as being at odds with qualitative logical analysis. Probability underlies many computational systems; it lies at the heart of game theory and dynamical systems theory, and in epistemology, probabilistic styles of analysis are at least as widespread as logic-based ones (see chapter 4.1 by Hájek & Staffel and chapter 4.7 by Dubois & Prade, both in this handbook).

The fruitful issue here is again one of combination. Qualitative and quantitative approaches naturally coexist, and the issue is just how. For instance, epistemic and doxastic logics, both static and dynamic, model uncertainty in terms of ranges of options (Adriaans & van Benthem, 2008), whereas Bayesian epistemology uses updates of probability functions (Bovens & Hartmann, 2003; Talbott, 2016). The compatibility of the two perspectives shows in combined systems (Halpern, 2003) that reason about both ontic and epistemic uncertainty, bringing together logic-based approaches with probabilistic conditioning. Other uses of probability concern action rather than information: witness the mixed strategies in game theory (for a logical perspective, see van Benthem & Klein, 2019). But there are also quite different interfaces of logic and probability, for instance, in the data-oriented parsing (DOP) architecture of Bod, Scha, and Sima'an (2003) and Bod (2008), which combines classical rule-based models of language and reasoning with probabilistic pattern recognition in a memory of earlier performance. Finally, the foundations of probability were still close to logic in the work of Boole and de Finetti, and various strands of research link the two realms in new ways. Harrison-Trainor, Holliday, and Icard (2018) study low-complexity qualitative reasoning systems that admit of introducing probability measures, while Leitgeb (2017) derives qualitative notions of belief from richer probabilistic models.

There are many further philosophical and technical issues to be explored at this rich and growing set of interfaces that we cannot cover in this chapter; the reader is referred to Spohn (2012) and chapter 5.3 by Kern-Isberner, Skovgaard-Olsen, and Spohn (this handbook).

9. Conclusion

This chapter has presented broad perspectives from logic and computation on rational agency. These ranged from high-level foundational insights into information and proof to specific studies of various abilities of information- and goal-driven agents. A rational agent, in this light, is a reasoner, information-processor, concept-crafter, and purpose-seeker: fallible but talented. Is it also a human cognitive agent? On connecting logic and computation to cognitive reality, we refer to Stenning and van Lambalgen (2008).

The main thrust of a logical approach as we see it is theoretical, but the deep entanglement of logic and computation over the past century has added practical dimensions. Rationality as studied here can be programmed and put into intelligent systems, even though the path to feasibility is not easy or trivial. It is this very distance that allows logical theories to also be normative, providing an essential tension between the real and the ideal in the study of rational behavior, which keeps sparking further investigation.

We have not hidden the fact that the classical logico-computational paradigm faces challenges, coming from probability theory, dynamical systems theory, and machine learning. But we think this is all to the good, since these challenges suggest new interface topics of interest to all.

Finally, it should be clear that we have not claimed that logic is the only game in town. Neither is computation. The approach surveyed in this chapter does not hold the unique key to understanding the rich phenomenon of rationality, but it does offer one valid and illuminating perspective.

References

Adriaans, P., & van Benthem, J. (Eds.). (2008). *Handbook of the philosophy of science: Vol. 8. Philosophy of information*. Amsterdam, Netherlands: Elsevier.

Andrews, G. R. (2000). *Foundations of multithreaded, parallel, and distributed programming*. Reading, MA: Addison-Wesley.

Arló-Costa, H., Hendricks, V. F., & van Benthem, J. (Eds.). (2016). *Readings in formal epistemology: Sourcebook*. Cham, Switzerland: Springer.

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior*, *8*(1), 6–9.

Baggio, G., van Lambalgen, M., & Hagoort, P. (2015). Logic as Marr's computational level: Four case studies. *Topics in Cognitive Science*, *7*(2), 287–298.

Balkenius, C., & Gärdenfors, P. (2016). Spaces in the brain: From neurons to meanings. *Frontiers of Psychology*, *7*, 1820.

Baltag, A., Boddy, R., & Smets, S. (2018). Group knowledge in interrogative epistemology. In H. van Ditmarsch & G. Sandu (Eds.), *Jaakko Hintikka on knowledge and game-theoretical semantics* (Outstanding Contributions to Logic, Vol. 12, pp. 131–164). Cham, Switzerland: Springer.

Baltag, A., Christoff, Z., Hansen, J. U., & Smets, S. (2013). Logical models of informational cascades. In J. van Benthem & F. Liu (Eds.), *Logic across the University: Foundations and Applications: Proceedings of the Tsinghua Logic Conference, Beijing, 2013* (Studies in Logic, Vol. 47, pp. 405–432). London, England: College Publications.

Baltag, A., Gierasimczuk, N., & Smets, S. (2011). Belief revision as a truth-tracking process. In K. Apt (Ed.), *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge, TARK XIII* (pp. 187–190). New York, NY: ACM.

Baltag, A., Moss, L. S., & Solecki, S. (1998). The logic of common knowledge, public announcements, and private suspicions. In I. Gilboa (Ed.), *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)* (pp. 43–56). San Francisco, CA: Morgan Kaufmann.

Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the future: Agents and choices in our indeterminist world*. Oxford, England: Oxford University Press.

Bergstra, J., Ponse, A., & Smolka, S. (2001). *Handbook of process algebra*. Amsterdam, Netherlands: Elsevier.

Beth, E. W. (1971). *Aspects of modern logic*. Dordrecht, Netherlands: Reidel.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Bod, R. (2008). The data-oriented parsing approach: Theory and application. In J. Fulcher & L. C. Jain (Eds.), *Computational intelligence: A compendium* (Studies in Computational Intelligence, Vol. 115, pp. 307–348). Berlin, Germany: Springer.

Bod, R., Scha, R., & Sima'an, K. (2003). *Data-oriented parsing*. Stanford, CA: CSLI Publications.

Bonizzoni, P., Brattka, V., & Löwe, B. (Eds.). (2013). *The Nature of Computation: Logic, Algorithms, Applications: 9th Conference on Computability in Europe, CiE 2013, Milan, Italy, 2013, Proceedings* (LNCS 7921). Heidelberg, Germany: Springer.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford, England: Clarendon Press.

Brafman, R. I., Latombe, J.-C., Moses, Y., & Shoham, Y. (1997). Applications of a logic of knowledge to motion planning under uncertainty. *Journal of the ACM*, *44*(5), 633–668.

Brandenburger, A. (Ed.). (2014). *The language of game theory: Putting epistemics into the mathematics of games* (World Scientific Series in Economic Theory, Vol. 5). Singapore: World Scientific.

Bratman, M. E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Cardon, A., & Itmi, M. (2016). *New autonomous systems* (Vol. 1). Hoboken, NJ: Wiley.

Chakraborty, P., Saxena, P. C., & Katti, C. P. (2011). Fifty years of automata simulation: A review. *ACM Inroads*, *2*(4), 59–70.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics, 58*(2), 345–363.

Ciardelli, I., Groenendijk, J., & Roelofsen, F. (2019). *Inquisitive semantics*. Oxford, England: Oxford University Press.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge*. Cambridge, MA: MIT Press.

Friedkin, N. E. (1998). *A structural theory of social influence*. Cambridge, England: Cambridge University Press.

Gabbay, D. M. (1999). *Fibring logics*. Oxford, England: Clarendon Press.

Gabbay, D. M., & Guenthner, F. (Eds.). (1983–). *Handbook of philosophical logic*. Dordrecht, Netherlands: Springer.

Gabbay, D. M., Hogger, C. J., & Robinson, J. A. (Eds.). (1993–1998). *Handbook of logic in artificial intelligence and logic programming*. Oxford, England: Clarendon Press.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural natural language inference models. CSLI, Stanford University.

Gödel, K. (1986). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I [On formally undecidable propositions of Principia Mathematica and related systems]. In S. Feferman, J. W. Dawson, Jr., S. C. Kleene, G. H.

Moore, R. M. Solovay, & J. van Heijenoort (Eds.), *Kurt Gödel: Collected works* (pp. 144–195). Oxford, England: Clarendon Press. (Original work published 1931)

Halpern, J. Y. (2003). *Reasoning about uncertainty*. Cambridge, MA: MIT Press.

Halpern, J. Y. (2016). *Actual causality*. Cambridge, MA: MIT Press.

Halpern, J. Y., & Pass, R. (2012). Iterated regret minimization: A new solution concept. *Games and Economic Behavior*, *74*(1), 184–207. (Original work published 2009)

Halpern, J. Y., & Rêgo, L. C. (2013). Reasoning about knowledge of unawareness revisited. *Mathematical Social Sciences*, *65*(2), 73–84.

Harel, D. (1987). *Algorithmics: The spirit of computing*. Reading, MA: Addison-Wesley.

Harrenstein, P., van der Hoek, W., Meyer, J.-J., & Witteveen, C. (2001). Boolean games. In J. van Benthem (Ed.), *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge* (pp. 287–298). San Francisco, CA: Morgan Kaufmann.

Harrison-Trainor, M., Holliday, W. H., & Icard, T. F., III. (2018). Inferring probability comparisons. *Mathematical Social Sciences*, *91*, 62–70.

Haugeland, J. (Ed.). (1997). *Mind design II: Philosophy, psychology, artificial intelligence*. Cambridge, MA: MIT Press.

Hintikka, J. (1973). *Logic, language-games, and information: Kantian themes in the philosophy of logic*. Oxford, England: Clarendon Press.

Hodges, W., & Väänänen, J. (2019). *Logic and games*. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2019/entries/logic-games/

Horty, J. F. (2012). *Reasons as defaults*. Oxford, England: Oxford University Press.

Ibeling, D., & Icard, T. (2018). On the conditional logic of simulation models. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)* (pp. 1868–1874). Stockholm, Sweden: AAAI Press.

Isaac, A., Szymanik, J., & Verbrugge, R. (2014). Logic and complexity in cognitive science. In A. Baltag & S. Smets (Eds.), *Johan van Benthem on logic and information dynamics* (pp. 787–824). Cham, Switzerland: Springer.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. Cambridge, MA: MIT Press.

Kelly, K. T. (1996). *The logic of reliable inquiry*. New York, NY: Oxford University Press.

Klein, D., & Rendsvig, R. K. (2017). Convergence, continuity and recurrence in dynamic epistemic logic. In A. Baltag, J. Seligman, & T. Yamada (Eds.), *Logic, Rationality, and Interaction: 6th International Workshop, LORI 2017, Sapporo, Japan, September 11–14, 2017, Proceedings* (pp. 108–122). Berlin, Germany: Springer.

Kneale, W., & Kneale, M. (1962). *The development of logic*. Oxford, England: Clarendon Press.

Kremer, P., & Mints, G. (2005). Dynamic topological logic. *Annals of Pure and Applied Logic*, *131*(1–3), 133–158.

Leitgeb, H. (2004). *Inference on the low level: An investigation into deduction, nonmonotonic reasoning, and the philosophy of cognition*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Leitgeb, H. (2017). *The stability of belief: How rational belief coheres with probability*. Oxford, England: Oxford University Press.

Liu, F. (2011). *Reasoning about preference dynamics* (Vol. 354 of Synthese Library). Dordrecht, Netherlands: Springer.

Liu, F., Seligman, J., & Girard, P. (2014). Logical dynamics of belief change in the community. *Synthese*, *191*(11), 2403–2431.

Lorini, E. (2018). In praise of belief bases: Doing epistemic logic without possible worlds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 1915–1922).

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge, England: Cambridge University Press.

Nisan, N., Roughgarden, T., Tardos, E., & Vazirani, V. V. (Eds.). (2007). *Algorithmic game theory*. Cambridge, England: Cambridge University Press.

Nishimori, H. (2001). *Statistical physics of spin glasses and information processing: An introduction*. Oxford, England: Oxford University Press.

Papadimitriou, C. H. (1994). *Computational complexity*. Reading, MA: Addison-Wesley.

Parikh, R., & Ramanujam, R. (2003). A knowledge based semantics of messages. *Journal of Logic, Language and Information*, *12*, 453–467.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London, England: Routledge.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *4*, 515–526.

Rao, A., & Georgeff, M. (1991). Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, & E. Sandewall (Eds.), *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning* (pp. 473–484). San Mateo, CA: Morgan Kaufmann.

Restall, G. (2000). *An introduction to substructural logics*. London, England: Routledge.

Russell, S., & Norvig, P. (1994). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Shoham, Y., & Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge, England: Cambridge University Press.

Skyrms, B. (2010). *Signals: Evolution, learning, & information*. Oxford, England: Oxford University Press.

Spohn, W. (2012). *The laws of belief: Ranking theory and its philosophical applications*. Oxford, England: Oxford University Press.

Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.

Talbott, W. (2016). Bayesian epistemology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian/

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*, 230–265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *236*, 433–460.

van Benthem, J. (2011). *Logical dynamics of information and interaction*. Cambridge, England: Cambridge University Press.

van Benthem, J. (2014). *Logic in games*. Cambridge, MA: MIT Press.

van Benthem, J. (2018). Computation as social agency: What, how and who. *Information and Computation*, *261*(3), 519–535.

van Benthem, J. (2019). Implicit and explicit stances in logic. *Journal of Philosophical Logic*, *48*(3), 571–601.

van Benthem, J., Ghosh, S., & Verbrugge, R. (Eds.). (2015). *Models of strategic reasoning: Logics, games, and communities* (Lecture Notes in Computer Science, Vol. 8972). Berlin, Germany: Springer.

van Benthem, J., & Klein, D. (2019). Logics for analyzing games. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/spr2019/entries/logics-for-games/

van Benthem, J., & Pacuit, E. (2011). Dynamic logics of evidence-based beliefs. *Studia Logica*, *99*(1–3), 61–92.

van Benthem, J., & Smets, S. (2015). Dynamic logics of belief change. In H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, & B. Kooi (Eds.), *Handbook of logics for knowledge and belief* (pp. 313–393). London, England: College Publications.

van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2007). *Dynamic epistemic logic*. Dordrecht, Netherlands: Springer.

van Harmelen, F., Lifschitz, V., & Porter, B. (Eds.). (2008). *Handbook of knowledge representation* (Foundations of Artificial Intelligence). Amsterdam, Netherlands: Elsevier.

van Rooij, R., & Schulz, K. (2019). Conditionals, causality and conditional probability. *Journal of Logic, Language and Information*, *28*(1), 55–71.

Wang, H. (1996). *A logical journey: From Gödel to philosophy*. Cambridge, MA: MIT Press.

Wooldridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). Chichester, England: Wiley. (Original work published 2002)

Wooldridge, M. (2010). *Reasoning about rational agents*. Cambridge, MA: MIT Press.

Xue, Y. (2017). *In search of homo sociologicus* (Unpublished doctoral dissertation). The Graduate Center, City University of New York.

Yamada, T. (2008). Logical dynamics of some speech acts that affect obligations and preferences. *Synthese*, *165*(2), 295–315.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Knauff, Markus, editor. | Spohn, Wolfgang, editor.
Title: The handbook of rationality / edited by Markus Knauff and Wolfgang Spohn.
Description: Cambridge : The MIT Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020048455 | ISBN 9780262045070 (hardcover)
Subjects: LCSH: Reasoning (Psychology) | Reason. | Cognitive psychology. | Logic. | Philosophy of mind.
Classification: LCC BF442 .H36 2021 | DDC 153.4/3—dc23
LC record available at https://lccn.loc.gov/2020048455