# Psychological and Philosophical Frameworks of Rationality—A Systematic Introduction

Markus Knauff and Wolfgang Spohn


## Summary

Rationality is a vast issue. It has occupied the most brilliant thinkers since the beginning of human culture. It mirrors our capacity for self-reflection, our desire to recognize the potentials and limitations of our mental abilities, and our determination to understand how our mind works. Such questions can be frightening, as they seem so complex that we may never find conclusive answers. But we can try. In this introductory chapter to *The Handbook of Rationality*, we unfold the concept of rationality and the relations between theoretical and practical rationality, normative and descriptive theories of rationality, individual and social rationality, and an output-oriented and a process-oriented perspective on rationality. We also describe some cognitive preconditions and domains of rationality and the different intellectual traditions in psychology and philosophy—where they intersect, fall apart, and converge. We hope that this overview helps readers to orient themselves in the widely ramified research on human rationality.


## 1. Why Study Rationality?

Humans are rational animals. Since Aristotle, this has been considered the essence of *Homo sapiens*. Of course, much about this view can be criticized: one objection might be that irrationality seems to prevail. The human species is about to make our planet uninhabitable. The globe is full of political, economic, cultural, and religious conflicts. Often, people seem to be driven not by rationality but rather by power struggles, anxiety, greed, prejudice, and so on. All of this can scarcely be called rational. However, it is hard to argue about the relative amounts of rationality and irrationality. Rationality attributions are delicate, as this handbook will richly display. We well know that a collective of rational individuals may end up in an irrational collective disaster (Janis, 1972; Moorhead, Ference, & Neck, 1991; Reason, 1987). And the extent of the irrationality often depends on the perspective of the assessor. In fact, to call humans rational is to say that they are *gifted* with rationality, but not that they *exercise* this gift all the time. In Chomsky's terms, rationality is a *competence*, and the *performance* may well be defective.

A second objection to the definition of humans as rational animals might be that rationality is not unique to humans. Today we know that many other animals have mental capacities that were traditionally assumed to distinguish *Homo sapiens*. Of course, it is important to study continuities and differences between potentially rational creatures, and there is no doubt that human rationality has evolved from more fundamental neural mechanisms of adaptation and cognition. However, our mental powers of representation, of differentiation, of inner and outer experience, are much larger than those of any other animal. We find full-blown rationality only in us, although it also exists in basic forms in other species.

A third objection to the definition of humans as rational animals might be that various other characteristics distinguish our species from other animals as well: language, religion, pedagogics, arts, aesthetics, humor—the list is endless. What is so special about rationality? Take language, for instance: one might say that language is not only for communication but also for so many other social and individual purposes. Hence, compared to rationality, it is the far more comprehensive phenomenon. However, the many uses of language depend on the serious and sincere exchange of information and expression of beliefs and desires (see D. Lewis, 1969, and chapter 10.3 by Meggle, this handbook). Without the formation of such beliefs and desires—an essentially rational process—language would not be what it is.

Still, some of these objections may be justified. However, if we really were to doubt that rationality is at the heart of our existence as human beings, we would have difficulties explaining how people think, reason, judge, make decisions, solve problems, and act. Only on this rationality assumption can we explain how we interact, communicate, and form societies. And only on this

assumption can we explain how humans developed science, culture, and modern technology. All these abilities and achievements rely on our mental powers of representation and processing, which are implemented in the neural structure of our brains. Thus, it is legitimate to question how rational humans are, but shifting priority away from rationality would be a mistake.

Today we are facing many vital challenges, from fighting starvation and curtailing war to coping with climate change, pandemics, worldwide diseases, and the drawbacks of new technologies and globalization. Indeed, for the first time in history, it seems that these challenges may concern humankind globally and substantially. One may well argue about what will help us most in meeting these challenges: empathy, justice, modesty, peacefulness, honesty, solidarity, the right values, religion? Surely rationality alone won't do. Yet there is no way to bypass rationality. Theories of justice refer to theories of rationality; we must act rationally to reach our moral, ethical, and humanist goals; and even religion cannot ignore rationality. So, whichever human features we promote, whatever measures we propose in order to meet our challenges, they should be well reasoned and withstand our critical reflection.

These first thoughts already reveal the two fundamental aims of this handbook. One aim is to bring together mainly psychological and mainly philosophical accounts of rationality and to display what both disciplines can learn from each other. The other aim is to substantiate the relation between the *normative* and the *descriptive* perspective on human rationality,[1] between what human thinking *ought* to be like and what it actually *is* like.

One might say that philosophy has been attempting to grasp rationality for 2,500 years, while psychology has been doing so only since the end of the 19th century. But this does not adequately represent the intellectual history. Large parts of this history, beginning already with the Presocratics (see Lorenz, 2009), were concerned with psychological issues, even if called philosophy, and were indeed labeled "philosophical psychology." Yet, for a long time, the methodology of this field was not clear, and hence the normative and the descriptive perspective were often not clearly defined and separated (see chapter 1.1 by Sturm, this handbook). This changed at the end of the 19th century, when psychology became an independent academic discipline, and one of its first topics was indeed to empirically investigate the specific conditions under which cognition and action conform to or deviate from what were at that time regarded as the norms of rationality (Störring, 1908; Wundt, 1896/2010). Interestingly, while this happened in scientific psychology, logic developed in the opposite direction and became determinedly antipsychologistic under the influence of Frege (1884). We will return to this below.

Nowadays, the roles of psychology and philosophy in the academic landscape are clarified. Psychology is now the empirical science of the mind, with its own methodology, while philosophy sees more clearly that the normative issues belong to its proper domain. Perhaps because each discipline first had to find its role, the discourse between the two disciplines was initially poor or appeared quite unproductive. However, this has recently changed to a high degree, as will be amply displayed in this handbook.[2]

Clearly, both perspectives, the descriptive and the normative one, are each important and valuable in themselves. Their relation, however, is utterly contested. Some scholars argue that psychology should ignore the normative work in philosophy and simply go ahead describing how people think and reason (Elqayam & Evans, 2011). Other scholars refer to the fundamental distinction between competence and performance and argue that it is in principle impossible to empirically confirm that humans are irrational (Cohen, 1981). If either of these camps is right, there is no tension between the normative and the descriptive perspective on rationality—they are either independent or basically coincide. However, both views are minority positions in the community of rationality researchers. The premise of this handbook is instead that a comprehensive account of human rationality requires both empirically evaluated descriptive theories and elaborated normative theories as a positive or a negative point of comparison.

The aim of this handbook is to give a comprehensive overview of the state of the art in philosophy and psychology, and of the areas where the two disciplines are connected, where they share interests and concepts, and where they have diverging interests and research agendas. The handbook has many contributions from both disciplines and also includes some approaches from economics, sociology, artificial intelligence, and cognitive neuroscience, even if it cannot do so exhaustively.

This introductory chapter provides some structure to our huge topic. It does not follow the table of contents, which we already explained in the overview of the handbook. Rather, the arrangement of the chapter reflects some basic distinctions, conceptual frameworks, and research lines in the philosophical and psychological exploration of human rationality. It is organized as follows:

In section 2 of this chapter, we start with a preliminary contour of the subject matter we are talking about

when we use the term "rationality." Then, in section 3, we describe the essential distinction between *theoretical* or *epistemic* and *practical* rationality. Psychologists are used to making a related distinction between *reasoning* and *decision making*. Next, in section 4, we will discuss the important tension between a *normative* and a *descriptive* perspective on rationality. As we just stated, this is one of the most crucial issues in our psychological-philosophical enterprise. In section 5, we introduce the distinction between *individual* and *collective* or *social* rationality, a topic that overlaps with topics from economics and other social sciences. Section 6 is concerned with the distinction between what we call *output-oriented* and *process-oriented* theories of rationality. While the former is the dominant approach in philosophy, the latter lies at the heart of cognitive psychology, which seeks to understand the processes that lead from the input given to the cognitive system to the output generated by the system. The four distinctions in sections 3 to 6 can be combined with each other and thus lead to a *systematic framework* that, for the first time, combines philosophical and psychological theories and empirical results on human rationality. We think that this framework not only guides the readers through this handbook but also provides a new perspective on research on human rationality. Then, in section 7, we will be concerned with some of the *preconditions* of human rationality, that is, the neural underpinnings of thinking and reasoning, as well as the relation between rationality and intelligence, memory, and other cognitive functions. In section 8, we discuss whether special forms of rationality exist, for example, in science, communication, or artificial intelligence (AI). We also discuss some connections of rationality to concepts such as emotion, morality, and culture. In the (final) section 9 of this chapter, we conclude with some open questions and problems that interdisciplinary rationality research still has to tackle.

## 2. Contours of the Concept of Rationality

Rationality constitutes a rich conceptual field. Broadly construed, it is about our higher cognitive faculties, about acquiring concepts, forming beliefs, gaining knowledge, inferring and reasoning, thinking, judging, decision making, planning, deliberating, calculating, and satisfying wishes, needs, and desires. In recent terms, we may also say that it is about our intentional and propositional attitudes. Being endowed with self-reflection, humans have surely wondered about these issues all along and so started thinking about how the human mind might work.

The field is also most confusing, tentatively tamed by many incongruous terms. If we look only at the five languages in which essential parts of early philosophical psychology were conducted—Greek, Latin, French, English, and German—we find many relevant terms, often accompanied by full doctrines. These terms are neither easily translatable, nor do they have a stable meaning. Think, for instance, of the distinction between *understanding* and *reason*, which acquired ever more significance in 17th- and 18th-century philosophy, culminating in Immanuel Kant, for whom that distinction between "Verstand" and "Vernunft" took a very specific form closely intertwined with his entire philosophical edifice. In the hermeneutic tradition, then, understanding was rather something opposed to explanation (von Wright, 1971). Nowadays, these distinctions are treated as spurious by most authors, with the exception, perhaps, of orthodox Kantians.

Or consider the English term "reason," which has three different meanings, for which German, for example, uses three different words. First, "reason" without a determiner (*Vernunft*) is just a general colloquial term for our higher cognitive faculties (which derives from medieval German *vernehmen* which meant *erfassen*, i.e., "to grasp"). Second, "a reason" (*Grund*) is what we have or give or accept in order to explain or justify whatever may be explained or justified. And finally, "to reason" (*räsonieren*, *schließen*) is something we do when we argue, make inferences, or arrive at conclusions. The connection presumably is that when we reason or give reasons, we *use* our reason. However, when we give reasons, we are not necessarily reasoning, and reasoning need not proceed from reasons (but only from premises the status of which may be left open). That's the English muddle, which will occupy us later. Other languages have their own difficulties.

The etymological origin of these terms is the Greek *raetos*, which, among other things, means "rational," in the sense of rational numbers, and thus displays another aspect of our conceptual field. The term ratio owed its tremendous career also to the fact that Cicero (106–43 BC) established it as the standard translation of the Greek *logos*. The term *rationalitas*, however, is first found in Tertullian (ca. 150–220 AD), for whom rationality was one of the essential attributes of the soul. Since possessing rationality simply amounts to being endowed with reason, the term *ratio*/"reason" was the widely used one, while *rationalitas*/"rationality" played a minor role in history.[3]

The term "rationality" became fashionable only at the end of the 19th century under the influence of economics and social science. In philosophy, this was perhaps due to the fact that the normative dimension was seen

more clearly in the 20th century and was more closely associated with the new term "rationality" than with old terms like "reason" and "understanding." In psychology, the term became important in controversies about the role of empirical psychological laws for logic and epistemology (Wundt, 1910). More details on the historical developments in philosophy and psychology are represented in chapter 1.1 by Sturm and chapter 1.2 by Evans (both in this handbook).[4] In the following, we just describe some milestones in the recent history of our topic that were particularly influential in shaping the current state of the art and still guide ongoing debates in philosophy and psychology.

### 2.1 Milestones in the Study of Rationality

Let us start our milestones with *George Boole* and *Gottlob Frege*, two eminent philosophers and logicians. Boole (1854/1951) developed a new logic, now called "propositional calculus," which forms the basic part of classical logic. The title of his book, *An Investigation of the Laws of Thought*, indicates that Boole was concerned with human thinking, not with the abstract truth of propositions. He wrote, "The design of the following treatise is to investigate the fundamental laws of those operations of the mind by which reasoning is performed" (Boole, 1854/1951, p. 1).

Frege later argued in exactly the opposite direction and became one of the most influential thinkers in the history of logic and philosophy. One reason for the importance of Frege's work is that his revolution of logic was not only a revolution of large areas of philosophy. It was also the seed of what later emerged as analytic philosophy. Another reason is that his famous *antipsychologism* was essential for the separation of philosophy and psychology, a gap that we want to bridge in this handbook. In his *Grundlagen der Arithmetik*, Frege suggested, as one of three fundamental principles,[5] that one must "separate sharply the psychological from the logical, the subjective from the objective" (Frege, 1884, p. x). For Frege, grasping a thought such as Newton's law of gravitation "is a process which takes place at the very confines of the mental and which for that reason cannot be completely understood from a purely psychological standpoint. For in grasping the law something comes into view whose nature is no longer mental in the proper sense, namely the thought" (Frege, 1979, p. 145). Here, one must observe Frege's peculiar use of "thought." For him, thoughts are objective entities belonging to the abstract, nonmental "third realm" of "senses" (i.e., meanings). Thus, logic is about the laws of truth, which resemble laws of nature. These laws entail laws of thought. "Rules for asserting, thinking, judging, inferring follow from the laws of truth" (Frege, 1918, p. 289). Here, "rules" translates *Vorschriften*; they are prescriptions. So, Frege's main claim was that the laws of thought are normative laws, not empirical laws of nature.

In the same year in which Frege published his *Begriffsschrift*, namely in 1879, *Wilhelm Wundt* founded the first German institute for experimental psychology at the University of Leipzig. Wundt, who called Frege's antipsychological approach *logicism*, argued that logic is a result of psychological functions and thus a branch of psychology. Although already the philosopher David Hume took psychology to be the backbone of philosophy, Wundt was probably the first psychologist who systematically applied this view to the study of logic and rationality (Wundt, 1910).

Wundt still retained the connection of psychology to philosophy. However, his work was also a milestone with regard to the academic independence of psychology. One of his many scholars, *Gustav W. Störring*, may be said to have started empirical research on cognition and rationality when he published a 127-page article (Störring, 1908) in which he described the first systematic experiments on human reasoning. In his studies, volunteers had to solve a battery of deductive reasoning problems, while their verbal responses, response times, eye movements, gestures, and even the expansion and contraction of their chest were measured. This work was part of Störring's attempt to develop psychology into a more experimental-physiological direction. He was occupied with the laws of thought in a descriptive, not a normative, sense.

Probably one of the most influential thinkers about the nature of human rationality was the developmental psychologist *Jean Piaget*. In his quite extensive publications, Piaget placed great importance on the development of rationality in children, culminating in his pioneering book *The Growth of Logical Thinking* (Inhelder & Piaget, 1955/1958). His approach was often referred to as "genetic epistemology," as he explained how people acquire knowledge and reasoning abilities by the biological functions of accommodation and assimilation to the environment. According to Piaget, people develop cognitively from birth throughout their lives in four different stages: the sensorimotor (birth to age 2), preoperational (2–7 years), concrete operational (7–11 years), and formal operational (11 years and onward) stages. For Piaget, rational reasoning is practically identical with logical reasoning and develops primarily in the formal operational phase, where children learn to think about abstract entities and concepts and develop logical abilities for reasoning, planning, and problem solving. For Piaget, these rational abilities rely on abstract formal rules of inference stored in long-term memory, which he

thought to be almost equivalent to the rules of formal logic. Later, this approach was developed further in rule-based theories of reasoning, one of which is described in chapter 3.2 by O'Brien (this handbook). This approach, however, has been attacked by the adherents of several other theories, notably the *theory of mental models*, which has been developed by Philip Johnson-Laird from the 1980s onward and is another milestone in the history of the psychology of reasoning (Johnson-Laird, 1983, 2006, 2010; Johnson-Laird, Khemlani, & Goodwin, 2015). This theory is described in detail in chapter 2.3 by Johnson-Laird (this handbook; see also chapter 3.3 by Khemlani and chapter 6.3 by Byrne & Espino, both in this handbook).

Another major episode in the discussion of rationality was the *Popper–Kuhn controversy*. With Frege's new logic as a firm base, *Karl Popper* (1934/1989) developed his falsificationist hypothetico-deductive scientific methodology, which purported to state how science has to rationally proceed. Popper had a dispute about this with his positivistic colleagues in the Vienna Circle (e.g., Rudolf Carnap). And there was further controversy, notably in the social sciences, where the Frankfurt school (Habermas, Adorno, etc.) accused Popper of being a positivist. However, the dispute widened dramatically when *Thomas S. Kuhn* (1962) meticulously showed that scientists actually do not follow Popper's methodology. One may conclude, as Popper did, that scientists proceed irrationally. But this is implausible. Isn't rationality the hallmark of the sciences? Thus, Kuhn's observations were taken rather as an empirical argument against Popper's normative methodology—a most remarkable fact—and as a challenge to find a more adequate normative methodology that does not accuse scientists of being irrational. *Imre Lakatos's* (1978) "methodology of scientific research programmes" was such an attempt. This gave rise to what is nowadays called "science and technology studies," an amalgam of history, sociology, and philosophy of science. However, the controversy also motivated more abstract investigations into the logic of theory change and the nature of defeasible reasoning, which surface in many places in this handbook.

Popper's logic-inspired falsificationism was a direct reference point for another milestone of cognitive rationality research: Wason's selection task, which is sometimes called "the Drosophila of the psychology of reasoning" (Beller, 1999). The task is named after *Peter Wason*, a cognitive psychologist who wanted to explore whether people indeed seek to falsify logical rules as required by Popper (Wason, 1966, 1968). The task is as follows:

> You are shown a set of four cards placed on a table, each of which has a number on one side and a letter on the other side. The visible faces of the cards show A, D, 4, and 7. The cards obey the rule: if there is a vowel on the one side of the card, there is an even number on the other side of that card. Which card(s) must you turn over in order to test the truth of this rule?


According to the norms of propositional logic, participants should check the validity of the given rule by turning over the cards with the A (modus ponens) and the 7 (modus tollens) on the front. Typically, fewer than 10 out of 100 participants can solve this problem correctly (J. St. B. T. Evans, Newstead, & Byrne, 1993). Why is that so? Psychologists have given many different answers to this question, some of which are discussed in chapter 1.3 by Schurz, chapter 2.3 by Johnson-Laird, chapter 4.5 by Chater and Oaksford, and chapter 15.1 by Markovits (all in this handbook).

Such tasks were enormously important for the psychology of reasoning. They were designed to study *pure* inference processes without the contamination of true or false prior beliefs. The participants were instructed to draw logically valid conclusions from the premises, which they had to consider as facts that could not be questioned. Since the content of the premises did not matter, the only possible error seemed to be that the conclusion was inferred from the given premises in an unreasonable way. Although this paradigm has many advantages, it also led researchers to pay little attention to the contents of the beliefs represented in the premises. Participants are simply forced to believe that the premises are true; otherwise their responses are often excluded from the analysis.

Later, the pendulum swung more toward the study of beliefs and away from processes of reasoning. This development proceeded roughly in two steps. The first step consisted in research paradigms invented to study interactions between logicality and believability. In such experiments, a conclusion that is logically valid or invalid in relation to the premises could be true or false in relation to the participants' prior beliefs. This introduces the *belief bias*, which arises if a participant in an experiment infers a logically invalid conclusion from the premises because it agrees with her prior beliefs or, conversely, if she rejects a conclusion because it is implausible, even though it is logically valid (J. St. B. T. Evans, 1989). Such *content-effects* became very prominent in cognitive psychology and have been identified in almost all areas of human reasoning (J. St. B. T. Evans, 1993).

The second step in developing more belief-related reasoning research was even more fundamental. Under

the label *new paradigm psychology of reasoning* (Oaksford & Chater, 2007, 2020), researchers started around the year 2010 to focus on the subjective degrees of belief that people might have in the contents of the premises and on how those affect the acceptability of the conclusion. Here, a different connection between psychology and Kuhn's theory of progress in science shows up. Proponents of this approach see their new paradigm as "revolutionary" in the Kuhnian sense,[6] as it replaces the norms of two-valued classical logic by those of Bayesian probability theory (Oaksford & Chater, 2007). While the previous research regarded the influence of beliefs as a bias, the new paradigm puts such subjective degrees of belief center stage. The positions of the leading figures in the field are represented in chapter 4.5 by Chater and Oaksford and chapter 6.2 by Over and Cruz (both in this handbook).

These developments in psychology go back to another milestone on the normative side of rationality: the development of probability theory by Reverend *Thomas Bayes* and *Frank P. Ramsey*. Thomas Bayes (1764/1970) is often credited with having established an important theorem in his groundbreaking work (but see Stigler, 1983), a theorem showing how to infer new posterior probabilities from old prior probabilities and the evidence. Today, this framework is called *Bayesianism* and is very popular in philosophy and psychology. An important later step was the work by Ramsey, who—in his 1926 essay "Truth and Probability" (in Ramsey, 1978)—strongly emphasized the subjective interpretation of probability. Unfortunately, Ramsey's writings are few, since he died in 1930 at the age of 26. However, the widespread popularity of Bayesianism in epistemology, philosophy of science, statistics, and cognitive psychology owes much to Ramsey, who made degrees of belief central to epistemology. Ramsey also insisted that degrees of belief must be conceived as subjective probabilities and even proposed, for the first time, a method for measuring those subjective degrees of belief. In all fairness, though, *Bruno de Finetti* (1937) must be mentioned as another pioneer in subjective probability theory. (See also Gillies, 2000; Krüger, Daston, & Heidelberger, 1987; Krüger, Gigerenzer, & Morgan, 1987.)

Nowadays, most psychological research in the "new paradigm" is within this Bayesian framework (see the chapters in section 4 of this handbook).[7] Slowly, however, the insight gains strength that subjective degrees of belief do not need to be modeled as probabilities. Models can also build on other systems for representing degrees of belief and their revision. These options will be richly dealt with in this handbook from both a normative and a descriptive point of view (see chapter 4.7 by Dubois & Prade and the chapters in section 5 of this handbook).

Ramsey was also important for another paradigm of rationality research, namely decision theory, which builds on probability theory and assumes decisions and actions to be guided by utilities. In his paper "Truth and Probability," he proposed for the first time a method for simultaneously measuring subjective probabilities *and* utilities, and thus laid the foundations of standard decision theory as it is taught today. This measurement method also offered novel ways of justifying decision theory, which later gained ever more importance (see chapter 8.2 by Peterson, this handbook). And this was only possible by interpreting probability in a subjectivist way. Nowadays, standard decision theory is the most widely accepted account of practical rationality and decision making. It serves as the basis of economics and of much of the social sciences, at least as far as it is pursued within the *rational choice paradigm* (see sections 8 and 9 and chapter 10.4 by Raub, all in this handbook).

Another milestone in rationality research was the work of *Herbert A. Simon* (Simon, 1957, 1959), who was as strongly influenced by standard decision theory as Wason was inspired by Popper's falsificationism. Simon criticized standard decision theory for its idealizations and claimed that human decision making is limited in various ways: by incomplete knowledge, by imperfect memory, by restricted capacities of representation and computation, and so on. Thus, he developed the concept of *bounded rationality*, which has by now ramified into various research fields. Simon was also the first to suggest that human behavior is determined by *heuristics* rather than by a decision calculus. He proposed the *satisficing heuristic*, according to which the search for options is stopped as soon as an option is found that reaches a preset achievement level such that no further optimization is needed. Simon's work first radiated into economics but was soon recognized in cognitive psychology, where it initiated much experimental work (Gigerenzer & Selten, 2002). As a pioneer of cognitive-oriented artificial intelligence (together with Allen Newell, he invented the general problem solver [GPS]), Simon was also a founding father of cognitive science (Newell & Simon, 1972), where rationality research looms large as well.

The idea of heuristics was extended in the work of *Amos Tversky* and *Daniel Kahneman* (1974), two psychologists who were also interested in economics. Their experiments showed that human judgment does not conform to the rules of the probability calculus but is rather guided by various heuristics, cognitive rules of thumb that are often helpful because they save cognitive

resources, but may also lead to systematic deviations from the norms of mathematical probability and decision theory. This approach was further developed in the program of *ecological rationality* by *Gerd Gigerenzer* and coworkers, although in this theoretical framework, heuristics have a less negative connotation. In fact, these researchers argue that heuristics are highly adaptive and rational (Gigerenzer, Hertwig, & Pachur, 2011). Kahneman and Tversky (1979) also developed *prospect theory* and presented many experimental findings showing that this psychological theory is empirically more adequate than standard decision theory from economics. Today, prospect theory and the *heuristics and biases program* can be found in most textbooks of psychology and behavioral economics (see chapter 8.3 by Glöckner and chapter 8.5 by Hertwig & Kozyreva, both in this handbook).

It should be noted, though, that most of the reported theories somehow oscillate between the normative and the descriptive view on rationality. Piaget was perhaps most explicit in using logic as both a descriptive and a normative framework for human rational reasoning. In contrast, standard decision theory was normatively motivated, and the hope was that it would not be too strongly empirically idealized. Kahneman and Tversky decidedly took the empirical point of view but also uncovered the many empirical inadequacies of normative decision theory. Simon's work was more ambiguous in this respect, as it attempted to define what rationality normatively requires, given the constraints of the cognitive system and the environment. Ecological rationality mirrors a similar view on human rationality. We shall return to these issues in section 4 of this introductory chapter and repeatedly in this handbook. Of course, these few milestones cannot stand in for a comprehensive history of our topic, not even for the past 150 years. But they should prepare the readers to better follow the further structure of this chapter.

### 2.2 Basic Concepts of Rationality Assessment

Let us turn to some basic concepts of rationality assessments. We already mentioned that the history of the terminology of rationality research does not provide a stable starting point. Neither is it advisable to start amid the current discussion—there is too much theory-ladenness. So, where should we start? Perhaps with ordinary language, which, albeit imperfect, is the best initial starting point we have: it preserves the insights of our ancestors, if also their confusions. Hence, let us briefly look at how we talk about rationality in our everyday life. Which are the things we call reasonable, rational, and so on? And what is it about those things that makes us call them rational?

First, there are many things that are not subject to our rationality assessment, for instance, the weather. We may call them *arational*. They are outside of our consideration. Many things, though, are assessable as *rational*—and these kinds of things may also be assessed as *irrational*. For instance, people are upset about the irrationality of certain traffic regulations and road constructions, and the presentation of products in a supermarket may be very reasonable, at least from the manager's point of view. Still, two categories stand out as primary objects of our rationality assessments: *beliefs* or *epistemic attitudes* in general[8] and *actions*. This is why we carefully distinguish theoretical rationality—which is concerned with beliefs—and practical rationality—which is concerned with actions (see section 3 of this introductory chapter).

Other things may then be called rational or irrational in a derived sense, insofar as they are caused by, or causally connected to, those primary objects, such as the road constructions or the presentation of products in the grocery store. Of course, we also assess persons and other animals as rational or irrational, insofar as they act and believe in a rational or irrational way. Whether we can evaluate emotions as rational or irrational as well is a matter of much controversy (see, e.g., de Sousa, 2011; D. Evans & Cruse, 2004; Helm, 2001). We will return to this question only briefly in section 8 of this introductory chapter. Our overall focus is on what we just called the primary objects. We will explain below that intentions and desires should be included as well.

We should emphasize, though, that it is *actions* and not behavior in general that are judged as rational. Psychologists typically say that action is behavior controlled by the mind. Philosophers typically say that action is intentional behavior (i.e., caused by an intention). Many psychologists study unintentional behavior, which is not, or hardly, under cognitive control. Sneezing, or salivation in the presence of food, is an arational, unintentional behavior. Sure, there is also intentional sneezing, which may or may not serve its purpose and may hence be assessed as rational—but this is fake sneezing. This restriction of rationality assessments to actions is important, but also delicate. Actions may be characterized as intentional and controlled even though the intention need not be consciously conceived and the control need not be actually exercised. So, actions may very well be purely habitual. Certainly, boundaries are vague. Moreover, the intentionality of a behavior is only a necessary, not a sufficient, condition for its being an action.[9]

Let us look a bit more carefully at our rationality assessments. A noteworthy point is that our judgments can take a *relative* and a *categorical* form.[10] In certain situations,

your belief that it will rain this afternoon may seem obviously unreasonable. This would be a categorical judgment. However, given that your only information is some outdated weather forecast, your belief may be reasonable after all, namely, relative to this information. Therefore, carrying an umbrella for your shopping tour may be categorically called unreasonable, but relative to your unreasonable belief, it is perhaps not so unreasonable after all.

Similarly, in psychological experiments, volunteers are typically instructed to assume that the premises of a reasoning problem are true. In the experimental setting, the premises are simply given and serve as input to be taken for granted by the participants. Thus, such experiments often assume a *closed world*, in which only the information from the premises is relevant for the conclusion. Relative to the closed world of the experiment, it would be reasonable to reach the conclusion. However, the assumption may be wrong. If the participant reaches a different conclusion, this need not be unreasonable—it may be reasonable relative to the participant's prior knowledge, which is not contained in the closed world but still taken to be relevant by the person. Psychologists then tend to speak of the *belief bias* (J. St. B. T. Evans, 1993), but in fact, such responses just demonstrate that human reasoning is often defeasible and, thus, does not conform to the monotonicity principle of classical logic, according to which no valid inference can be turned invalid by simply adding premises. Many experimental findings and different theories of defeasible reasoning are described in chapter 5.4 by Gazzo Castañeda and Knauff (this handbook).

Relative judgments of rationality seem more basic, while categorical judgments seem derived: they are something like judgments *all things considered* or relative to the total evidence. We will yet look at how they might be derived. But let us first discuss what the relative judgments are relative *to*. In the case of beliefs, the answer seems straightforward: a belief is assessed as rational relative to the evidence the agent has, relative to her reasons or the information she has, or more generally relative to her other beliefs or the epistemic state the belief is embedded in. This raises the question whether a belief can also be rational relative to other than epistemic matters, desires perhaps, or emotions.[11] This is a difficult issue. It seems that a certain desire or emotion may be a *cause* of a belief but never a reason for it. We would disapprove of such a belief as wishful thinking. But what might be the psychological mechanisms behind such phenomena? We shall return to this point in subsection 3.3 of this introductory chapter and when we discuss the distinction between the normative and the descriptive perspective and that between output-oriented and process-oriented investigations of rationality (section 6 of this chapter).

If we accept that only beliefs can be reasons for a belief,[12] it seems clear how we arrive at categorical judgments about beliefs: a belief is categorically reasonable if it is reasonable relative to other epistemic items that are in turn categorically reasonable. So, two things may go wrong with a belief: it may be based already on unreasonable premises, or it may be inferred from reasonable premises in an unreasonable way.[13] Again, both can happen in psychological experiments: participants can either deviate from a logical conclusion because they do not accept the premises or proceed from other premises, or they take the premises for granted but commit errors in the reasoning process itself.

So much for how we assess the rationality of beliefs. But what about actions, the other main class amenable to rationality assessments? Somehow our actions have to serve our goals according to our own lights. That is, the rationality of our actions is assessed *relative* to our beliefs and desires. This is the familiar and highly important notion of *instrumental rationality*, which is at home everywhere, in philosophy, psychology, and economics. "Beliefs and desires" must not be taken literally here. In fact, the term "beliefs" stands for an entire epistemic or cognitive complex, while "desires" stands for the aggregate of volitional, optative, conative, or buletic attitudes ("pro-attitude" is a more neutral but less familiar term). In ordinary language, we speak of drives, wishes, wants, interests, inclinations, goals or aims, norms and values, and the like. Psychologists distinguish between relatively stable behavioral dispositions, such as the needs for achievement, affiliation, or power, and short-term driving forces in a particular situation (Heckhausen & Heckhausen, 2018). However, let us here stick to "desire" as a generic term for this plenitude of concepts and classifications for the driving forces for actions.

Is there also a categorical rationality assessment of actions? Clearly, an action is categorically irrational if it derives from the given beliefs and desires in an irrational way. However, it may be categorically irrational even if it derives rationally from the given beliefs and desires. This would be the case if the beliefs or desires are themselves irrational. We have already seen how beliefs may be judged to be categorically irrational. So, this judgment immediately enters the practical assessment of actions. But are desires themselves also subject to a rationality assessment? Instrumental rationality refuses to judge desires in this way. It simply takes them as given. Recall the famous dictum of Hume (1739–1740/1975c, p. 415) that reason is and ought only to be the slave of

the passions. This leaves no room for a rationality assessment of the passions themselves. If so, the rationality of actions would ground only on the rationality of beliefs.

However, the case is a bit more complicated. We need to distinguish between *intrinsic* desires, the objects of which are desired in themselves, and *extrinsic* desires, the fulfillment of which only serves some other (extrinsic or intrinsic) desires. A similar distinction in moral philosophy is that between entities good as such, or as an end, and entities good as means. Relatedly, psychologists distinguish between extrinsic and intrinsic motivations. An action is extrinsically motivated when it serves to avoid something unpleasant or to get a return, a good grade, money, or the like. By contrast, it is intrinsically motivated when it is rewarding or enjoyable for its own sake, such as solving a puzzle or editing a book.

With this distinction at hand, we can extend rationality assessments also to desires: extrinsic motivations or desires may be irrational given certain intrinsic motivations or desires. If I want to become a millionaire (in order to satisfy my intrinsic desire for a carefree life) and *therefore* want to become a philosopher, then something is wrong with me. Either I have weird beliefs about the profitability of philosophy, or my leaning toward philosophy is simply irrational. That much we can say even from the point of view of instrumental rationality.

This points to the real issue. Can we assess as rational also *intrinsic* desires, motivations, or values? Many analytic philosophers still deny this and accept Hume's position. They would only grant that actions and (intrinsic) desires can be judged as moral or immoral. However, the majority disagrees. A crucial point is that it becomes difficult here to separate considerations of rationality and of morality (see chapter 12.1 by Fehige & Wessels, this handbook). The difficulty already commences with Kant, for whom the *categorical imperative* is the first a priori principle of practical reason. As such, it is a principle of rationality. But at the same time, it is the highest moral principle (see also chapter 1.1 by Sturm, this handbook). Max Weber (1921–1922) famously developed a concept of value rationality as opposed to instrumental rationality. These threads of Kant and Weber were taken up by many, notably by Jürgen Habermas (1981), who sees practical rationality as going beyond instrumental rationality, for instance, in communication, which is characterized by treating one's fellows as ends in themselves (in chapter 10.3 in this handbook, Meggle argues that communicative rationality is just instrumental rationality). Habermas (1973) criticized Weber's concept of instrumental rationality as a subordination to a fundamentally opportunistic lifestyle, while proper practical rationality also means becoming aware of and changing one's own role in society. Quite different attempts to make sense of rational (and not already moral) assessments of intrinsic desires may be found in Nozick (1993), H. S. Richardson (1994), and Kusser and Spohn (1992). However, this is an obscure topic in philosophy and not treated in cognitive psychology. Therefore, we will not further pursue it in this handbook. In any case, it is important to be clear about the ways in which our rationality assessments are relative and can be categorical.[14]

Having thus elucidated the objects of rationality judgments and their relational character, we have still not addressed the most essential question: What precisely makes one object—belief, desire, action—rational relative to other objects? What is the content of this relation? In short, what *is* rationality? In a way, the entire handbook is about this relation. Ordinary language is of little help here. We might say that the things relative to which some belief or desire or action is rational are the *reasons* for that belief or desire or action. But what is this reason relation? We certainly have a good intuitive understanding of it. We might say that some kind of *reasoning* or *inference* leads from the former to the latter. But what are the mechanisms governing such inferences? Questions like these are omnipresent in this handbook.

David Marr (1982) has famously introduced the distinction between the computational, the algorithmic, and the implementational level of explaining cognitive phenomena, a distinction governing cognitive science ever since. On the *computational* level, the goal of a computation is set: what is its function, and what is to be achieved by it? The *algorithmic* level inquires *how* this computational goal is reached: which processes or algorithms are used to reach the goal defined on the computational level? The *implementational* level then addresses how these processes are physically realized. Searle (1992) argued that mental processes can be explained only on the basis of the biological, not just the functional, properties of the brain, a position he calls "biological naturalism." By contrast, functionalist philosophers of mind from Putnam (1967/1975) and Fodor (1975, 2000) onward have claimed that the implementational level is irrelevant for understanding the functioning of the mind: it is largely irrelevant whether the processes on the algorithmic level are realized in a biological system such as the human brain or by the silicon chips of a computer. And there are even modern dualists, like Chalmers (1996) and others, who argue that some mental characteristics cannot be reduced even to functional characteristics. We briefly discuss this topic in section 7 of this introductory chapter about the preconditions of

rationality. A more detailed description of the cortical basis of human rationality is given in chapter 1.4 by Goel (this handbook). However, in the main, this handbook sticks to the computational and the algorithmic level of description, because we think that these are the most relevant levels if one really wants to understand what makes people rational—and sometimes irrational.

However, Marr's terminology is a bit confusing. His conceptual distinctions are immensely important, but the terms are misleading. For instance, a computation consists just in running an algorithm, and the algorithmic level also determines the kind of representations on which the algorithm works. Therefore, we here use the terminology already indicated at the end of section 1 of this introductory chapter, where we called Marr's computational level the *output-oriented* and the algorithmic level the *process-oriented* level of explanation.[15] This terminology is less biased toward computer science and also signals a slight shift of meaning. In particular, we want to make room for the idea that the computational goals are set by normative and not by empirical considerations. We will return to this issue in section 6 of this introductory chapter. But before that, we now come to the top-level distinction in most conceptions of rationality, in both psychology and philosophy.

## 3. Theoretical and Practical Rationality

Let us start with an example. Many current public and scientific debates revolve around the increased mortality of honeybees. This Sunday, there is a demonstration against the use of the herbicide glyphosate, which is feared to be harmful to animals, including bees. Should you participate? If you want to make a rational decision, your thinking should proceed in two steps: in the first step, you should weigh the arguments that speak for or against the assumption that glyphosate kills bees, evaluate the inferences made by the supporters and opponents of this claim, consider empirical evidence from scientific research, and so on. Let us assume that, based on these considerations, you conclude that glyphosate indeed can cause increased mortality among honeybees. Now, in a second step, you have to decide whether or not you will go to the demonstration. On the one hand, you would prefer a relaxed Sunday on your sofa, it is likely to rain, and you fear that there may be some violence at the demonstration. On the other hand, you think, based on your previous conclusion, that glyphosate should be banned and that it is important to fight for nature, and many of your friends will be there too. Both considerations are important for your decision to attend the demonstration.

Of course, in daily life, these kinds of thoughts are not so clearly separated, your decision might be influenced by many other internal and external factors, and so on. Yet, our example makes clear that we must carefully distinguish two aspects of rationality. The first part of the example refers to *theoretical* rationality, which is about the rational justification of beliefs, inferences, and explanations, or of our epistemic states in general. This issue lies at the center of epistemology and constitutes one of the main topics in the philosophy and psychology of rationality. Thus, the Wason selection task and the large amount of studies on human conditional, syllogistic, probabilistic, counterfactual, causal, or relational reasoning are concerned with theoretical or epistemic rationality. The so-called Linda problem, which is almost as famous as the Wason selection task (see chapter 1.2 by Evans, chapter 4.3 by Merin, and chapter 8.5 by Hertwig & Kozyreva, in this handbook), is also concerned with the theoretical rationality, or irrationality, of human reasoners.

The second part of the example above is concerned with *practical* rationality, which is about assessing actions or pro-attitudes in general. Here, psychologists explore how people choose between different alternatives that have different values for them. Thus, it is concerned with what people decide in a particular situation where different options are considered.

John Searle, who is just as important for philosophy as for the cognitive sciences, expressed the distinction in terms of "direction of fit": theoretical rationality is about how to *represent* the world, how to make our mind correspond to it, whereas practical rationality is about how to *shape* the world, how to make it correspond to our mind (Searle, 1983). The distinction can also be made with respect to the kinds of reasons that can be adduced for the attitudes in question. Beliefs, or epistemic attitudes in general, can only be justified with reference to further epistemic elements—knowledge, beliefs, evidence, perceptions, or testimony—while the reasons for actions and intentions lie not only in those epistemic elements but also in our motives and desires or pro-attitudes. So, to account for theoretical rationality, we need to talk only about epistemic matters, whereas accounting for practical rationality also requires talking about motivations, aims, values, and so on.

Some psychologists might have problems with this sharp separation of theoretical and practical rationality. Their main argument is that in daily life, the functions of reasoning and decision making are intertwined. Both kinds of rationality, theoretical and practical, are strongly interlinked, because reasoning typically serves good decision making. This is also substantiated in the

account of *utility conditionals* in chapter 6.4 by Bonnefon (this handbook). Recently, the proponents of the "new paradigm" in psychology have argued in this direction, as they consider subjective psychological value, or utility, and social pragmatics as important for reasoning. Accordingly, this approach aims to integrate the psychology of reasoning with the study of decision making (Elqayam & Over, 2013).

It is certainly true that we can find reasoning on both sides. There is even an influential philosophical position to the reverse effect, namely, that we can find decision making on both sides: we decide not only what to do but also what to believe. This doctrine, *doxastic voluntarism*, goes back to Descartes (1641, Fourth Meditation). It tries to understand belief formation as deciding which beliefs best fulfill our epistemic aims, where truth is the central epistemic aim, but may be accompanied by other aims. The result is what is called *epistemic decision theory* (Konek & Levinstein, 2019; Levi, 1967). We can even say that participants in reasoning experiments decide between conclusions having different values—a reasoner may value the valid conclusion more than the invalid one. In these psychological and philosophical lines of thought, there is no clear distinction between theoretical and practical rationality.

Here we disagree. In our view, practical rationality *presupposes* theoretical rationality. We think that we can study theoretical rationality independently from practical rationality but not the other way around. The reason is that to account for theoretical rationality, we need to talk only about epistemic matters, whereas accounting for practical rationality also requires talking about motivations, aims, values, and so on. The activity of reasoning is necessary on both sides. However, theoretical reasoning proceeds from factual or empirical premises to factual or empirical conclusions. Practical reasoning takes motives or values or pro-attitudes as additional premises and arrives at practical conclusions (intentions or actions). Of course, these are different reasoning tasks requiring different cognitive processes. In the following two subsections, 3.1 and 3.2, we unfold the distinction in more detail. In subsection 3.3 of this introductory chapter, we will return to the issue of what may be special about theoretical rationality.

### 3.1 Theoretical Rationality

Theoretical rationality is often also called "epistemic rationality," even though, as mentioned in subsection 2.2 of this introductory chapter, "doxastic rationality" would be the more appropriate term. There, the fundamental issue concerns the rational form of epistemic attitudes, states, or processes. Concerning this shape, the first question is which entities those attitudes are about, that is, what it is that we believe, take to be plausible, and so on. Usually, these are taken to be linguistic entities. And usually, these entities are taken not to be sentences or utterances themselves—otherwise the Frenchman and the Spaniard would not be able to share beliefs—but rather the *contents* or meanings of sentences and utterances. Philosophers and cognitive psychologists call the latter "propositions," while there is some disagreement about a suitable characterization of propositions as objects or contents of epistemic attitudes.[16] This handbook is silent on those issues—deliberately, because treating them fairly would mean engaging deeply the most fundamental questions of the cognitive sciences and the philosophy of mind and language, which are not our topic. The consequence of our decision not to enter into issues of mental contents is that this handbook remains quite vague about the objects or contents of our beliefs and, when it comes to practical rationality, about the contents of our motives and desires.

The next important point pertains to the many possible forms of epistemic states. Basically, these forms may be conceived in a *qualitative*, a *comparative*, and a *quantitative* way. In many situations in everyday life (and in the psychological lab), people just believe or disbelieve something, for example, that glyphosate is harmful to bees. For many decades, psychologists have mainly explored inferences based on such qualitative beliefs. Usually, participants had just two response options: belief or disbelief, entailed by the premises or not, logically valid or invalid, yes or no, and so forth. Today we have a relatively clear empirical understanding of how people deal with such inferences, although there is less agreement regarding the underlying cognitive processes (J. St. B. T. Evans et al., 1993; Johnson-Laird & Byrne, 1991). Thus, an epistemic state (of a certain person at a certain time) may be qualitatively characterized simply as a set of beliefs, a set of contents that are accepted (or "endorsed" or "maintained") by the person at that time. If this set may be an arbitrary set, philosophers call it a *belief base*; if it is to satisfy the fundamental and strong but contested rationality postulates of consistency and deductive closure, it is called a *belief set*. Philosophers also discuss weaker rationality requirements. This conception is unfolded in several chapters of this handbook (see in particular chapter 5.1 by van Ditmarsch and chapter 5.2 by Rott).

However, this characterization neglects that beliefs usually come in degrees. Our beliefs are often more or less uncertain. For instance, you think that glyphosate probably kills bees, but you are not certain. We have a

very rich vocabulary for describing this *quantitative* level; we speak of probability, plausibility, believability, uncertainty, credibility, likeliness, and so on. In rationality research, the dominant quantitative account by far is *Bayesianism*, the doctrine that degrees of uncertainty or belief are subjective probabilities. This holds for psychology just as much as for philosophy (Jeffrey, 1992; Oaksford & Chater, 2007). However, one should observe that the label "Bayesianism" is only legitimate when these degrees conform to the axioms of mathematical probability theory.

Probabilistic thinking has evolved for about 350 years (cf. Hacking, 1975).[17] It has pervaded most scientific disciplines (Krüger, Daston, & Heidelberger, 1987; Krüger, Gigerenzer, & Morgan, 1987; for a philosophical introduction, see Gillies, 2000), and so it comes as little surprise that Bayesianism has become a very strong paradigm in psychology and philosophy as well. For this reason, we devote the entire section 4 of the handbook to it. From the Bayesian perspective, the qualitative characterization of belief states looks hopelessly vague, if not useless. Bayesianism has thus developed imperialistic tendencies, claiming that it is the *only* basic characterization of epistemic attitudes. If so, this would bring laudable unification to the field. However, this imperialism has come under pressure from two sides.

One reason is that one cannot simply discard the qualitative level, as advocated by the "radical probabilism" of Jeffrey (1992). The qualitative notion of belief seems too deeply entrenched in everyday discourse. So, the only option is to somehow reduce the one to the other level. However, there is no good reductive account. How weak may a belief be and still count as a belief? Can we accept the so-called Lockean thesis that something is believed if and only if the degree of belief in it is above a certain vague and perhaps contextually given threshold, just as we might say that a man is tall within the entire population if his size is above a certain threshold, say 6′3″, and tall among basketball players if he is above 7 feet? In fact, the relation between belief and probabilities is currently the matter of many controversies.[18]

The second reason why it is not cogent to represent degrees of belief as probabilities is that the past 50 years have seen a plethora of alternative proposals (some of which have earlier precursors). An important driving force was artificial intelligence, which needed algorithms for uncertain reasoning, found probability theory infeasible, and hence looked for alternatives. In the meantime, Markov and Bayes net theorizing have much improved the computational manageability of probabilities (Pearl, 1988). However, alternatives are available (Halpern, 2003). The need was also raised from the psychological and the economic side, which found subjective probabilities descriptively wanting. Even the theory of mental models, which for a long time was dominated by binary classical logic, has more recently developed approaches for how to deal with degrees of belief without using Bayesian probabilities (Hinterecker, Knauff, & Johnson-Laird, 2016; Johnson-Laird et al., 2015; Johnson-Laird & Ragni, 2019). This increases justificational pressure on the normative side—why this rather than that mathematical format?—as well as on the descriptive side—why model human uncertainty in this rather than in that way? Discussions of such alternatives may be found in chapter 4.7 by Dubois and Prade; chapter 5.3 by Kern-Isberner, Skovgaard-Olsen, and Spohn; chapter 8.3 by Glöckner; and chapter 8.4 by Hill (all in this handbook).

Between the qualitative and the quantitative, there is, moreover, the *comparative* level, where one proposition is taken as more plausible or credible or certain than another. For instance, you may take it to be more credible that glyphosate harms bees than that sugar does. One may even take the comparative level as basic, since it also allows one to state the threshold idea of qualitative belief and certainly underlies any quantitative measure of uncertainty (like one may base judgments of tallness as well as measurements of height on the relation "taller than"). Indeed, one may wonder which formal properties of those plausibility comparisons lead to which numerical measure. Again, this opens a large space of normative dispute as well as issues of empirical adequacy. In Knauff (2013), a comparative level has been developed for beliefs about spatial relations, which also plays a role in chapter 13.2 by Ragni and chapter 13.3 by Knauff (both in this handbook). However, this comparative level is not further developed in this handbook.

So far, our remarks about how to conceive of epistemic states were made only from a *static* or *synchronic* perspective. They have dealt with how epistemic states are at a given moment. This is basic—and only preparatory for taking a *dynamic* or *diachronic* perspective. It is important, we think, to distinguish two different aspects of the epistemic dynamics, *internally* and *externally* driven dynamics. Both are subject to considerations of empirical and normative adequacy, and they lead to different kinds of dynamic issues, as we will describe now.

There is, first, an *internally driven dynamics*, which is, roughly, brought about by *thinking*. It runs on the process-oriented (or "algorithmic") level. As indicated at

the end of subsection 2.2 of this introductory chapter, this is a crucial issue for cognitive scientists but also for logicians. It is one thing to analyze the formal structure of what a reasoner receives as input—typically the set of premises and their formal structure—and then to look at the output the reasoner generates—typically a conclusion that is evaluated as (more or less) justified according to some normative standard. However, the history of psychology shows that approaches occupied only with input–output associations have limited explanatory and predictive power. It is another and much more demanding thing to understand how the input to the human cognitive system is *processed* and why this leads to the observed output, for instance, a particular conclusion or belief. What mental machinery, involving which cognitive operations, lies behind these internally driven dynamics of human rationality? Here, the main questions are: On the basis of which mental processes are people *competent* to reason rationally? Why are these processes sometimes *error-prone*? And how do reasoning processes *interact* with other cognitive and noncognitive psychological mechanisms? This is a more dynamic description of the main questions that have occupied the psychology of reasoning since its very beginning (Wilkins, 1928; Woodworth & Sells, 1935). The most influential theories from the psychology of reasoning are sketched in sections 2 and 4 of this handbook, which compare normative and descriptive theories of rationality in reasoning.

The logical side also deals with a multitude of issues. There are not only actual but also many potential reasoning processes, which may or may not be followed and which may or may not be correct (as is presupposed by speaking of errors). For instance, logic is full of different (but equivalent) sound calculi or proof procedures, whether or not they are actually used. In a qualitative picture, reasoning is, above all, logical inference. Piaget would have said the same about psychology. It is amply clear, though, that logical inference must not be restricted to deductive or monotonic inference. The logical zoo has become quite diversified, and there is a growing field of nonmonotonic logic, default logic, and defeasible reasoning. This field is partially represented in sections 5 and 6 of this handbook. Each logic comes with its own sound procedures or derivation rules and possibly with a semantic justification. And each of them may or may not be adequate for describing how the human mind works, and may or may not be approximated by mental models or heuristics. In a quantitative picture, reasoning is first and foremost probabilistic reasoning. Again, though, one must emphasize that all the other quantitative formats come with their own accounts of reasoning. Again, this opens a rich field of normative and empirical assessment.

However, this does not exhaust the dynamic perspective. There is, second, an *externally driven dynamics*. This is about how epistemic states change under the influence of external input: experience, perception, learning, or information in general. Again, one may say that this is about drawing inferences from this input. Cognitive psychologists have developed several descriptive theories of how people account for new information, how they detect and resolve inconsistencies with prior beliefs, how they take into account the order in which new information comes in, how this affects the person's entire set of beliefs, how they consider the trustworthiness and reliability of the source of new information, and how people deal with the fact that new information is inconsistent with their prior beliefs (Elio & Pelletier, 1997; Johnson-Laird, Girotto, & Legrenzi, 2004; Politzer & Carles, 2001; Revlin, Cate, & Rouss, 2001). Such theories are either motivated by approaches from artificial intelligence and philosophy, for example, principles of minimal change or epistemic entrenchment (Alchourrón, Gärdenfors, & Makinson, 1985; Gärdenfors, 1984, 1992), or they have genuine roots in psychology, for instance, when people consider the trustworthiness of information sources or try to explain conflicts between prior beliefs and new facts (Khemlani & Johnson-Laird, 2011; Wolf, Rieger, & Knauff, 2012).

While these process-oriented theories seek to reconstruct how the cognitive system deals with new information, we can also take an output-oriented perspective (Marr's computational perspective) on the externally driven dynamics of belief. This is not about how the input is processed, but about how the goals of computation change through the input—they have their dynamics as well. Put plainly, it is about *what* to believe upon receiving some data and not about *how to arrive* at it. According to Spohn (2012, chapter 1), this problem of revision is tantamount to the venerable problem of induction. Above, we mentioned the limitations of looking only at the input–output associations. However, this must not distract from the fact that these associations have their own dynamics, which is not treated by focusing on the internally driven dynamics but rather needs to be studied on its own. The traditional probabilistic account of this issue consists in learning by conditionalization or, equivalently, by Bayes' theorem; it has by now been considerably refined. However, each

epistemic format comes with its own learning theory. Philosophers tend to emphasize the externally driven dynamics; it is addressed in chapter 4.1 by Hájek and Staffel, chapter 4.2 by Hartmann, chapter 5.2 by Rott, and chapter 5.3 by Kern-Isberner, Skovgaard-Olsen, and Spohn (all in this handbook).

Of course, the opposition between internally and externally driven dynamics is not as clear-cut as it may seem. For instance, suppositional reasoning also belongs to the internally driven dynamics, because a person does not require external input in order to work out what she *would* believe given a certain supposition. Then, however, there seems to be a close connection between internally and externally driven dynamics. Isn't what we accept under a given supposition precisely that which we would accept after getting the information (through whatever channels) that the supposition actually holds? This is a topic that has been intensely discussed under the heading "supposing vs. updating" (see Spohn, 2012, chapter 9; Zhao et al., 2012), but we cannot go into the details here. We just want to emphasize once more that all theories of the dynamic aspects of rationality can be assessed in a normative and a descriptive dimension. We return to this in section 6 of this introductory chapter.

### 3.2 Practical Rationality

We have quite extensively discussed theoretical rationality. Many points, however, apply not to theoretical rationality specifically but similarly to practical rationality, that is, to the rationality of actions or intentions and motivational attitudes in general. Recall that practical rationality presupposes and thus includes theoretical rationality. For this reason, we can carry over most of the conceptual distinctions we have already made, and we can be much briefer even though the topic is wider. The topic is wider because we now deal with a more comprehensive conceptual field. We now deal not only with epistemic attitudes but also with decision making and motivational matters, which we have already unfolded in subsection 2.2 of this introductory chapter.

A first issue of practical rationality is whether it is legitimate to lump together drives, wishes, motivations, wants, goals, aims, norms and values, etc., as the topics of practical rationality. In a theoretical spirit, many philosophers tend to do so by simply speaking of desires, and many economists do so by measuring all of them on a single utility scale. Basically, they treat drives and motivations in the same manner as other-regarding preferences and moral obligations. However, other philosophers, economists, and psychologists emphasize the differences between the various kinds of such pro-attitudes. We do not want to take a stance here, because in the end, it is a theoretical issue whether or not the various kinds of desires work according to different theories. Most psychologists would probably argue for such a separation. Perhaps the "imperialism" of utility theory here is as problematic as that of Bayesianism in theoretical rationality. Still, here we ignore these potential differences, as our conceptual points about desires presumably apply across the board.

So, let us focus on the same questions as we already did on the side of epistemic rationality. The first question is again which entities those desires are about, that is, what it is that we want, desire, etc. On the epistemic side, we already mentioned some problems regarding propositions as the contents of beliefs. Now, in practical rationality, the issue is even more obscure with respect to desires and motivational attitudes. If we treat the contents of desires as propositions, this might not fully capture our ordinary talk. We may also say that the entities we desire are objects or goods or actions, and so on. There is perhaps a unification: "I want this bike" is the same as "I desire that I possess this bike," and "I intend to do *a*" is the same as "I intend that I perform action *a*" (Jeffrey, 1965, chapter 4). One may think that objects, goods, and actions are clearer contents of desires than propositions. But this is not true. For instance, I love Wonder Woman, but she does not actually exist. So, what is the object of my desire here? Another example: I may want to meet Dr. Jekyll but not Mr. Hyde. This is not possible, because both are the same person. There are no two different objects here that I could desire to meet or not to meet. Philosophers then tend to say that desires are about intentional objects or about objects under a description. And they prefer to consider the contents of desires to be propositions because propositions can contain objects in such a modified understanding.[19] We are aware that this may sound weird to people who are not usually involved in such theoretical considerations. Hence, we better avoid delving into those problems and simply stick to propositions, for ease of exposition.

The next distinction that we made within the realm of theoretical rationality was the one between qualitative, quantitative, and comparative shapes of mental states. We have the same distinction on the side of practical rationality. In a qualitative framework, propositions can simply be or not be desired, wanted, mandatory, or the like. Recall, though, that we need to account for the interaction between desires and beliefs or epistemic attitudes in general, because we must respect the basic distinction between intrinsic and extrinsic desires, which we introduced in subsection 2.2 of this introductory chapter. Extrinsic desires involve such interaction, since

they are directed at propositions somehow believed to be conducive to intrinsically desired propositions. Within a purely qualitative framework, however, it is very difficult to account for this interaction (for a proposal and further references, see Spohn, 2020).

So, just as on the epistemic side, where we talked about degrees of belief, we should allow desires, too, to come in varying strength. If we do so, we again have two options. We can either express the strengths of desires quantitatively, or we can express them merely on a comparative level. In fact, the comparative level is more prominent on the practical than on the epistemic side. Thus, we enter the large field of preference theory. The relation between the comparative and the qualitative level has been intensely studied in the *theory of revealed preferences* (Samuelson, 1938), which originally served as a way of operationalizing preferences by means of choice behavior. The comparative and the quantitative level are also closely tied together, namely, by the famous von Neumann–Morgenstern utility theory (von Neumann & Morgenstern, 1944, chapter I.3), which showed how we can measure utility functions on the basis of preferences in a sufficiently unique way. Before this theory, numerical strengths of desires seemed to be elusive things of doubtful scientific status, but afterward, talk of numerical utilities seemed legitimate (for all this, see chapter 8.1 by Grüne-Yanoff, this handbook).

Thus, talk of utility functions measuring motivational strengths by real numbers is by now well established and even dominant, at least in economics and philosophy. The great advantage of utility functions is that they combine so smoothly with probability measures, the prevalent numerical representation on the epistemic side. This combination culminates in the principle of maximizing expected utility, the fundamental rationality principle of standard decision theory (see chapter 8.2 by Peterson, this handbook). No wonder there is a tremendous debate about the normative foundations of this quantitative framework and its empirical adequacy. In the past decades, three cognitive and social scientists—Herbert Simon, Daniel Kahneman, and Richard Thaler—were awarded the Nobel Prize in Economics for showing how and explaining why people deviate from these norms of standard decision theory. This stimulated a number of alternative accounts, which are represented in chapter 8.3 by Glöckner, chapter 8.4 by Hill, and chapter 9.4 by Dhami and al-Nowaihi (all in this handbook). Today, this field is quite variegated and heterogeneous. It is sometimes called "qualitative decision theory," where "qualitative" does not have the meaning from above, but only signals that the field moves to a less fine-grained way of description than offered by probabilities and utilities, although it does not belong to the qualitative level as we use the term here.

Let us now come to the next distinction we made on the side of epistemic rationality. Is there also an *internally and an externally driven dynamics* of desires, wishes, motivations, and so on? Yes, and in principle we can make the same distinctions, although the algorithms of practical reasoning certainly differ from those of theoretical reasoning. On the side of the internally driven dynamics, it is tempting to interpret decision theory as an algorithmic theory about practical reasoning processes at the process-oriented level. Decision theory provides general rules for calculating expected utilities and thus for determining which options maximize expected utility.[20] However, thus interpreted, decision theory seems descriptively inadequate—people apparently do not follow these rules and do not do any of these calculations. Hence, we may prefer to locate decision theory on what we call the output-oriented level. In the attempt to be more adequate on the process-oriented level, cognitive psychologists have developed various accounts of so-called fast and frugal heuristics and bounded rationality (see chapter 8.5 by Hertwig & Kozyreva, this handbook). However, the goal of developing psychologically plausible algorithmic theories of practical rationality is still daunting, probably even more demanding than in the area of epistemic rationality.

What about the externally driven dynamics? In practical rationality, this becomes relevant as soon as we conceive decision theory not as a process-oriented but as an output-oriented theory that only states which intention should result from the given utilities and probabilities. Here again, epistemic rationality comes into play. Desires and intentions are also affected by the externally driven dynamics of epistemic states. We learn—and thereby change our extrinsic desires and indeed our intentions. This is a very important point, which many people forget when talking about desires and practical rationality: we often wait for, or indeed seek, information, and depending on what we learn, we intend to do either this or that. For instance, once I have learned the weather forecast for this afternoon, I may change my desire to go to the demonstration in a T-shirt. There are sophisticated theories elaborating on *strategic rationality*, that is, on plans or strategies optimally responding to various possible pieces of information (see Raiffa, 1968).

So far, we have only discussed the dynamics of extrinsic desires or intentions. Is there also an externally driven dynamics of intrinsic motivational states? Sure. Psychologists have extensively studied this, although not under the heading of rationality. In motivation research,

our actions are seen as either driven by relatively stable behavioral dispositions or triggered by stimuli from the environment, for instance, when you see appetizing food or a physically attractive person (Heckhausen & Heckhausen, 2018). However, even the relatively stable dispositions are not immutable. Our pro-attitudes change and evolve all the time, due to all kinds of influences apart from information. Their dynamics is also driven by education in general and moral education in particular, by the social environment, by fashion and advertising, by saturation, boredom, and curiosity, simply by maturing and aging, etc. So, this is usually considered an issue of descriptive theories of rationality.

However, the dynamics of desires is not only a factual matter to be studied empirically. It also raises normative issues. And many of them are of a moral kind. We want people to acquire the *right* values and try to prevent them from adopting wrong ones, whatever they are. These topics fall outside the scope of this handbook, since they are about morality, not rationality.

Another field that is not represented in this handbook, although it has to do with externally driven dynamics of desires, is *dynamic decision theory* (McClennen, 1990). It deals with what economists call "endogenous preference change" (Bowles, 1998; Loewenstein & Elster, 1992), that is, precisely with our present issue. The question there is not which changes would be rational but how to behave rationally in view of preferences changing due to whatever influences.[21] Thus, the change itself need not be assessed as rational, but when deciding, we may take such changes into account in a rational way. However, such considerations seem to require optimizing according to two different points of view, the old preferences and the new preferences. From the perspective of standard decision theory, which is about optimizing within only one point of view, this is a difficult, if not unsolvable, question. So, this field has thus far remained esoteric and will be neglected here.

### 3.3 Truth and Rationality

At the beginning of this section, we argued that it is important to distinguish between theoretical and practical rationality—a distinction that roughly matches the distinction between reasoning and decision making in psychology. But then we saw that these two types of rationality have many things in common and that theoretical rationality is a precondition of, and thus part of, practical rationality. So, why still sharply distinguish the two types of rationality?

The answer seems simple: because epistemic rationality is about the rational justification of beliefs, whereas practical rationality is about the rationality of means to an end. But this is too simple. A crucial point is that epistemic rationality is not merely about the justification of beliefs. Things are a bit more complicated because we need to distinguish between epistemic and nonepistemic justifications of beliefs. A standard example, slightly modified, comes from BonJour (1985, p. 6): a mother ought to, and does, believe in the innocence of her son even in the face of overwhelming evidence that he committed a severe crime. One may take it to be morally required that at least the mother backs her son, and given this moral requirement, she may be rationally justified in having that belief. However, given the evidence, the mother would be epistemically irrational to believe in her son's innocence. This shows that the assessment of beliefs as rationally justified may involve more criteria than merely epistemic justification. The justification of beliefs may also take nonepistemic and specifically moral forms. This is not per se a reason to call such beliefs irrational. They are, however, epistemically irrational. Theoretical rationality is restricted to the epistemic part of justification.

But then the question arises whether nonepistemic justifications of beliefs are still acceptable. We agree with BonJour (1985) and many other philosophers that the rational formation of beliefs and doxastic states is subject only to the norms of epistemic rationality. This also means endorsing the normative principle that any external considerations demanding epistemic irrationality must be rejected. It was a long historic fight to establish this principle, starting with the progress from myth (*mythos*) to reason (*logos*) in Ancient Greek philosophy, and continuing with the rise of scientific methods, with Cartesian methodical doubt (which called into question any dogmatic truth), with Kant's characterization of Enlightenment: "Have courage to use your own mind!" or with the claim of Peirce (1877) that the scientific method is the only appropriate one for fixing our beliefs. Certainly, the fight is not over.

The task, then, is to more substantially describe this narrow sense of epistemic rationality. Philosophers discuss various specific aims of our epistemic activities, and, as mentioned at the beginning of this section, some proceed to formulate a specific epistemic decision theory. Among these aims may be simplicity, systematicity, relevance, explanatory power, and perhaps even aesthetic values. For sure, though, the primary aim is *truth*. This first maxim was stated by William James (1896/1956, section VII) in the shortest possible way: "Believe truth! Shun error!" Some philosophers prefer to identify *knowledge* instead of merely truth as the aim of belief (see Chan, 2013; Williamson, 2000, chapter 11). However, in

subsection 2.2 of this introductory chapter, we already decided to leave the difference between knowledge and true belief undiscussed.

Of course, belief formation does no more than *aim* at truth. Rational belief formation in no way guarantees truth. This is part of our skeptical heritage. We may accidentally hit the truth in irrational ways. And we may arrive at false beliefs in a perfectly rational way, not only individually but also collectively. In principle, the possibility of misleading evidence and false theorizing can never be excluded. This is why philosophers say that the reasons we have for our beliefs are no more than truth-*conducive*. The point of the normative principle stated above is only that there is no alternative way to approach the truth.

It is a difficult issue whether all other epistemic goals can be reduced to the aim of truth or knowledge. Willard Van Orman Quine once famously claimed that "normative epistemology is a branch of engineering," namely, "the technology of truth-seeking." "The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed" (Quine, 1986, pp. 663–664). We doubt this (see Spohn, 2012, chapter 1) but need not expand on the issue now (see also subsection 4.3 of this introductory chapter). Moreover, it is quite unclear how to transfer the aim of truth to the other epistemic formats mentioned above. For instance, it makes no sense to call subjective probabilities true or false. A key notion discussed instead is that probabilities may be justified by their greater or lesser *accuracy* (see chapter 4.1 by Hájek & Staffel, this handbook). In fact, this point speaks in favor of the indispensability of the notion of belief, because beliefs can plainly be called true or false. However, these problems cannot distract from the fact or the norm that our epistemic activities essentially aim at truth.

But then the question is: what is truth? Aristotle's correspondence theory is the traditional paradigm. It has been amended by many modern versions: Tarski's semantic theory, deflationary and disquotational truth theories, the redundancy theory, and so on (see Kirkham, 1992; Künne, 2003). All these theories are epistemologically not enlightening and do not intend to be so. Because of this deficit, a host of further truth theories have been proposed: Peirce's and James's pragmatic theory, Habermas's consensus theory, idealistic and constructivist truth theories in various forms, the coherence theory of truth, again in various forms, and so forth.[22] Already the labels indicate that we are moving into a very controversial terrain that we cannot discuss here.

The main difficulty is that psychologists and philosophers tend to have quite different opinions and theories about the notion of truth. On the one hand, the opinions of most psychologists are shaped by countless experimental findings showing how deeply people's understanding of the world is mediated by their experience and constructed by their brains. This may suggest that the idea that we perceive the "real world" is just an illusion, a concern that can be traced back to the work of Piaget, the learning theory of Bruner (1957), and the social constructivism of Vygotsky (1978).[23] The skepticism goes even further back to the groundbreaking work of Sir Frederic Bartlett, who already emphasized the reconstructive character of the human mind and the social factors that play an important role in what we think is true of the world (Bartlett, 1932/1995).

On the other hand, philosophers could say that such findings about the constructive achievements of the brain inform us about how people come to take something to be true, but not about truth itself; they may find the empirical views of psychologists only indirectly relevant to the multitude of philosophical truth theories mentioned above. However, it is important for philosophers that the notion of truth is distinct from, and does not collapse into, the notion of belief (of *taking* something to be true). This leads to an objective or at least intersubjective notion of truth, which most psychologists might accept as an ideal but often question on empirical grounds.

Maybe psychologists and philosophers should try to collaborate more on the topics of truth and objectivity. This is also important for the distinction between practical and epistemic rationality. If beliefs aim at truth, as said above, and do so in some not entirely subjective sense, then epistemic rationality serves a distinguished, not entirely subjective, aim. By contrast, it is very doubtful that there are corresponding aims on the side of practical rationality. We rationally pursue our practical goals, which may be many. And we all should behave morally. However, whether there is an (objective) moral truth is the highly contested issue of moral realism, which is still undecided (see, e.g., Sayre-McCord, 1988; Schroeder, 2018). And even if there is something like moral truth, it is highly contested in turn whether the pursuit of this moral truth is a demand of rationality (see also chapter 12.1 by Fehige & Wessels, this handbook). By contrast, the pursuit of truth might be the key issue of epistemic rationality. This is what sets apart the two domains and makes epistemic rationality special.

## 4. Normative and Descriptive Theories of Rationality

Whenever we talk about theoretical and practical rationality, about beliefs, knowledge, desires, pro-attitudes, etc., this has a *normative* and a *descriptive* dimension. The distinction is at the heart of our philosophico-psychological enterprise and was already implicitly used in the previous sections. The core of the distinction can be made explicit in three questions: We *ought* to be rational, but what exactly does this require from us? We think we *are* rational (at least to some extent), but how do we *actually* reason? And what is the relation between these two questions?

The first question is more at home in philosophy. It is concerned with the *normative* standards of rationality. The second question has been at the center of psychological research since the very beginning of the discipline. It is concerned with the development of *descriptive* theories of human rationality. The third question might be the most controversial. It is motivated by the seemingly huge gap between normative and descriptive approaches to human rationality: people seem to commit many errors and suffer from biases in their thinking, reasoning, and decision making, when measured against standard normative systems such as classical logic, probability theory, or decision theory. Our aim in this section is to illustrate the normative–descriptive distinction by explaining some central issues and results on both sides. We also want to discuss the relation between the two perspectives, where they diverge, where they converge, and how we conceive of this relation, to obtain a comprehensive theoretical and empirical understanding of human rationality.

We should start by clarifying what we mean when we speak of "ought" or "norms." These terms are systematically ambiguous. We may speak about empirical or about genuine normativity. *Empirical* normativity concerns the norms we empirically find in a certain community. For instance, when we come to Great Britain, we learn that one ought to drive on the left; this is an empirical fact about Great Britain. Another example is religion. We can empirically study which religious norms are accepted in certain communities and thus mostly followed. In a *genuinely* normative perspective, by contrast, the empirical norms can never settle whether we should really drive on the left or whether we ought to follow certain religious norms. We cannot find out merely by empirical investigation what ought to be the case and what we genuinely ought to do or believe. It is a matter of normative deliberation and of taking or accepting a normative stance. In this sense, it is basically a fallacy to derive *ought* from *is* (and also *is* from *ought*).

This is an important distinction, which is often blurred. Within legal philosophy, it has been strongly emphasized by Hart (1961, pp. 54ff.), who distinguished normativity viewed from an *external* and from an *internal* perspective. From the external perspective, it is just an empirical question what is normative in a given group ("In the UK, they recognize the law to drive on the left . . ."). In this perspective, there is no opposition between the descriptive and the normative. An outside observer can recognize an empirical norm to hold just by noticing that it is mostly followed and that deviations are sanctioned by law or by the fellow people. In the internal perspective, by contrast, we ourselves accept norms as standards governing how we ought to act. Only in this internal perspective do we have a real distinction between normative and descriptive theories of rationality. It is this real distinction that we want to discuss in this section. Here, normativity is always understood in the genuine or internal sense.

### 4.1 Normative Theories

Genuine normativity is a huge topic for philosophers. But, we argue, it is also highly relevant for psychology: empirical research needs some normative reference points as benchmarks for human thinking and action. This is sometimes denied by cognitive scientists (Elqayam & Evans, 2011; Elqayam & Over, 2016), a position that we will discuss in subsection 4.3 of this introductory chapter. However, psychological research is full of examples where empirical findings are compared to normative standards: memories are compared with actual events, responses to visual stimuli are classified as "hits" or "misses," spoken sentences are evaluated as syntactically correct or incorrect, and even emotional states are judged as either "normal" or pathological, to name just a few examples. In rationality research, normative standards usually come from logic, probability theory, and decision theory. So, let us start by discussing these theories, although we shall see that the normative discussion is much broader.

Let us first look at the norms of *theoretical* rationality. Already classical logic is a complicated case. If we follow Frege, logic is not a normative theory at all. It states the abstract laws of truth, which are what they are independently of any thinker. Logical truth is like mathematical truth. $2+3=5$—this is an atemporal mathematical truth. Logic becomes normative only when we add that the laws of truth are at the same time the laws of correct thinking, that our reasoning ought to follow the deductive rules of logic (see also chapter 3.1 by Steinberger, this handbook).

Indeed, this seems to go without saying. Didn't we say that truth is the central aim of rational belief? Now, the crucial feature of logical deduction is truth preservation: it inevitably carries us from true premises—if they *are* true—to true conclusions. This is the standard argument for the epistemic value of classical logic. But it does not necessarily speak in favor of classical logic. It could also speak for other systems of logic, such as intuitionistic, relevance, or paraconsistent logic.

Moreover, the fact that logic is truth-preserving does not imply that only logical deduction is rationally legitimate. Recall James's maxim quoted above: "Believe truth! Shun error!" Of course, logic is perfect for the second aim, shunning error: if we do not start from an error, logic does not introduce one. However, concerning the first aim, believing truth, logic does very poorly. It does not tell us what to believe, apart from the logical consequences of the given information. In other words, what is not implicit in the premises never becomes believed; logic is not ampliative. Obviously, we need more than logic. We shall return to this point.

Another point is that classical logic cannot handle degrees of belief. Yet we need a normative theory that tells us how we ought to deal with uncertainty. And there, probabilities are the natural choice. Again, though, probability theory is just a mathematical theory to be used for many and variegated purposes and under different interpretations. One of them is to interpret probabilities as subjective degrees of belief, and then it is obviously a normative requirement that these degrees should behave like mathematical probabilities. Indeed, philosophers have devised a plethora of justifications for considering degrees of belief as probabilities, the most important of which are represented in chapter 4.1 by Hájek and Staffel (this handbook). One type of justification, accuracy arguments, even attempts to justify probabilities in terms of truth approximation (= accuracy), that is, with reference to truth as the epistemic goal (Joyce, 2009; Pettigrew, 2016). These justifications extend to learning rules, that is, rules for how to revise one's degrees of belief after receiving new information or evidence. Bayes' theorem, which is tantamount to the rule of conditionalization, is the main rule. There are, however, various other learning rules studied in the literature (see again chapter 4.1 by Hájek & Staffel, this handbook).

Note that this is a purely normative discussion. If there should be other normatively defensible conceptions of degrees of belief, the entire argument cannot be cogent. Indeed, some alternative conceptions of degrees of belief will be mentioned below. It is therefore unfortunate that psychologists widely take the normative requirements of subjective probability theory for granted and thus focus only on empirical evidence for, or counterevidence against, this assumption.

It should also be noted that the requirement to obey the mathematical axioms of probability is still very weak. It is, for instance, compatible with the strange anti-inductive inference to ever *less* expect the next swan to be white, the *more* white swans one has seen. So, epistemic rationality requires much more than these axioms, but it is highly contested what this might be (see, e.g., the attempts of Carnap, 1971, 1980, at an inductive logic, or D. Lewis, 1980; see also chapter 4.1 by Hájek & Staffel, this handbook).

Let us now turn to the norms of *practical* rationality. For a long time, this field was dominated by decision theory. This is, however, as problematic as the use of classical logic and Bayesian probability for epistemic rationality. The standard decision theory has been codified by Savage (1954). Its central principle of maximizing expected utility, often called "Bayes' principle" (not to be confused with Bayes' theorem), is primarily a normative principle. Again, some psychologists seem to see it as the only normative option and then raise their empirical objections. Economists and philosophers, by contrast, have devoted much effort to justifying the standard theory. The main argument is: If your preferences have certain normatively commendable features, then your utilities can be measured on an interval scale such that your most preferred option is one that maximizes expected utility. Savage (1954) was the first to extend this kind of argument to probabilities and utilities simultaneously (see also chapter 8.2 by Peterson, this handbook). Indeed, probability and utility theory support each other, and neither is easily replaced by some alternative in the presence of the other.

There are many empirical criticisms of standard decision theory, as we will see below. However, there is also considerable normative criticism, as illustrated, for example, by Ellsberg's paradox (Ellsberg, 1961). There is an urn before you with 90 balls of various colors. More precisely, it contains 30 red balls and 60 other balls. The 60 other balls are either black or white in an unknown ratio. You have to guess the color of the ball drawn next. If you are right, you get a reward; otherwise, you get nothing. Now, you first have a choice between guessing that the ball is red and guessing that it is white. Most subjects prefer to bet on red—the known risk regarding red seems preferable to the uncertainty regarding white. Second, you are offered a choice between guessing that the next ball is black-or-red and guessing that it is black-or-white. Here, most people prefer to bet on

black-or-white, perhaps because this time, black-or-white has a known risk while black-or-red has not. But this seems irrational—and is so according to standard decision theory. The first preference seems to take red as more likely than white, while the second preference seems to take red as less likely, which is inconsistent.

The point of the example is not to show that people are somehow irrational. Rather, it intends to demonstrate that the allegedly inconsistent preferences are at least plausible and that there is something wrong with standard decision theory. Thus, we also find attempts to *rationalize* (not merely explain) these preferences, for instance, by introducing a nonadditive kind of probability with a corresponding alternative way of calculating expected utilities. Generally, such normative criticisms tend to generate alternative theories that can cope with the objections, and we could enter here a variegated landscape of normative disputes in decision theory (see chapter 8.4 by Hill, this handbook).

So much about the difficulties with the norms most often used in psychological research. We should, however, emphasize that the normative discussion about rationality is much richer and has produced many more proposals than we can report here. In fact, the entire handbook is supposed to give such an overview, but even this is incomplete. We have briefly mentioned an alternative to standard decision theory, and we have already announced that probability theory is by far not the only account of degrees of belief. Alternatives are plausibility theory, the Dempster–Shafer theory of belief functions, possibility theory, ranking theory, imprecise probabilities, and so on (see the overview in Halpern, 2003). Some of these accounts are represented in this handbook (see chapter 4.7 by Dubois & Prade and chapter 5.3 by Kern-Isberner, Skovgaard-Olsen, & Spohn). Moreover, these alternative approaches often come with their own decision theory (see Halpern, 2003, chapter 5; Spohn, 2017, 2020). We mentioned that, with the exception of chapter 8.1 by Grüne-Yanoff, treating preference theory, the comparative level is neglected in this handbook. However, this is an area of rich normative theorizing as well, concerning both practical and theoretical rationality. Certainly, cognitive research would benefit from more seriously considering such approaches, rather than limiting itself to the traditional reference points of logic, probability theory, and decision theory.

The gap between the normative and the descriptive perspective is perhaps most apparent in the realm of logic. Above, we mentioned the strengths and the weaknesses of classical logic. It is obvious that most of our reasoning is ampliative. This is not something to be criticized but to be accounted for. The need to go beyond classical logic was perhaps most strongly felt concerning the conditional "if–then." The long-known paradoxes of material implication[24] clearly showed that material implication as provided by propositional logic does *not* represent the conditional.[25] This point had dramatic effects. One may even say that the philosophy of logical empiricism failed due to the inability of classical logic to cope with the conditional.[26] The situation changed only with the discovery of conditional logic by Stalnaker (1968),[27] which in turn was based on developments in modal logic about 10 years earlier. Since then, the field has exploded.

Another strong influence came from AI in the 1970s, which also saw the need to go beyond classical logic in order to build ampliative inference into the computer. To that purpose, AI researchers developed various kinds of nonmonotonic, default, and circumscription logics. These developments are not represented in this handbook (but see Gabbay, Hogger, & Robinson, 1994). However, the field soon merged with philosophical attempts, and today a quite confusing multitude of nonmonotonic reasoning systems is available. This is partially a play with formal possibilities, but it is also a field of serious normative dispute. There are stronger and weaker systems, and it is a normative issue whether the axioms and rules of these systems are acceptable. This field is incompletely represented in sections 5 to 7 of this handbook (but see also Koons, 2017). Again, we feel that this is a domain where the distance between psychologists and philosophers should be reduced. A recent attempt in this direction can be found in Ragni, Eichhorn, Bock, Kern-Isberner, and Tse (2017). The work of Stenning and van Lambalgen (2008) is another attempt that uses logic programming to model human nonmonotonic reasoning.

Is there a common methodology behind all these ways of normative theorizing? Indeed there is, although as far as we know, it has received little systematic discussion. What one finds are somewhat vague approaches in which the methodology of normative theorizing is assimilated to the methodology of empirical theorizing. The idea of such approaches is not that normative theorizing would be in search of something like normative truth. This would be a problematic idea. Rather, it is in search of normative agreement, of the better and maybe decisive argument. Normative theories of rationality are just *hypotheses* about what is the best way to come to rationally justified beliefs or actions.

This normative theorizing is much the same as cognitive or empirical theorizing in general. Typically, cognitive psychologists use data from controlled experiments and develop theories that agree with these data.

Normative theories also build on a kind of basic data, although not on empirical or experimental data, but rather on basic normative reference points, such as normative intuitions or primitive normative assessments. For instance, imagine you know that, so far, I have seen very many white and very few black swans, and now you see me betting that the next swan will be black. Spontaneously and indignantly, you comment, "That's silly!" This is a primitive assessment of epistemic rationality, not yet backed by a theory, but very likely to be respected by any systematization of such assessments. Or imagine you give me a voucher for a movie ticket and I throw it away. You are upset and shout, "You are crazy!" This expresses an intuition of practical irrationality. Or you might reproach me: "You can't say that we should host the refugees and at the same time vote for the nationalists!" Obviously, some of these primitive normative assessments are very strong, while others are weaker and fleeting: what at first looks unreasonable perhaps turns out after closer inspection not to be so.

These primitive normative intuitions are not data in the sense typically used in science, but they are similar. Experimental data are the reference points for empirical theories: the data should be predicted and explained by the theory. Similarly, normative theories must agree with and justify those normative reference points or primitive assessments. So, from there on, normative theorizing proceeds just like cognitive theorizing. We try to find rules of moderate generality, we think about first principles, and we systematize our normative assertions. Whenever there is an incoherence in our structure—for example, when a rule or principle has counterintuitive consequences—we try to solve the problem, and so forth. Perhaps the situation is more fluid than in empirical theorizing, because our primitive normative assessments are not fixed and may sometimes change under the influence of convincing theories. In the end, however, we are satisfied when we have brought everything into a good and stable *reflective equilibrium*.[28] Yet this equilibrium is not a matter of taste; it is a matter of careful and ever more embracing argument, even though the result cannot be called "the truth." We will return to this issue in subsection 4.3 of this introductory chapter.

### 4.2 Descriptive Theories

Let us now turn to *descriptive* theories of rationality. Developing empirical theories about the cognitive processes underlying human rationality is arguably one of the most challenging endeavors of the cognitive sciences. Research from the past decades shows that people often draw unjustified conclusions from given premises, that is, they do not adhere to the principles of theoretical rationality. Similarly, most people and organizations accept more losses for a potential high gain than the rules of rational choice theory seem to allow; national lotteries and equity trading are good examples. Hence, people deviate from the alleged norms of practical rationality. So, does the design of the human mind really provide the necessary cognitive abilities to solve the complex problems we are facing, as individuals and as society? What are the limitations of our rational capacities? Many chapters of this handbook seek to give answers to such questions; some are empirically robust and well founded, and others are more tentative and still under debate.

Let us first look at the descriptive side of *theoretical rationality*. Almost all research in this area is concerned with reasoning, that is, people's ability to draw rationally justified conclusions from given premises. As we already said, the three main cognitive issues are competence, errors, and the interaction of reasoning with prior knowledge and beliefs.

Different descriptive theories explain human reasoning *competence*. Some accounts assume that people reason epistemically by syntactic, language-based mental derivations. On the output-oriented level, such accounts explore whether or not the output of the reasoning process is logically justified by the input the reasoner received. On the process-oriented level, the key idea consists in a repertoire of inference rules represented in long-term memory. These rules are derived from general knowledge and refer to sentential connectives such as "if–then" and quantifiers like "all" and "some." Reasoning is a process of transferring the inference rules into working memory and applying them to the given premises, which are also represented in a language-like format. The result is a language-based conclusion (e.g., Rips, 1994). Obviously, such accounts, in order to model reasoning processes, use logic as both a normative standard and a descriptive framework. The core idea is that human reasoning proceeds in analogy to the proofs of formal logic, although some logical connectives might be understood in a way diverging from propositional logic. The natural-logic account in chapter 3.2 by O'Brien (this handbook) is an instance of this theoretical framework. Many versions of this account were inspired by the developments in AI since the 1960s, where most AI systems were so-called production systems and inspired by the idea of a general problem solver (Newell, 1973; Newell & Simon, 1972).

Other descriptive theories assume that human reasoning competence consists in processing subjective probabilities (e.g., Oaksford & Chater, 2007, 2020). We already mentioned this account in our list of milestones

in subsection 2.1 of this introductory chapter. Such theories primarily focus on the output-oriented level. They take the premises as input to the cognitive system and then compare the output to the alleged normative standards of Bayesian probability calculus. Although most proponents of this account say that they are not interested in explaining mental operations on the process-oriented level, it is clear that such accounts must rely on some kind of rules for the computation of subjective probabilities. An important rule in this context is Bayes' theorem (Oaksford & Chater, 2001, 2007). However, since proponents of this account just formulate (output-oriented) computational-level theories, they do not explain how the cognitive system actually computes conclusions according to this theorem.

Yet other theories adopt the position that people use mental simulations to draw epistemically rational inferences. On the output-oriented level, people use the input to create models of what would be the case if the input is true and produce the output as a result of a mental simulation performed to find new information not explicitly contained in the input. These mental models capture possibilities of how the world is, or could be, under certain conditions (e.g., Johnson-Laird, 2001, 2010; Johnson-Laird & Byrne, 1991). The key assumption on the process-oriented level is that reasoning does not rely on syntactic operations, as in rule-based approaches, but on the construction and manipulation of mental models. A mental model represents a possible "state of affairs" described in the premises of an inference problem. It only represents what is true according to the premises but not what is false. The common assumption of most mental-model accounts (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) is that reasoning is a cognitive process in which spatially organized or iconic models of the premises are constructed and then alternative models are sequentially generated and inspected. A conclusion is true if it holds in all models that agree with the premises (Johnson-Laird, 1999, 2001; Johnson-Laird & Byrne, 1991; Johnson-Laird & Khemlani, 2013). This theory is supported by many experimental results and is also implemented in several computer programs (Bara, Bucciarelli, & Lombardo, 2001; Khemlani & Johnson-Laird, 2013; Krumnack, Bucher, Nejasmic, Nebel, & Knauff, 2011; Ragni & Knauff, 2013). Chapter 2.3 by Johnson-Laird, chapter 3.3 by Khemlani, chapter 6.3 by Byrne and Espino, chapter 13.2 by Ragni, and chapter 13.3 by Knauff (all in this handbook) are related to this theory.

Rule-based theories and the model theory also explain why human reasoning is sometimes error-prone—the second question that a descriptive theory of reasoning should account for. In one respect, both theories agree: both assume that, in theory, humans have the competence to think and act rationally according to certain normative standards, but that, in practice, this competence is limited by many internal and external conditions. Apart from this commonality, however, there are large differences between the different theories.

According to rule-based theories, errors can creep in due to the variety of processes that are necessary for reasoning. Where exactly they may creep in is determined by the core assumptions of the theory. Before a mental representation of the premises can be stored in working memory, the premises must first be encoded by processes of understanding. Here, interpretation errors may occur because the use of logical expressions in natural language differs from that in logic. Once the premises are available in working memory, abstract reasoning schemata must be applied to the premises to derive valid inferences. Here, the wrong rules might be selected or the coordination of different rules might fail. And invalid conclusions may be produced by still other processing errors (Braine, 1990; Braine & O'Brien, 1998; Rips, 1994).

In the model theory, reasoning errors are mainly caused by the limited capacities of working memory: reasoning becomes more difficult when multiple models need to be generated from the premises, and it takes additional time to discover inconsistencies between tokens in a model. Errors can occur when potential alternative models are not generated or inconsistencies are overlooked. A crucial assumption is that inferences derived from initially formed explicit models are simpler than inferences that can be performed only by elaborating other implicit models. The distinction between implicit and explicit models is described in chapter 2.3 by Johnson-Laird (this handbook). In the past years, the *preferred models theory* has explained why people usually prefer to construct just a single easy, simple model of the given information, why they ignore alternative interpretations that are also logically valid, and why this leads to inferences that deviate from the norms of classical logic (Knauff, 2013; Ragni & Knauff, 2013).

We should also mention that other theories claim that humans are fundamentally irrational. Some of these theories are about theoretical rationality, but they are particularly prominent in research on practical rationality, notably in areas overlapping with behavioral economics. The most prominent accounts are the different versions of dual-process theories, according to which people reason by means of two different systems or processes. Kahneman (2011) argues that one system, System 1, processes the incoming information fast, largely

unconsciously, and relatively effortless. It relies on heuristics that often lead to good solutions but sometimes also to violations of rational norms. The other system, System 2, processes the incoming information slowly, consciously, and relatively effortfully. This more "rational" system might rely on logical rules, mental models, probabilities, and so on. Because people tend to use System 1 more often, they are considered largely irrational. Although the account is intuitively appealing—which might be one reason for its success inside and outside psychology and even in the public—it has many empirical and theoretical problems (Gigerenzer, 2010; Osman, 2004). Today, most reasoning researchers prefer to speak of two "processes" rather than "systems" and also admit that the simplified functional distinction in earlier theories cannot be upheld (J. St. B. T. Evans, 2018). In chapter 2.5 of this handbook, Klauer gives an up-to-date review of dual-process theories of human reasoning.

How do reasoning processes interact with prior beliefs, background knowledge, or attitudes? We have already said a bit on this topic in subsection 2.2 of this introductory chapter, but we want to emphasize again that no descriptive reasoning theory denies that in daily life, all of these strongly interact. Proponents of the Bayesian approach often allege that other theories do not pay attention to knowledge and prior beliefs. Yet, this is not really true. In fact, a more adequate description is that present reasoning theories differ in the role they assign to prior beliefs. Broadly construed, one position is that reasoning should be studied and described in its pure form. In such accounts, prior knowledge is a moderating factor that can influence reasoning by supporting or hindering it. This is where the term "bias" comes from. The other position is that people do not do much reasoning anyway but mostly use their beliefs to deal with inference problems. "Reasoning" is basically the retrieval of beliefs from memory and some less relevant cognitive procedures to process this knowledge. Most serious theories lie in between the two extremes, and there are several approaches that can deal with the effects of prior beliefs and the resulting nonmonotonicity of human reasoning within logical theoretical frameworks (see chapter 5.4 by Gazzo Castañeda & Knauff, this handbook).

Let us also briefly look at the descriptive side of *practical rationality*. Again, we can be shorter here, although the issue is broader, because we now deal not only with epistemic matters but also with how people actually judge and decide based on their desires, attitudes, values, and so forth. That is, we discuss the topics of judgment and decision making under the heading of practical rationality, even though philosophers would carefully distinguish between decisions and judgments. Here, we just stay away from vocabulary and briefly look at the psychology of judgment and decision making. Nor do we look at what some psychologists and neuroscientists call "decisions" too, for example, to lift one's arm or not, or to choose between two equally meaningless stimuli on the left and right side of the visual field (Heekeren, Marrett, & Ungerleider, 2008). Such "perceptual decisions" are largely *arational* and thus not considered here.

In principle, we could ask the same questions for practical rationality as we did for theoretical rationality: how can we explain people's competence, their performance, and the interaction of the underlying cognitive processes? The situation is a bit more complicated, however. One reason is that prior knowledge and beliefs are now complemented, and maybe even overruled, by values, attitudes, and so on. Another reason is that the discussions that we already mentioned on the normative side carry over to descriptive theories of human judgment and decision making as well. In fact, there is hardly any consensus concerning the best norms by which the judgments and decisions of people should be assessed. Accordingly, it is quite hard to say what a "good" judgment or decision is and what we should call an "error." So, many deep questions touching upon moral, ethical, cultural, economic, political, and ideological issues are involved in such appraisals. Hence, we better stay away from such evaluations in the next few paragraphs.

Broadly construed, a judgment occurs when somebody assigns a value to an object on a particular dimension. The objects of judgment can be things, situations, actions, people, abstract stimuli, and so on. In a judgment, the attitudes, values, preferences, etc. of a person become visible. So, philosophers would rather call it a "value judgment."

The most important theoretical framework for human judgment is the *lens model*, which was developed around the 1950s by Egon Brunswik (1956) and promoted by Kenneth R. Hammond (1996) in research on social judgments. The basic idea of this quite metaphorical model is that people cannot perceive the environment directly and objectively but instead use cues to make inferences, judgments, and choices. Some of these cues are more obvious in the environment while others must be inferred from further cues. People must select the cues they consider relevant, assign relative weights to them, and finally use these cues as aids in the judgment process. Since the real properties of the object are not directly accessible to the perception of the judging person, the cues' usefulness depends on how well they represent the real properties of the judged object. This is why Brunswik is considered

one inventor of the concept of *ecological validity*, which indicates how accurately the cues represent the actual properties of the object within the given environment. The higher the correlation is, the more the cue helps to come to an appropriate judgment.

The lens model is also popular in descriptive theories of human decision making. When individuals decide, they must choose between at least two different alternatives, for example, going to the anti-glyphosate demonstration or having a relaxed Sunday on the sofa. Whereas judgments stand for evaluations and preferences, decisions indicate an intention to pursue a particular course of action (Hardman, 2009, p. 3). The common feature is that both processes are based on cues that reflect more or less well the actual characteristics of the judged object or decision alternatives.

Descriptive theories of decision making seek to explain how people identify different choice options, how they gather and weigh information and cues, how they evaluate the feasibility and desirability of the alternatives, etc., in order to make choices that lead to the best outcome, including all costs and benefits. It is obvious that this is the place where all the normative questions reappear, which we ignore in this subsection. What we can say, however, is that an important turnaround happened when Kahneman and Tversky (1979) developed their prospect theory (see chapter 8.3 by Glöckner, this handbook), in which they argued that the formal structure of probability and utility functions is empirically incorrect and hence proposed more adequate alternatives. Since then, several descriptive theories have been developed that also account for decisions that they may or may not want to call "errors" because they deviate from the classical economic norms. Good examples are the approaches of bounded rationality (see chapter 8.5 by Hertwig & Kozyreva, this handbook), which criticize the standard decision theory as computationally overdemanding. And since classical economists are prone to load utility functions with egoistic or self-concerned connotations, it is easy to denounce the thus-laden theory as highly ideological, which in a way questions its normative status (see subsection 5.1 of this introductory chapter).

### 4.3 On the Relation between Descriptive and Normative Theories of Rationality

How are normative and descriptive theories of rationality related to each other? For Piaget, the answer was clear: typically developed adolescents and ultimately adults should attain the ability to reason according to the rules of formal logic. In this view, human reasoning research actually is, and should be, driven by the norms of logic and by a comparison of mental reasoning with these logical norms (Inhelder & Piaget, 1958). Certainly, this is a strong idealization, which not many people would support today. We cannot go into all the details of this complicated methodological issue.[29] What we can do, however, is to briefly present our opinion on the matter, which, we feel, does not accord with any of the opinions we find in the literature.

First, the normative and the descriptive perspectives are both legitimate and important. This goes without saying for the descriptive perspective. And it holds for the normative perspective precisely in the genuine, internal sense explained at the beginning of this section. We are bound to act, and thus we are bound to take a normative stance—even the decision to let things go is a normative stance. So, the normative perspective is absolutely unavoidable. And it extends to rationality, if rationality is normative. This raises the issue of how the normative and the descriptive perspective on rationality are related.

One possibility is that the two perspectives are simply independent. Apparently, they are at least logically independent. Anyone inferring an *ought* from an *is* commits a naturalistic fallacy—this philosophical lesson from Hume and Moore seems to stand.[30] And surely, reversely inferring an *is* from an *ought* is no less of a fallacy. These assertions are cleared up from the logical point of view in Schurz (1997), although there are quite a few logical subtleties. As a consequence, Elqayam and Evans (2011) argue that much of cognitive rationality research commits these fallacies, to its detriment. The authors do not deny the legitimacy of the normative point of view, but they suggest that empirical research better makes itself completely independent from it.

Another possibility is that "rationality" simply means different things in the two perspectives: there may be rationality$_1$ and rationality$_2$ (J. St. B. T. Evans & Over, 1996). Then, however, our joint handbook would merely rest on a big equivocation. Neither possibility gives us the positive relation we are looking for.

A natural way of building a bridge between the normative and the descriptive side, particularly in our scientifically minded times, is to try to *naturalize* rationality. "Naturalized epistemology" is a slogan introduced by Quine (1969), and if the quote from Quine (1986) in subsection 3.3 of this introductory chapter were true, epistemic rationality would indeed reduce to "technology," to a branch of empirical science. In the same vein, Schurz and Hertwig (2019) explain rationality in cognition to be instrumental for the predefined aim of cognitive success, which is measured in terms of ecological validity and applicability.

Cohen (1981), although not avowedly in the naturalistic camp, in effect argues in a similar way. As mentioned in subsection 4.1 of this chapter, he distinguishes a narrow and a wide reflective equilibrium in normative theory construction. The narrow equilibrium takes normative intuitions or primitive normative assessments as fixed reference points and attempts to systematize only them. By contrast, the wide equilibrium accounts for more comprehensive considerations even of an expert nature, e.g., logical theorizing or only theoretically explicable principles of minimal change. Then normative intuitions may be negotiable. At the end of subsection 4.1 of this chapter, we, too, used the metaphor of a reflective equilibrium. There, we definitely intended it in the wide sense. Cohen, by contrast, ties rationality theory firmly to what he calls the narrow equilibrium. Thereby, however, normative theory construction amounts to empirical theory construction over the fixed normative reference points: it determines our cognitive competence. The cognitive performance may deviate—a little. But the competence can never turn out to be irrational. This is Cohen's way of empirically controlling normativity. Again, we do not find genuine normativity in Cohen's picture.

In our view, these strategies of naturalistic reduction do not properly respect the autonomy of the normative discourse, which is so amply displayed in this handbook. Let us explain this with respect to the three naturalistic strategies just mentioned.

First, if rationality judgments were just a matter of a priori normative intuition, as suggested by Schurz and Hertwig (2019), their criticism of the normative perspective would be justified. But they are not; they are a matter of normative argument. They are also right when they complain that the metaphor of reflective equilibrium entails circular justification. But that's perhaps the nature of justification, and we simply need further criteria to distinguish good from bad circles. In their alternative account of rationality as instrumental for cognitive success, they simply take the aim of cognitive success for granted, and thus rationality research reduces to the exploration of this instrumental relation. We, by contrast, want to maintain the autonomy of normative discourse by interpreting their account as a normative one. It may well be normatively convincing to strive for cognitive success in their precise sense, weighing ecological validity and applicability, and to include this success in our normative reflective equilibrium. But this requires further normative argument.

Second, the quote from Quine (1986) in subsection 3.3 of this introductory chapter underrates this autonomy in a similar way. Beliefs aim at truth. Granted, but perhaps this slogan does not assign to epistemology the task of providing a "technology" for reaching a predetermined aim. Perhaps the aim itself is explicable only within the reflective equilibrium for epistemic rationality (see Spohn, 2016).

Third, Cohen, too, seems to underrate the autonomy when he firmly ties rationality to normative intuitions. But these intuitions, even though we have no better starting point, are often dim or confused, and the task is not just to quasi-empirically systematize but to normatively develop and explicate them. Thereby, they become an indispensable ingredient of the wide normative reflective equilibrium.

Our rejection of naturalistic reduction and our emphasis on the autonomy of the normative discourse can be seen to reflect Kant's doctrine of the autonomy of pure practical reason (Kant, 1788/1908): it has the positive freedom of self-legislation, and it has the capacity and the mission to fill that freedom. However, this emphasis at the same time deepens the gulf separating it from the empirical realm. As stated, logic cannot bridge the gap between the normative and the descriptive. What else can?

The crucial point, we think, is that there are plenty of *defeasible* (not deductively valid) inference relations between the normative and descriptive perspectives on rationality. In practice, we find these inferences everywhere, but we do not find this point clearly addressed in the methodological literature. In moral philosophy, another chapter of genuinely normative theorizing, people talk of prima facie reasons. For instance, if something is (descriptively) pleasing, then, prima facie, it is (normatively) good. This assertion is not analytically true—it is an open question whether something pleasing really is good. But the assertion is plausible—and acceptable when there are no opposing reasons.

The general reason for these defeasible inference relations is that humans are receptive to norms. We tend to follow the norms, not only the norms somehow externally given but also the genuine norms internally endorsed. This receptivity is sometimes strong, sometimes weak. There are uncountable confounding factors. Still, the receptivity exists. The point is often expressed[31] by stating that the normative theory simultaneously serves as an empirically idealized theory. This entails, quite generally, the following defeasible connection: whenever something that is under human control *ought* to be the case, there is some plausibility that it actually is the case. And reversely, whenever something under human control is the case, then there is some plausibility that it *ought* to be the case. This may sound far too

strong, and you will be able to immediately cite hundreds of counterexamples. Granted, but the abstract point could only be put so starkly. Of course, it must be applied with all due caution and reservation. But this does not make the point go away.

In rationality research, this point is ubiquitous. Philosophers explicate reasoning in the hope that people then obey their explications. And psychologists refer to normative standards in order to specify precisely this defeasible relation. Now, whenever there seems to be a discrepancy between norm and empirical fact, an exception to our defeasible rules, there are several ways to go. There is, first, the standard way of sticking to the norm and finding additional explanations why observed facts deviate from the norm. The Wason selection task, for example, leaves the norms of deductive logic unshaken and postulates certain biases disturbing the performance. In this case, the criticism concerns the people: they are somehow irrational. There is, second, the possibility of modifying the normative reference point in order to reduce or possibly eliminate the discrepancy. This was the Bayesian strategy: perhaps people's inferences should not be tested against logic but against probability theory. Then the criticism turns against the scientist for choosing an inappropriate normative reference point.

So far, the discrepancies had no consequences for the normative discussion. Neither logic nor probability theory are demoted in their normative status by the Wason task. It is just a matter of finding out empirically which normative ideal we should take people to adhere to and how much deviation from the normative ideal we should admit. The default assumption is perhaps: the less deviation, the better. But this is not so clear—the psychology of reasoning might still have to find its own descriptive reflective equilibrium.

Then, however, there is the third case, where the discrepancy is taken as a normative criticism. Take the Popper–Kuhn controversy (subsection 2.1 of this introductory chapter): Kuhn described the behavior of scientists. Popper took the first way and criticized this behavior as irrational. But this response seemed implausible, since the scientists did not accept the charge of irrationality. So, it rather seems that something is wrong with Popper's normative theory. Or take the Ellsberg paradox (subsection 4.1 of this chapter): people turn out to be stubborn. Even when we explain to them why their preferences are allegedly irrational, they stick to them. So, are they stubbornly irrational? Perhaps something is wrong with standard decision theory if it does not distinguish between known risk and a numerically identical subjective probability. This is an empirical input, which needs to be respected in the normative discussion. Here we have indeed a defeasible inference from *is* to *ought*.

It is unnecessary to give further instances of the various ways of this defeasible inference. If we are aware of its possibility, we easily see it everywhere. Defeasibility is not a one-way bridge. Defeasible inference carries us both ways, from the normative to the descriptive, and inversely. Empirical theorizing has to find its own reflective equilibrium (most descriptions of scientific methodology are much more detailed, though). Normative theorizing has to find its own reflective equilibrium as well, as sketched above and exemplified in the handbook. However, the two equilibria are correlated. One cannot optimize the one in disregard, or at the cost, of the other.[32] This is very vague but already gives an idea of our main point. It is methodologically important to observe the logic of descriptive–normative reasoning and to keep this *double equilibrium* in mind.

Elqayam and Evans (2011) fear to thereby get entangled in the arbitration problem, the problem of deciding between competing normative accounts: "Psychologists can . . . and do get involved in arguments about which norm is right—perhaps an odd activity for empirical scientists" (p. 239). The arbitration problem is indeed severe and will perhaps never be conclusively decided. Still, we think this activity is not odd at all but unavoidable.[33] It is an activity, or discourse, that philosophers and cognitive psychologists should jointly engage in. It is necessary in order to make progress in our understanding of human rationality. In fact, this handbook is one large plea for this joint enterprise.

## 5. Individual and Social Rationality

There is another basic classification of theories of rationality, namely, into those about individual and those about social or collective rationality. This classification cuts across the distinctions we have already introduced. We find a lot of empirical research on both individual and social rationality, as well as on many normative issues. And just like individual rationality, social rationality has an epistemic as well as a practical dimension.

At first glance, the distinction is clear: individual rationality is about the epistemic or practical reasoning processes of a single person and their outcomes, while social rationality is about group processes, their results, and their rational organization, concerning the purely epistemic formation of a joint opinion as well as the agreement on a joint strategy or the settlement of practical conflicts. We have seen many issues and instances of individual

rationality in the previous sections. In fact, those sections were severely biased toward individual rationality, while neglecting the social dimension of rationality.

There are reasons for this bias. The main reason is that philosophy and psychology both have an individualistic bias. Our Humean heritage gives a nice little example of this bias. On the one hand, one should think that communication provides the foundation for all social epistemology. On the other hand, Hume (1748/1975b, section X, "Of Miracles") classified testimony (through listening and reading) as a mode of perception among others, to be compared with perceptions in other modes, and the processing of any kind of perception is a matter of individual rationality. If so, social epistemology seems reduced to individual epistemology. Not that we would maintain this nowadays, but the bias persists. Another reason is that the notion of social rationality is much less clear than that of individual rationality, as we will indicate below. So, the field of social rationality is still more tentative.

For these reasons, this handbook has an individualistic bias as well. However, the social dimension of rationality is so important that we also wanted it to be represented in this handbook, even if only in a more exemplary way (see sections 9, 10, 12, and 14 in this handbook). In the following, we want to discuss at least briefly some central issues of social rationality.

### 5.1 Individual Rationality in a Social Context

Game theory is the classical paradigm of rational choice theory in a social context.[34] It has developed into *the* foundational economic theory and pervades economic theorizing. It is clearly about practical rationality, and it is clearly about how to behave rationally in social contexts, that is, when other agents are involved. However, is it also about collective rationality?

There is no clear answer. In order to understand this, we must look at the basic distinction between cooperative and noncooperative game theory. Cooperative game theory is not necessarily about cooperation. But it doesn't turn cooperation into a problem. It assumes that communication among all participants or players is free and available and that agreements can be reached, which may or may not be enforceable. This may, but need not, result in cooperation. Perhaps you only compromise with your opponents, or you find partners in order to defeat your opponents, and so on. Large parts of cooperative game theory are indeed about suitable coalition-forming. Thus, it certainly addresses important forms of collective rationality.

Similarly, noncooperative game theory is not necessarily about noncooperative agents. Clearly, players are allowed to cooperate also in noncooperative games. However, cooperation requires action, e.g., communicating, bargaining, entering into agreements, etc., and from the point of view of the noncooperative theory, each kind of action must be explicitly represented as a move in the game. This is why game theorists consider the noncooperative part as basic and the cooperative one as derived. And it is for this reason that only noncooperative game theory is represented in this handbook.

Now, it is important to see that noncooperative game theory is not about collective rationality. It is rather about individual rationality in a social context.[35] How can you pursue your interests in an environment of other agents who also pursue their interests? This is not simply a matter of instrumental rationality, of manipulating our fellows like instruments. It means recognizing that our fellows try to reach their aims as well. It is, one might say in Kantian terms, a matter of treating our fellows as ends and not as means.

In this perspective, the notion of a Nash equilibrium takes center stage. The actions or strategies of the players are in equilibrium if, given the choices of the other(s), no one can improve by changing her choice. Only then does everybody individually optimize. Also, only such equilibria can be publicly known and shared—no public advice can deviate from such an equilibrium without giving to at least one player a reason to reject the advice.

It may be asked whether such individual rationality in a social setting is really something special. Classical game theory assumes so up to the present day and considers standard decision theory as the limiting case of "one-person games" or "games against nature." Bayesian game theory, as developed by Harsanyi (1967, 1968a, 1968b), and its successor, epistemic game theory (Bernheim, 1984; Pearce, 1984; Spohn, 1982), try to reverse the direction of derivation and to conceive of equilibrium behavior as ordinary expected utility maximization under certain special conditions (see chapter 9.2 by Perea, this handbook, for how far this strategy carries). Quite a different interpretation of noncooperative game theory is found in *evolutionary game theory* (see chapter 9.3 by Alexander, this handbook), which is particularly germane for evolutionary explanations of various cognitive features (see chapter 1.3 by Schurz and chapter 10.6 by Cosmides & Tooby, both in this handbook).

The fact that individual rationality in a social setting does not per se amount to social or collective rationality becomes obvious in the *prisoners' dilemma*, which is perhaps the most famous paradigm of game theory:[36] two prisoners in two separate cells may or may not confess an alleged common crime. In other words, they can

either defect (from his partner in crime, $D$ = confess) or cooperate (with his partner in crime, $C$ = not confess). If a prisoner is the only one to confess, he is set free as a key witness (this has the highest utility, 3—the exact figures are not relevant), while the other one is imprisoned for a long time (this has the lowest utility, 0). If both confess, they get imprisoned for a slightly shorter time: at least they have admitted their guilt (utility 1). And if both don't confess, they can be convicted only of a lesser crime, entailing a small punishment (utility 2). The story has thousands of variations and occurs to us every day. The only important matter is the structure of the situation, the distribution of utilities, as given, for example, in table 1. There, the utilities of the row chooser are given in the lower left corner of each field and those of the column chooser in the upper right corner. Obviously, the only Nash equilibrium is ($D$, $D$). However, the justification of $D$ is even stronger: each player fares better by defecting, not only given that the other defects, but *whatever* the other does: either he gets 3 instead of 2 or 1 instead of 0. In other words, cooperating is strictly *dominated* by defecting. Hence, there is almost unanimous agreement that defecting is the only rational thing to do, at least if the game is played only once; iterated playing is a more complicated issue. From the point of view of individual rationality, this seems undisputable. From the point of view of collective rationality, however, this is silly. Both would fare better by cooperating: they would get 2 instead of the 1 from the equilibrium.

The prisoners' dilemma is often characterized as a conflict between individual and collective rationality. But what is collectively rational? In the prisoners' dilemma, it seems obvious that joint defection is collectively irrational because it is strictly Pareto-dominated, as economists say: both could fare better by jointly cooperating. However, not being Pareto-dominated (i.e., being "efficient") provides at best a necessary condition for collective rationality, indeed a very weak one. Note that the $C$/$D$ combinations are also efficient. In fact, it is quite unclear what collective rationality might mean beyond the criterion of efficiency, and in some contexts, the latter is even doubtful as a necessary condition (e.g., in the liberal paradox invented by Sen, 1970, chapter 6). However, this only shows that social rationality is a much more difficult and obscure notion than individual rationality. We will return to this later.

**Table 1**
Prisoner's dilemma

| | | Player B | |
|---|---|---|---|
| | | $C$ | $D$ |
| Player A | $C$ | 2, 2 | 0, 3 |
| | $D$ | 3, 0 | 1, 1 |

What is the normative status of game theory? It is quite unclear. The vacillating interpretation of rationality as individual or social adds to the uncertainty. However, the uncertainty persists even if we focus on individual rationality. Noncooperative game theory clearly has strong normative foundations, but there are also many situations, like the iterated prisoners' dilemma and the ultimatum game (Güth, Schmittberger, & Schwarze, 1982), where the recommendations of game theory are ignored by agents and indeed doubtful even from a normative point of view. Therefore, many economists have become quite guarded about game theory as a normative theory and prefer to characterize it as an idealized model that can claim validity only under restricted conditions. How far this validity extends, though, has become quite contentious in turn.

Most of the descriptive work in the field uses experimental paradigms inspired by noncooperative game theory. Dozens of experiments use different versions of the prisoners' dilemma, the ultimatum game, and many other games, to study the roots of rational cooperation and conflicts between individuals and social groups (Axelrod, 1997).

We should emphasize, however, that such experiments were criticized for many reasons (Bowles, 1998; Gomberg, 1989; Ostrom, 1998, 2010). The main criticism is that they derive from the tradition of classical economic thinking and the self-adjusting mechanisms of free markets. Indeed, it may be argued that an indoctrination with rational choice theory biases people toward acting more selfishly than they would otherwise. Therefore, one may suspect that the theory and its related paradigms create the type of selfish people they axiomatically assume (Frey & Meier, 2002). The main cause for this criticism is that economic theory tends to load the notion of utility with egoism and to assume that people decide like selfish maximizers, promoting only their individual benefit. However, many empirical results in the social and behavioral sciences, particularly in behavioral economics, show that this draws a too one-sided picture of human rationality, downplaying the fact that collaboration lies at the heart of the human species and our society (Tomasello, 2009a, 2009b).

### 5.2 Social Rationality

Let us now turn to proper social rationality. This is about the attempts of social groups to form joint beliefs and joint preferences and to perform joint actions in a

rational way. When the group cooperates to arrive at rationally justified beliefs, conclusions, or explanations, we speak of *epistemic* social rationality. When the group cooperates to form joint decisions, we talk about *practical* social rationality. A group cooperates if each group member coordinates its activities with that of other group members, and all group members work toward a goal that benefits the entire group (Witte & Davis, 1996).

There are many situations in our society where groups of people work together to achieve shared epistemic goals. For instance, a group of jurors wants to reach the verdict "guilty" if the defendant is guilty and "not guilty" if he is innocent. The members of an ethics committee, in psychology or medicine, reason together to anticipate possible risks resulting from experimental treatments or medical therapies. Scientists working together on a joint publication draw inferences from empirical data and want to come to justified conclusions and explanations. Politicians and parliaments seek to find the right means to reduce climate change or to fight pandemics. It is obvious that we urgently need a better understanding of this form of social rationality, since it lies at the heart of democratic societies, which assign central legislative, executive, and judicial decisions to *groups* of people, ranging from boards, cabinets, commissions, and parliaments to the entire electorate.

The easiest way of responding to normative questions here might be to use the same standards as for the epistemic rationality of individuals. The problem with this transfer, however, is that individual rationality is usually defined with respect to the formation, combination, and revision of individuals' beliefs, but it is unclear whether it makes sense to attribute beliefs to collectives (e.g., List & Pettit, 2011; Theiner, Allen, & Goldstone, 2010). Some accounts have been suggested, but none of them are as prevalent as the logical and Bayesian frameworks of individual epistemic rationality (see also chapter 10.1 by Dietrich & Spiekermann, this handbook).

So, let us look at the descriptive side of epistemic social reasoning, where interacting groups evaluate the "correctness," "truth," and "believability" of epistemic inferences and beliefs. Only a small number of studies investigate this topic. One stream of research started from work by Patrick R. Laughlin and coworkers, who showed that the mere announcement of a purely "intellectual task" with an "objectively correct solution" triggers other social combination processes than the announcement that different opinions should be merged to come to a joint decision (Laughlin & Adamopoulos, 1980; Laughlin & Ellis, 1986). The authors use the concept of "demonstrability" to explain which beliefs are shared by a cooperating group: a "demonstrably" correct solution is the one for which the group members can demonstrate that it is correct. They also found a correlation between demonstrability and the number of group members required for a collective solution (Laughlin & Ellis, 1986). Similarly, a study by Moshman and Geil (1998) on deductive reasoning in groups indicates that groups are better at avoiding typical reasoning biases. They asked 143 college undergraduates—32 individuals and 20 groups of 5 or 6 interacting peers—to solve the Wason selection task and found a clear group advantage. After intensive deliberation, groups turned out to be less prone to the confirmation bias than individuals. There are a few more studies on the epistemic rationality of groups, but overall, this field has started to flourish only recently (e.g., Claidière, Trouche, & Mercier, 2017; Jern, Chang, & Kemp, 2014).

A second branch of research is in the tradition of Gricean and neo-Gricean theories of communication (Grice, 1975, 1989; Levinson, 2000), the "relevance theory" (Sperber et al., 2010; Sperber & Wilson, 1995), or the "argumentative theory of reasoning" (Mercier & Sperber, 2017). All these approaches to communicative rationality could be said to view human conversation as a form of social epistemic rationality. Some of these approaches are discussed in chapter 5.5 by Hahn and Collins and chapter 5.6 by Woods (both in this handbook).

Research on *practical* social rationality is concerned with many topics that are not important when groups perform epistemic tasks. This research has in fact a much longer tradition; there is a huge amount of literature from philosophy, psychology, economics, political science, etc. Nevertheless, the area is still quite diffuse, and the empirical and the normative side seem less well integrated than in the case of individual rationality. One reason is perhaps that in the field of practical social rationality, normative benchmarks stand out much less clearly than in that of individual rationality. For this reason, we have treated the topic in this handbook rather in an illustrative than in a fully representative way.

On the normative side, about the earliest attempts to get a theoretical grip on the topic come from democratic voting and social choice theory, which are not represented in the handbook (see List, 2013). They stand in the utilitarian tradition and try to determine (procedures for finding) the common good or a social preference on the basis of individual preferences. Group decision making has the similar problem of integrating diverging aims into one common aim and of finding a joint strategy for proceeding. Such group decision processes often presuppose the sharing of information and the integration of

diverging opinions into a common one. This is a central issue in social epistemology; see chapter 10.1 by Dietrich and Spiekermann (this handbook).

Clearly, these fields are full of normative issues. One of the earliest and most baffling ones was Arrow's impossibility theorem, showing that there is no way of aggregating arbitrary individual preferences into a social preference, provided this aggregation is to satisfy certain apparently very plausible conditions, one of which is non-dictatorship (i.e., that the social preference is not identical with the preference of a given individual or "dictator"; List, 2013, section 3). It is much less clear, though, whether such normative issues are issues of rationality. In the case of group opinion and group decision making, the answer seems to be positive: they seem quite analogous to individual decision making. However, already the case of preference aggregation seems different. Is the non-dictatorship condition (if it can be interpreted as the absence of a dictator in the ordinary sense) a demand of rationality? Hardly. It rather appears to be a moral requirement. In other words, in these social matters, normative issues tend to take on a moral character. And the extent to which morality is a question of rationality is very much open. If we could reduce morality to rationality, moral maxims would be just as well justified as postulates of rationality. Presumably, though, morality should keep its autonomy. This unclear situation is another reason why this handbook does not fully engage in normative social issues. At least section 12 of this handbook is devoted to the relation between rationality and morality.

Cognitive psychologists have also investigated practical rationality and decision making in groups. Most of this research is motivated by the hope that social rationality is in principle superior to that of individuals. However, the empirical findings suggest otherwise: while several experimental studies showed that sometimes groups do indeed perform better than individuals, other studies did not find any group advantage, or even reported that individuals outperform groups.

The superiority of groups can have several reasons. One prominent finding is that groups often outperform individuals because groups tend to weight the input from more competent members more strongly. This often leads to a "truth wins" principle in a social combination process in which the existence of a single knowledgeable group member is necessary and sufficient for a correct group response (Laughlin & Ellis, 1986). However, Schultze, Mojzisch, and Schulz-Hardt (2012) have provided an alternative explanation, namely, that group members actually become more accurate individually during group interaction. Many other comparisons between groups and individuals are reported and often suggest different explanations for group advantages (Liang, Moreland, & Argote, 1995). The "wisdom of crowds" literature (Surowiecki, 2004) shows that collectives can decide rationally, or optimally, even when the individuals in the collective are irrational (Lyon & Pacuit, 2013). Yet, this only works when the decisions of the group members are independent from each other. Good overviews about the advantages of groups in judgment and decision making can be found in Esser (1998) and Witte and Davis (1996).

Inferiority of groups can also have several reasons. For instance, classical studies have demonstrated that *conformity* can lead to suboptimal group performance (Asch, 1951, 1956). Janis (1972, 1982) showed that members of groups often avoid raising controversial issues due to the desire to preserve harmony. This "groupthink" can lead to irrational decisions in groups of potentially rational agents—and has been considered the main reason for the fact that in 1986, critical information was neglected when planning the launch of the space shuttle *Challenger*, which ultimately cost seven astronauts their lives (Moorhead et al., 1991). Research in the hidden-profile paradigm—in which some information is shared among group members, whereas other pieces of information are unshared—demonstrates the potential disadvantages of group collaboration (Brodbeck, Kerschreiter, Mojzisch, & Schulz-Hardt, 2007; Jönsson, Hahn, & Olsson, 2015; Kerr, MacCoun, & Kramer, 1996; Kerr & Tindale, 2004; Mojzisch & Schulz-Hardt, 2006; Stasser & Titus, 2003). Research in social-choice and game theory demonstrates that groups of fully rational agents can end up making decisions that are collectively irrational (Bacharach, 2006; McAdams, 2009). Today, we know many more biases that appear in groups and prevent them from making good decisions. For instance, we could hardly call it rational if the order in which group members speak influences the joint decision. In group meetings, however, those who speak first often have a higher impact on the outcome of the group decision. This often leads to suboptimal output of the group (Bazerman, 2012; Hartmann & Rafiee Rad, 2020) and may polarize the group members (Zhu, 2013). Again, this can make it hard for groups to reason and decide optimally or rationally.

In sum, the study of social rationality, both from the normative and from the descriptive perspective, is a highly complex issue that we are just beginning to understand. It is also related to many other research fields in social psychology, economics, political science, and even anthropology and legal theory. This does not

make the topic simpler. It is obvious, however, that such questions are immensely important for our society, and it is good that recently more psychologists and philosophers have become interested in these topics.

## 6. Rationality in a Process-Oriented and an Output-Oriented Perspective

We now want to return to the important distinction between output-oriented and process-oriented approaches to human rationality. At first glance, the two perspectives seem quite closely attached to the two disciplines at the center of this handbook: output-oriented approaches dominate in philosophy, and process-oriented approaches are crucial for the cognitive psychology of human rationality. Although these connections to the two disciplines are not entirely wrong and can largely be seen in the historical developments of both disciplines, we will later show that the distinction also cuts across the disciplines: some psychologists are also satisfied with output-oriented approaches, and process-oriented approaches are, to some extent, also pertinent in philosophical approaches to rationality.

Our distinction between theories that focus either on the output or on the process of cognition is of course not new. Marr's (1982) well-known distinction between the *computational* and the *algorithmic* level of description is certainly the most influential version of this distinction, at least in cognitive science. In subsection 2.2 of this introductory chapter, we already explained why we nevertheless decided to use another terminology here: Marr's terminology often leads to misunderstandings, it is too closely attached to the computer metaphor of human cognition, and our terminology leaves more space for the vital normative–descriptive distinction.

Of course, Marr's approach had precursors. Already in the early days of modern logic, there was, on the one hand, the so-called syntactic approach, according to which logic just consists in the manipulation of symbols in conformity to certain rules, for example, in writing strings of symbols (i.e., formulae) that have the shape of axioms and then transforming them in ways defined by inference rules; no understanding of the formulae is involved. On the other hand, there was the so-called semantic approach, according to which formulae receive truth conditions (i.e., meanings), and inference is defined by necessary truth preservation, which leaves open how the inference is syntactically, that is, algorithmically, realized. The two approaches indeed diverge, as became clear through the epoch-making incompleteness results of Kurt Gödel (1931): logical or mathematical provability does not exhaust logical or mathematical truth! This in turn led to investigation of what is computable, i.e., algorithmically accessible, in the first place. Three very different theories—the theories of Turing machines, recursive functions, and λ-definability—were proved to be equivalent, which corroborated the *Church–Turing thesis* that each of these theories indeed captures our intuitive notion of computability (see Kleene, 1967, chapter V; Kripke, 2013). This was the birth of all our modern conceptions of algorithms.

Another, related terminological distinction is the separation of the *symbol level* and the *knowledge level* in the classical work by Newell and Simon (Newell, 1982; Newell & Simon, 1972). On Marr's computational level, we ask, "What is the goal of the computation?" We ask "why" a cognitive process is performed by the cognitive agent. This is similar to Newell's *knowledge level*, which is an abstract level of description for knowledge, intentions, and reasons. On Marr's algorithmic level, the question is: "How are the goals and strategies at the computational level algorithmically realized?" which is similar to Newell's *symbol level*. On this level, we ask "how" the cognitive system works: which properties the representations have, how the relationship between input and output is mediated, and which transformations are made to get from the former to the latter.

The distinction is also reflected in more recent philosophical writings, where we find the distinction between *reason-based* and *mechanistic* models of a given phenomenon. Bechtel and Abrahamsen (2005) state that process-oriented explanations in the life sciences often consist of models of the mechanisms taken to be responsible for a given phenomenon. This differs from the output-oriented, nomological explanations commonly presented in philosophy. Bechtel and Abrahamsen identify several differences between these approaches. First, whereas in philosophy, the focus on language is quite typical, scientists who develop mechanistic explanation are not limited to linguistic representations. Second, the fact that mechanisms involve assumptions about component parts and operations provides direction to both the discovery and testing of mechanistic explanations (Bechtel & Abrahamsen, 2005, p. 421). Finally, mechanistic approaches are developed for specific cognitive phenomena and are initially not phrased in terms of universally quantified statements. Later generalization then involves the investigation of both the similarities of a new phenomenon with those already studied and the differences between them. In the past decades, Marr's approach and the related suggestions to distinguish between different levels of explanation have received

many variations, extensions, and criticisms (Peebles & Cooper, 2015). A good overview is given in the collection of articles in Colombo and Knauff (2020), which brings together about a dozen authors from philosophy, psychology, biology, AI, robotics, anthropology, and other fields.

We have already mentioned many examples of output- and process-oriented approaches to rationality. In principle, when cognitive psychologists explore rational reasoning, be it of a logical, conditional, causal, probabilistic nature or whatever, they aim at understanding the underlying cognitive representations and processes. The difficulty of this attempt, however, arises from the fact that human reasoning belongs to the area of complex cognition, a particularly challenging subfield of cognitive psychology. One crucial feature of complex cognition is that it is embedded in a multitude of cognitive processes interacting with one another and with other, noncognitive processes. For instance, reasoning involves a large amount of strongly interlinked processes of language processing, perception, working memory, long-term memory, etc. (Knauff & Wolf, 2010). We return to some of these subsystems in the next section on the preconditions of rationality. In the past, a small number of models tried to account for this plurality of processes, but this was often at the expense of the theoretical rigor of these approaches (Dörner, 1999). Most approaches nowadays focus instead on subcomponents of reasoning and decision-making processes, with the goal of complementing them stepwise with additional processes. Most cognitive approaches in this handbook belong to this class of theories. Another process-oriented approach is the theory of *computational rationality*, which combines models from AI, cognitive science, and neuroscience to reconceive processes of reasoning and action under uncertainty through the lens of computation (Gershman, Horvitz, & Tenenbaum, 2015). We had a chapter on this approach planned, but, unfortunately, it did not work out.

We have also mentioned many examples of philosophers' attempts to state and justify normative standards for these reasoning processes. Typically, those accounts took an output-oriented perspective: They were about establishing a criterion for logically valid deductive inference, about probabilistic norms for degrees of belief, about establishing maximizing expected utility as the rational way to act, and so on. And they were about all the possible alternatives and amendments to these standards. Such approaches also make clear that the two perspectives apply to practical rationality just as well as to theoretical rationality. Of course, there are reasoning processes on both sides. However, the process-oriented perspective may be more difficult to implement on the practical side.

We mentioned already in subsection 3.1 of this introductory chapter how the distinction between the process- and the output-oriented perspective relates to the distinction between internally and externally driven dynamics of epistemic states: the internal dynamics is about how our thinking does or should proceed, and it is thus bound to take a process-oriented perspective. By contrast, the external dynamics is about how we do or should respond to external input; in determining this response, we primarily take the output-oriented perspective, although we may and should investigate by which internal dynamics the response is mediated.

We would like to point out, though, that there is a deeper question underlying our considerations, namely, the question of how to conceive of an epistemic state (or a mental state in general) at all. This question is not pursued in the handbook and will only be touched on here.

Philosophers are used to distinguishing two conceptions of epistemic (or mental) states: as something occurrent or as a disposition. For them, a state is *occurrent* if it presently "goes through our mind" or "is before our eyes," that is, if we are presently aware of it. Thus, philosophers tend to conceive occurrent epistemic states as conscious. For instance, when asked for your opinion about glyphosate, or any other question, you form an occurrent thought or belief, which did perhaps not occur to you a moment before, and you are aware of this belief. We treat an epistemic state as *dispositional*, by contrast, if we take it to consist in a behavioral disposition that may manifest in our verbal and other behavior, to be elicited by questions or tests, but without further input that could change the disposition. The disposition lies, so to speak, in the background of the occurrent beliefs, and we are typically not aware of it.

To illustrate this distinction: in subsection 3.1 of this introductory chapter, we distinguished between belief bases and belief sets. A belief *base* can be understood in the *occurrent* sense, as just a few beliefs of which we are currently thinking and which we may then algorithmically evolve. By contrast, a belief *set* is deductively closed, that is, it contains all logical consequences of the belief base. Thus, we can never have it wholly present in our conscious mind. Belief sets can be understood only in the *dispositional* sense. Similarly, if the Bayesian conceives an epistemic state as a probability measure, he can do so only in the dispositional sense. At any given moment, at most tiny parts of such a measure can be grasped.

The distinction between occurrent and dispositional belief has been developed independently of cognitive

psychology.[37] It is tempting, however, to identify occurrent and dispositional beliefs with representations in working memory and in long-term memory, respectively. Occurrent beliefs might be those that are momentarily active in working memory, often conceptualized as a "global workspace" (Baars, 2002; Dehaene, Changeux, & Naccache, 2011). Only the representations of beliefs in working memory are consciously accessible and available for cognitive processing. Process-oriented approaches from cognitive psychology are concerned with the format, representation, update, and revision of such temporarily active, occurrent beliefs in working memory. Dispositional beliefs, in contrast, may be seen as contents of long-term memory to which we do not have immediate access. Here, our epistemic states are represented as a kind of accumulation of myriads of occasions in which the relevant topics have been presented to our mind. So, consistency checks, for instance, can be performed only for the occurrent beliefs in working memory but not for all beliefs represented in long-term memory, even though this might be normatively desirable.

Some of these ideas about the relationship between the philosophical concepts of occurrent and dispositional beliefs and the psychological conceptions of memory can also be found in Goldman (1978). Here we may leave open whether this cognitive account perfectly matches the philosophical intentions. In any case, the lesson for us is: When our study of rationality pursues an output-oriented or a process-oriented perspective and explores an internally or an externally driven dynamics, we implicitly presuppose this or that conception of epistemic states, and we must be clear beforehand which conception we apply.

Many of our previous thoughts may suggest that philosophers are only interested in the output-oriented perspective and psychologists only in the process-oriented one. Yet, this is not entirely true. Many *micro-theories* from the early stages of psychological reasoning research took an output-oriented perspective (for an overview, see J. St. B. T. Evans et al., 1993). Even today, most psychological studies about choice are still about the resulting choices and not about the underlying thought processes. Similarly, we have mentioned that the "new paradigm," which is closely associated with Bayesianism, is also just output oriented—and our criticism was precisely that it does not attempt to move onward to the process-oriented level. So, clearly, we also find output-oriented approaches in psychology.

Conversely, philosophers are also interested in the process-oriented perspective, and this is where approaches from philosophy overlap with research in AI, for which the process-oriented perspective on algorithms is even the defining characteristic. This overlap is due to the common ancestry in logic. Indeed, in logic, a precise definition of logical validity (output oriented) was preceded by an account of valid proof procedures (process oriented). Many surprisingly different proof procedures have been invented for logics with all kinds of purposes (see chapter 3.1 by Steinberger, this handbook). For instance, axiomatic calculi are very convenient for metamathematical purposes, while they are hardly manageable for actual proofs. No empirical claim is associated with such algorithms. Their point is, rather, to establish the feasibility of the standards assumed, for example, those of logical validity in the case of logic. This feasibility is not guaranteed at all, as is displayed by the incompleteness results in logic (see Shoenfield, 1967, chapter 6), and recursion theory tells us that the computable functions form only a small subclass of all possible functions (Shoenfield, 1967, chapter 7).

The interest in algorithms goes far beyond deductive logic. AI researchers try to find efficient algorithms, and philosophers belabor the wide field of philosophical logics, where we find epistemic, conditional, counterfactual, and nonmonotonic logics, among others (see chapter 5.1 by van Ditmarsch, chapter 5.2 by Rott, chapter 6.1 by Starr, and chapter 7.1 by Pearl, all in this handbook). There is also probability logic (see chapter 4.4 by Pfeifer, this handbook). And the so-called roll-back analysis of decision trees (Raiffa, 1968) is an algorithm for finding a strategy with maximal expected utility. It works perfectly, but precisely its unrealistic nature has motivated the search for more realistic accounts in the field of bounded rationality (Rubinstein, 1998).

This interest in algorithms is not an empirical one, and it is not backed up by any empirical research on thought processes. Indeed, today few AI researchers share the concerns of cognitive science. They are rather interested in the space of possibilities for specifying processes and finding algorithms that arrive at the intended results and besides have other kinds of desirable properties. Similarly, philosophers often discuss normatively plausible axioms before having a semantic benchmark, and if they possess it, they are interested in its computational feasibility—algorithmic inaccessibility might, but need not, be regarded as a blemish on the benchmark. In any case, we may conclude that the process-oriented and the output-oriented perspective are not exclusively assignable to psychology or, respectively, philosophy.

This brings us to the main issue of this section, the relation of the output- and process-oriented perspectives to the normative–descriptive distinction, which

is so central to our enterprise. Certainly, one may suspect strong relationships between the two dichotomies. One obvious connection is that the conclusions of our reasoning processes should satisfy certain normative benchmarks. Such normative evaluations always take place within the output-oriented perspective. Since the normative evaluations lie mainly in the scope of philosophy, this would explain why philosophy is mostly engaged with this perspective. Conversely, the study of our actual reasoning processes can, it seems, only be an empirical, that is, descriptive, matter.

By and large, this is correct. However, we should note that the normative perspective also extends to the process-oriented level. Of course, we normatively evaluate also algorithms in themselves and not only the results they deliver. Computer scientists do this all the time. However, the relevant normative considerations in AI differ from those in philosophy. Typically, the former are about computational complexity classes, that is, about the time and storage space needed by an algorithm to solve a problem. For computer scientists, complexity considerations are based on the behavior of an algorithm in worst-case scenarios (Papadimitriou, 1994). The nightmare for computer scientists are NP-hard problems, which cannot in general be solved efficiently. Although such results from complexity theory take a process-oriented perspective and are obviously relevant for software development, it is questionable whether computational complexity is indeed relevant from a cognitive point of view. One criticism is that it might be better to look at the average performance of algorithms (Sedgewick & Flajolet, 2013). Another criticism might be that people usually do not have to deal with an infinite number of choice alternatives, which typically is the basis of complexity theory. An example is the traveling salesman problem (finding the shortest possible route that visits a given set of cities), which is an NP-hard problem but is manageable when just a small number of cities must be visited. Moreover, people often do not aim for optimality but rather are satisfied with a good solution. Altogether, complexity theory provides potentially interesting process-oriented normative insights for rationality research, but it is questionable how relevant such findings are for descriptive theories of rationality. Some psychologists emphatically deny their relevance (Gigerenzer et al., 2011), while others consider complexity theory highly relevant for the cognitive sciences (Otworowska, Blokpoel, Sweers, Wareham, & van Rooij, 2018).

A different example of a normative process-oriented perspective is given by the attempts in decision theory to consider decision costs. The various ways of reaching a decision require efforts of varying costs, which should enter into the expected utilities of the available options. Yet this idea is apparently threatened by circularity. The costs of a decision procedure depend on the decision problem at hand, and to determine them, one would need to already have completed the decision procedure for the given problem. For this reason, the problem seems quite intractable (Bossaerts & Murawski, 2017; Gottinger, 1982). A radical response to it consists in satisficing: by fixing in advance an appropriate satisfaction level, we avoid the costs of further optimization.

The literature on bounded rationality is quite ambiguous with regard to the normative–descriptive distinction. Treatments from the economic side can also be read as normative recommendations for how to get along with all kinds of restricted resources (Rubinstein, 1998), while psychologists usually explore how people proceed under actual restrictions concerning time, working memory, etc., without evaluating the thinking processes themselves (see also chapter 8.5 by Hertwig & Kozyreva, this handbook). Such evaluations come up only when we ask how to improve our thinking or our rational capacities in general (see chapter 15.2 by Stanovich, Toplak, & West, this handbook).

So, what is the connection between normative and descriptive theories, on the one hand, and output- and process-oriented accounts of rationality, on the other? The lesson from this section is that that the two distinctions are independent and that all four areas generated by them are important and indeed pursued in both psychology and philosophy. The two disciplines have different preferences; this is nothing to complain about. However, it would be particularly helpful if the field were to make progress also on the normative side of process-oriented theories and explain what good thinking ought to be like, not only regarding the output but also regarding the thought processes themselves. Such a development would be helpful not only for education and the development of curricula in schools and universities: it would indeed help all of us if we better understood *how* we should think in order to develop and update rational beliefs and come to good decisions and actions.

## 7. Preconditions of Rationality

This section is concerned with a topic that also could have stood at the very beginning of this introductory chapter: What are the preconditions of human rationality that distinguish us from other animals? What

enables people to think and act (more or less) rationally? Of course, the answer to these questions is: our larger brains! Certainly, the increase in brain size is the most striking feature of human evolution. Probably no other organ in mammalian evolution has evolved as quickly as the human brain: the brain volume of the prehominid *Australopithecus africanus* was about 400 to 500 $cm^3$, which is about the brain size of today's chimpanzees; the brain size of the first tool-making *Homo habilis* was between 500 and 750 $cm^3$; and that of *Homo erectus*, who lived about 1.5 to 0.3 million years ago, reached a volume between 880 and 1,100 $cm^3$. The now living humans have a very wide range of brain volumes. According to Beals et al. (1984), the average worldwide brain volume is 1,350 $cm^3$, although the existing differences are substantial—the brain volume of a healthy adult can lie between approximately 900 and 2,000 $cm^3$.

Obviously, brain volume alone is not decisive for cognitive ability. In absolute terms, the sperm whale has the largest brain, and even the brains of horses and elephants are larger than those of humans. In relative terms—when we relate brain size to body mass—the shrew mouse has the largest brain, although it does not have the cognitive abilities of humans. Among humans, males have about 10% larger brains than females—but on average, there are no differences in cognitive abilities between men and women as assessed by traditional intelligence tests. Nowadays, most researchers say that about 10% of cognitive ability can be explained by brain size (Pietschnig, Penke, Wicherts, Zeiler, & Voracek, 2015).

In fact, the connection between brain size and cognitive abilities is much more complex. Some studies show that the correlation is higher when specific areas of the brain are considered. For instance, cognitive abilities are more strongly correlated with the size and thickness of the prefrontal cortical areas (Menary et al., 2013), they are correlated with the connectivity of neurons in the brain (Cole, Yarkoni, Repovš, Anticevic, & Braver, 2012; Genç et al., 2018), and there is also evidence that the brains of people with lower cognitive abilities consume more energy in particular brain areas than the brains of people with higher cognitive abilities (Neubauer & Fink, 2009). However, even hard-core neuroscientists acknowledge that we need much more research to better understand the connection between cognitive abilities and the anatomy, structure, and functioning of the brain (Luders, Narr, Thompson, & Toga, 2009). This may be one reason why, in trying to understand rationality and other cognitive capacities of humans, many cognitive psychologists for a long time refrained from studying the brain.

Another reason for the long-lasting skepticism of cognitive psychologists toward brain research is even more fundamental. It lies in the assumption of *multiple realizability*, which was the core argument for the rise of *functionalism* in cognitive psychology and philosophy of mind. As a rule of thumb, we can say that the more complex the cognitive ability was that the researchers were interested in, the less they believed that these abilities could be understood by studying how the brain works. Since rationality ranges very high on this scale of cognitive complexity, this position was particularly prominent in rationality research. Right around the millennium, this changed, and today many researchers use functional brain imaging and other neuroscientific methods to study the cognitive and cortical foundations of human rationality. In the following sections, we discuss which of these systems appear to be so fundamental that they can actually be regarded as essential preconditions of rationality. We also argue that the development of *Homo sapiens* into a highly social being was an essential precondition for the evolution of rationality.

### 7.1 Cortical Preconditions of Rationality

The first neuroscientific studies on rationality focused primarily on epistemic rationality. Meanwhile, however, there are also many findings on the neural preconditions of practical rationality. The early experiments with functional brain imaging, which we can assign to the field of epistemic rationality, have primarily investigated logical thinking. The pioneering work was carried out by Vinod Goel (see chapter 1.4 by Goel, this handbook) and our own group (e.g., Knauff, Mulack, Kassubek, Salih, & Greenlee, 2002). It would not be helpful to present all the complex neural activation patterns that have since been reported in many neuroscientific articles. These results can be found in excellent meta-analyses of the several dozen published studies on logical reasoning with conditionals, relations, and quantifiers, which used quite different methods for presenting premises and asking for conclusions (Goel, 2007; Monti, Osherson, Martinez, & Parsons, 2007). Instead, we believe, it is much more useful to bring some structure to the field by describing the *core networks* associated with logical thinking.

We think there are three: the first core network covers large areas of the brain that have often been associated with the processing and production of language. These include areas in the left frontal and prefrontal cortex (PFC), also comprising Broca's area, a region involved in semantic tasks and language production (Goel, Gold, Kapur, & Houle, 1998; Prado et al., 2015). Although this system is highly sensitive to the format of

the presentation and the like (hence the activations vary enormously in the studies), the fact that many imaging studies found activations in these brain regions is usually interpreted as indicating that logical thought processes are based on linguistic processes and representations. It is obvious that this is consistent with one of the central theoretical assumptions of rationality research, to which we return later.

The second core network involved in logical reasoning covers areas in the parietal cortex, the inferior temporal cortex, and the occipital cortex, which are involved not only in visual perception but also in visual mental imagery (Fangmeier, Knauff, Ruff, & Sloutsky, 2006; Knauff, 2009; Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003; Knauff et al., 2002; Ruff, Knauff, Fangmeier, & Spreer, 2003). These findings are often interpreted as indicating that in logical thinking, not only linguistic processes are involved but also processes in which people vividly imagine what would be the case if the premises of a given logical problem were true. These mental simulations—or models, or images—can be of very different precision. Some may be almost like representations caused by visual perception, while others are less vivid but still much more concrete than purely linguistic representations. While linguistic representations agree with the intuitions of many rationality researchers, visual representations fit well with the intuitions of laypeople, who often report experiencing their own thinking as "seeing before the inner eye" (Knauff, 2009, 2013). We will also return to this topic in a moment.

The third core network includes areas of the brain that are associated with executive control processes and conflict resolution. These cover areas in the right lateral/dorsolateral PFC, detecting conflicts between the logical structure of a task and its content. These areas are activated, for example, when a conclusion is logically valid but implausible, because it does not agree with our beliefs. Typically, this is associated with content effects at the cognitive level and the so-called *belief bias*. Since these experiments were carried out mainly by Goel and his collaborators, they take up a lot of space in his chapter (chapter 1.4 in this handbook). In addition, these findings are related to other results showing that in conflicts between form and content, the normatively correct answer is available in the brain—even if it cannot subsequently prevail against the intuitively more plausible response. De Neys interprets this as indicating that people have logical intuitions that are deeply rooted in our brains (De Neys, 2012).

A methodological problem with imaging techniques is that they only establish correlations (but no causal relations) between cortical and cognitive processes. There is a certain irony in the fact that, despite these relatively weak statistical accounts, the results of brain-scanning experiments are so overrated by many people, scientists and laypeople alike (McCabe & Castel, 2008; Munro & Munro, 2014). One way to more directly study the causal connections between cortical and cognitive functions is to examine patients suffering from damages of certain brain regions due to strokes, tumors, accidents, etc. In fact, a variety of studies have yielded interesting results that, fortunately, match well with the results from imaging studies. For instance, a strong argument for the necessity of certain brain regions for a specific cognitive capacity is given when a specific kind of brain damage leads to impairment in tasks of type *A* but does not in tasks of type *B*, while another kind of brain damage leads to impairments in *B* but not in *A*. Such *double dissociations* have also shown that damage to certain brain regions leads to impairment in linguistic or imagery-based reasoning, while other regions are necessary for abstract or concrete reasoning processes that may or may not agree with our beliefs (Goel, Shuren, Sheesley, & Grafman, 2004). Brief overviews of patient studies on rational reasoning and its impairments can be found in Knauff (2009) and Reverberi, Shallice, D'Agostini, Skrap, and Bonatti (2009).

Patient studies, however, also have problems because they are often based on a small number of participants, sometimes just on a single case. In addition, the brain damage is often not clearly delimited or associated with other damages. Recently, new methods have become available in which specific brain regions can be selectively blocked for a short period of time (by applying a strong magnetic field). This corresponds to a temporary lesion—even if this description is greatly simplified and not really correct. In any case, with this method, it is possible to investigate which cognitive tasks can and cannot be performed by an individual whose information processing in the particular brain region is temporarily disrupted. Such studies have also shown that the application of the disrupting magnetic field to areas in the parietal and visual cortex can affect logical thinking, which indicates that mental imagery indeed plays a causal role in logical thinking (Hamburger et al., 2018; Ragni, Franzmeier, Maier, & Knauff, 2016).

Recently, some Bayesians have criticized the fact that the background knowledge and prior beliefs of participants have not been sufficiently taken into account in previous experiments. Moreover, they argued that the instructions were always biased toward the standards of classical logic and did not allow the participants to sufficiently consider their various beliefs when solving a

problem (Oaksford, 2015). However, this camp of rationality research has not yet carried out its own brain-imaging experiments to support these arguments. To take up their criticism, Gazzo Castañeda, Sklarek, Dal Mas, and Knauff (2021) conducted a combined cognitive and brain-scanning experiment in which conditionals with high and low conditional probabilities as well as abstract conditionals were embedded in valid and invalid inferences. Some participants had to evaluate the probability of the conclusion, while others received deductive instructions. Both groups of participants could freely choose between a dichotomous and a graded response. During brain scanning, the participants having received deductive instructions showed elevated cortical activity in regions typically associated with conflict detection and inhibition of prior knowledge. In contrast, when given probabilistic instructions, additional activity was found in areas correlated with the influence of prior knowledge. Moreover, the results revealed that many participants, even those having received probabilistic instructions, preferred dichotomous responses. Content only affected inferences under probabilistic, but not deductive, instructions. These findings suggest that people's consideration of prior knowledge and their preference for graded responses are not universal. In fact, people can flexibly activate or suppress background knowledge when they reason probabilistically or deductively. This calls into question how meaningful and empirically justified the disputes between the different camps in the psychology of epistemic rationality are.

There are also many neuroscientific studies on practical rationality. It has been known for decades that damage to the brain's frontal lobe impairs people's ability to think and make choices and decisions (Stuss & Levine, 2002), and these findings were supported by functional brain-imaging studies of the intact brain (Koechlin, Ody, & Kouneiher, 2003). Many of these studies were concerned with the neural correlates of dealing with moral dilemmas, the prisoners' dilemma, and several other paradigms and tasks from social psychology and behavioral economics (Sanfey, Loewenstein, McClure, & Cohen, 2006; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). It is not surprising that the pattern of neural activation depends strongly on the given paradigms and tasks. Here, again, we just want to give a summary of the cortical networks whose functioning seems to be a precondition of decision making and practical rationality (although neuroscientists probably wouldn't use the latter term).

Neuroscientific studies have usually distinguished between two components of decision-making processes: the first core network is related to cognitive control functions, including task switching, response inhibition, error detection, response conflict, and working memory (Miyake et al., 2000). Several studies have found correlated activity in the dorsolateral PFC and the anterior cingulate cortex, as well as other parts of the PFC (Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; MacDonald, Cohen, Stenger, & Carter, 2000; Stuss, 2011). The second core network is connected to assigning values to choices. These functions have been mainly associated with ventral and medial parts of the PFC (Gläscher et al., 2012; Grabenhorst & Rolls, 2011).

What is interesting about these studies are not the exact cortical localizations but rather the finding that the cortical networks for "cognitive control" and "value-based processes" seem to be quite distinct, even though they interact in many ways. This is interesting, because it has implications for our understanding of practical rationality: the involved processes are not a muddle; rather, they can be clearly separated and have differential effects on the outcome of the decision process. This finding has also been supported by brain-imaging studies with patients who had lesions to one or the other network (Gläscher et al., 2012). That these damages in fact resulted in an impairment of either the cognitive or the value-based processes supports the assumption that these networks are not only correlated but both necessary for rational decision making and practical rationality.

### 7.2 Cognitive Preconditions of Rationality

Having a brain is obviously a necessary precondition for rationality. But which cognitive preconditions must a "system" meet to enable rationality? Certainly, the system must be *intelligent*, at least to a certain degree. However, recent research has demonstrated that the concept of intelligence should not be identified with the concept of rationality, for many reasons. In particular, "intelligence" is an individual-difference concept for distinguishing between cognitive abilities of individuals, whereas "rationality" is typically not used in this way. There is the essential distinction between epistemic and practical rationality but no corresponding distinction in most intelligence tests. And the concept of intelligence frequently served as a door-opener for racism, eugenics, and the justification of social injustice, whereas rationality is not associated with such ideological matters.

In the present context, however, the most striking problem with equating intelligence and rationality arises from the many experimental findings signifying that even very intelligent people are prone to irrationality. The present handbook is full of such findings. There

are even studies in which classical reasoning problems were combined with IQ tests and other measures of cognitive ability, and they often found that performance on such tasks correlates relatively weakly with cognitive abilities and often only under very specific task instructions. For example, people with higher cognitive abilities performed better only if there was an advanced warning that biased processing must be avoided (Stanovich & West, 2008). Based on this evidence, Stanovich and coworkers argued that rationality is a very different concept than intelligence. A certain level of intelligence is a necessary but not sufficient precondition of rationality, which also relies on the ability for reflective thought (Stanovich, 2009) to control and monitor one's own thinking and to correct it from faulty beliefs and reasoning biases. Chapter 15.2 by Stanovich, Toplak, and West (this handbook) further explains why rationality is substantially different from intelligence.

Psychologists have extensively studied control and monitoring processes under the label *metacognition* (Koriat, 1993, 2018). This research largely relies on the investigation of metamemory, which has long recognized the difference between the processes responsible for retrieving information from memory and the processes responsible for monitoring that information (Dunlosky & Bjork, 2008; Koriat, 1993). According to Koriat, monitoring refers to the "subjective assessment of one's own cognitive processes and knowledge" (Koriat, Ma'ayan, & Nussinson, 2006, p. 38). This assessment can rely on implicit cues, for example, the ease with which a memory comes to mind (Koriat & Ma'ayan, 2005), or it can be based on explicit cues, for example, on the person's assessment of her own abilities (Prowse Turner & Thompson, 2009). For example, a basic metacognitive control function is confidence judgment (Koriat, 2008): following an initial response, a second response about the correctness of the previous response is generated (Fleming, Dolan, & Frith, 2012). This is often accompanied by a metacognitive experience, called the *feeling of rightness* (FOR), which can signal when additional cognitive work is needed (Thompson & Johnson, 2014; Thompson, Prowse Turner, & Pennycook, 2011; Thompson, Therriault, & Newman, 2016). Another metacognitive experience is *fluency*, the ease with which information is processed (Unkelbach & Greifeneder, 2013).

While most of the past metacognitive research was concerned with the regulation and monitoring of retrieval processes from memory, there has recently been a growing interest in the metacognitive processes that accompany more complex cognitive tasks, such as problem solving and rational reasoning (Ackerman & Thompson, 2017). For instance, Thompson and coworkers have shown that metacognitive control processes support individuals in solving conditional reasoning problems (Thompson et al., 2011). The difference between classical reasoning experiments and these studies was that participants were asked to generate an initial response, then a FOR rating, and then a final conclusion. Thompson et al. (2011) found that individuals use FOR as a metacognitive control function, which can signal when additional analysis and cognitive effort is needed, which finally leads to better reasoning performance in individuals. More on the essential connection between rationality and metacognition can be found in chapter 8.6 by Thompson, Elqayam, and Ackerman (this handbook).

Psychological research on metacognition mainly proceeds from what we have called the process-oriented perspective. However, the issues may also be tackled from an output-oriented perspective. This has been done by philosophers, although they do not use the label "metacognition" but rather speak about "higher-order attitudes," which are concerned with beliefs about one's own beliefs or also about one's own desires, or with desires concerning one's own desires, etc. Higher-order attitudes are hence concerned with all kinds of self-referential propositional attitudes. The origin of this topic lay in the fact that constructions like "It is necessary that $p$," "It is permitted that $p$," "$a$ wants that $p$," "$a$ thinks that $p$," etc. can be iterated, just as we can form arbitrarily long conjunctions. For instance, I can say, "I think that I think" or "I believe that I want," etc. So, the question arose what the logic of these constructions is, a question that is extensively treated in modal and intensional logic (Montague, 1974; Zalta, 1988). It became clear, however, that this question is not just a matter of logic, of the linguistic rules governing these constructions, but also a matter of rationality requirements. For instance, it is disputed whether we are always aware of our beliefs, i.e., whether when I believe that $p$, I also *believe* that I believe that $p$. And conversely, the question is whether a second-order belief is rationally infallible, i.e., when I *believe* that I believe that $p$, then I do in fact believe that $p$ (Kemmerling, 2017, chapters 13–20). The topic is not restricted to epistemic matters. For instance, Frankfurt (1971) has famously proposed that free will is essentially connected to our ability to form effective second-order desires and volitions.

Higher-order attitudes have also a dynamic dimension. So-called auto-epistemology deals with the connections between one's present first-order beliefs and one's (present) second-order beliefs about one's future (or past) beliefs. For instance, if you believe that tomorrow

you will believe that it will rain the day after tomorrow, shouldn't you already *now* believe that it will rain the day after tomorrow? One is inclined to say, "Yes, you should" (Binkley, 1968). The issue has interesting consequences, particularly in probabilistic terms (van Fraassen, 1984). Finally, interpersonal versions of the topic loom large in everyday life: we permanently think about the attitudes and beliefs of our fellows, about what they think about us, and how common knowledge emerges in social or public processes. This is particularly relevant in social discourse and communication (see chapter 10.3 by Meggle, this handbook). It is also related to the *false-belief task*, to which we return in the next section. Although the task is mainly used to study the development of "theory of mind" in children, it has also been used to investigate the development of rationality in children, for instance, when they learn to reason counterfactually (Rafetseder, Schwitalla, & Perner, 2013).

Although the false-belief task is a good example of the overlap between philosophy and psychology in the study of epistemic rationality, there are still many ideas in philosophy that have not yet been taken up by psychologists. For instance, the philosophical work on one's present beliefs and one's beliefs about one's future beliefs would also be very interesting from a cognitive point of view. Overall, research on metacognition and second-order attitudes and beliefs would have many implications, for example, concerning the rationality of nonhuman animals: other animals seem to have little or no reflective metacognitive capacities. So, does rationality require such capacities? This raises not only empirical but also conceptual questions.

Let us turn to the relation between rationality and *memory*, one of the grand concepts and research topics of cognitive psychology. We can hardly doubt that rationality would be impossible without an ability to learn, store, represent, and remember knowledge, beliefs, attitudes, values, and all kinds of mental states. An interesting observation, though, is that theories of rationality are often not connected with accounts of the structure, functions, and constraints of human memory. This applies particularly to normative and output-oriented theories of rationality, notably in philosophy, but not only there. These theories presuppose a highly idealized model of the agents' memory. A good example for such an idealization is the rational requirement of the consistency of one's belief system. Thus, the evaluation of one belief requires the evaluation of others to maintain consistency. While this is a normative demand, the belief systems of actual human beings are often not adjusted in that way. This becomes apparent when we consider that knowledge and beliefs must be maintained over certain periods of time, must be recalled when required, and must be consciously manipulated to draw inferences and to make decisions. Beliefs can also change over time, and, importantly, they can be forgotten or may sometimes not be retrievable from memory.

All these processes take place in two subsystems of human memory: working memory and long-term memory. *Working memory* is conceived as a memory system that allows us to maintain information for certain relatively short periods. Even though older theories of short-term memory specified particular time frames in seconds, the really critical issue is how much information needs to be maintained and whether new information is coming in while you are still occupied with maintaining the previous information. In theory, information in working memory can be rehearsed for a very long period if no new information is coming in. The limiting factor is thus not time but capacity. An early quantification of the capacity limit was suggested by Miller (1956), who claimed that the information-processing capacity of working memory is around seven elements. Later research, however, showed that this number depends on many other factors; for instance, working memory span is lower for long than for short words (Baddeley, 1986). The "elements" can be words or visuospatial information or any other kind of meaningful verbal or perceptual units. They can come bottom-up, from the perception of external stimuli, or top-down, from thought processes. Forgetting in working memory can have different causes. One cause is that elements in working memory decay over time, unless decay is prevented by rehearsal. Another cause can be that new elements replace older ones, and different elements interfere with each other or might compete during the retrieval process. Since all these can lead to forgetting, elements in working memory can have two fates. If they are not rehearsed, they are irretrievably lost. Or they can, via consolidation, become elements of long-term memory.

*Long-term memory*, in contrast, is the storage system for all we have learned through our lifetime. The elements in long-term memory might not currently be used but are needed to enable rational thinking and actions. Elements in working memory can become elements of long-term memory when we make meaningful associations to what is already stored in long-term memory. One subsystem of long-term memory—called *semantic memory* by psychologists since Tulving (1972)—is a repository for the long-lasting storage of world knowledge, ideas, concepts, beliefs, attitudes, and everything that we have accumulated throughout

our lives that helps us to make sense of the world. For example, the knowledge that glyphosate is a pesticide is stored in semantic memory and also the belief that it may kill bees. All such memory traces are strongly interconnected in a network-like structure, which follows several complex organizational principles (for an overview, see J. R. Anderson, 2000). Another subsystem of long-term memory—which psychologists call *episodic memory*—is related to biographical information and events, which are represented including the temporal and spatial context. Surely, the memory traces of such personal experiences are of great importance for the formation and revision of beliefs, but we cannot discuss this topic here.

The structure and functioning of working and long-term memory are often the limiting factors of human rationality. They can explain the large gap between how people actually reason and how they *should* reason according to certain normative theories. For example, when a person is confronted with the set of premises of a reasoning problem, the amount of information can simply exceed the capacity of working memory. Hence, the individual might forget some of the relevant propositions and thus come to incorrect conclusions. Or the information from an earlier premise might interfere with new information, and so forth. There are many other reasons for human irrationality that are related to the structure and limitations of human memory. People might also use strategies to deal with these limitations, for instance, by using heuristics, which we have already mentioned several times. Mental model theory, too, is inspired by the limitations of working memory, since deviations from the norms of rational reasoning may occur because the number of mental models to be considered exceeds the capacity of working memory (Klauer, 1997; Knauff, Strube, Jola, Rauh, & Schlieder, 2004; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002; Vandierendonck & De Vooght, 1997). Other rationality theories are more strongly related to long-term memory, for instance, by assuming that we solve problems by retrieving domain-specific knowledge that we have acquired in the past (Cheng & Holyoak, 1985; Oaksford & Chater, 2012, 2020).

It is essential to understand that working memory is the window to our beliefs, knowledge, attitudes, opinions, and so on. It brings together the outer and inner world. Only what is momentarily active in working memory can be consciously experienced, reflected upon, and checked for consistency or correspondence to the actual world. It is the "global workspace," corresponding to all the momentarily active, subjectively experienced mental states of a person (Baars, 2002; Dehaene et al., 2011). In section 6, we already argued that occurrent beliefs may be seen as elements in working memory, whereas dispositional beliefs are (sets of) elements stored in long-term memory. This connection between occurrent beliefs and working memory, on the one hand, and dispositional beliefs and long-term memory, on the other, fits well with the fact that we are aware of what is currently in working memory but do not have direct conscious access to long-term memory, just as we are not permanently aware of the countless beliefs that we have accumulated throughout life. We can, of course, become aware of a small subset of beliefs when they are transferred into working memory in the service of ongoing tasks. In principle, such a task can also consist in evaluating and readjusting beliefs in the way normative accounts propose. In working memory, though, only a small number of beliefs can be considered at the same time, which obviously provides strong constraints on consistency checks as suggested in many theories of belief systems and belief revision in philosophy and AI research (Alchourrón et al., 1985; Gärdenfors, 1992). This is an idea that deserves more empirical investigation. In any case, there can be no doubt that having a functioning memory is a central, if not *the* central, precondition that a system must meet to enable rationality.

You may object that *language* is at least equally important for rationality. This is certainly true. Indeed, many philosophers and psychologists share the view that language and thought, and thus rationality, are closely related. Already Plato lets Theaetetus consent to the claim, "Well, then, thought and speech are the same; only the former, which is a silent inner conversation of the soul with itself, has been given the special name of thought" (*Sophist* 263e, quoted from Plato, 1921). Similarly, Aristotle (1984, book III) saw thinking as an ability that only those living beings can have who can "consult with themselves". This is a widespread position in contemporary philosophy and psychology that we cannot spell out in detail here. However, the opposite position is that language, thinking, and rationality might not be as closely linked as it seems. Pinker (1994), for instance, says that the idea that thought and language are the same is a "conventional absurdity" (p. 49). With this position, of course, we are drifting into a huge field of research, where philosophers and psychologists have suggested many different theories of human thought that are less closely linked to language and more to mental models, spatial representations and processes, mental simulations, bodily experiences, and so forth (de Vega, Glenberg, & Graesser, 2008).

Why is the role of language for human rational thinking so controversial? One reason is that many people, unnecessarily, associate the topic with their opinions about the influence of culture, cultural relativity, and the universality of cognitive principles. This leads directly to opinions about the relation between nature and nurture that have always been more ideological than scientific. The other reason, although perhaps not independent of the first point, is that empirical evidence is equivocal. Indeed, the famous Sapir–Whorf hypothesis, that the structure of a language determines its speakers' way of thinking and conceptualizing the world, has led to an enormous amount of empirical research. It will not come as a surprise that many linguists interpret the results as support for linguistic relativism, whereas cognitive psychologists tend to think more in terms of universal principles that all people have in common, independent from the language they may use. The universalistic position can be found in Li and Gleitman (2002), and a relativistic reply can be found in Levinson, Kita, Haun, and Rasch (2002).

So, is language a precondition of rationality? Currently, the best answer is: yes and no. In the previous section, we reported findings from cognitive neuroscience indicating that one cortical system involved in rational thinking comprises several language-related brain areas. And there are tons of findings from cognitive psychology that point in the same direction. On the other hand, this does not mean that language is necessary for all kinds of rational thinking. Can't children think rationally before they learn to speak? Don't nonhuman animals also exhibit forms of rationality? Maybe the equation "thinking = speaking" is too strong, but it also would be mistaken to deny that language plays an important role in human rationality.

*Visual imagination* is another capacity of the human mind that many cognitive scientists consider an important precondition of human rationality. And if we accept this, it would again challenge the position that thinking is determined by language. Sometimes, as we just said, people indeed experience their thinking as inner speech, but they also very often report seeing "mental pictures" during thinking, "seeing before their inner eye," for instance, what would be the case if the premises of a reasoning problem were true. They often say that they can scan such visual mental images to find new information that is not explicitly given. This idea is as old as the idea of thinking in language. Aristotle, for instance, used the term *phantasma* for ideas closely akin to perceptual experiences. Today, most psychologists accept that visual mental images are a kind of mental representation that closely resembles perceptual experience. Empirical evidence shows that people can scan visual images to obtain information that has never been intentionally stored, and it shows that the experience of visual mental imagery is accompanied by activity in the primary and secondary visual cortex, which supports the idea of a strong overlap between visual imagination and visual perception (Kosslyn, 1994, 2005).

Today, most cognitive scientists accept that visual mental images are a special kind of mental representation in human memory. However, as with language, there is still disagreement about the actual importance of visual imagery for human rationality. On the normative side, skepticism prevails. Among philosophers and logicians, the idea is prominent that rational thoughts must be based on some kind of formal language. Although it is generally accepted that pictures and diagrams are suitable for explaining logical relationships, they are rarely accepted as an independent method of proof. This way of thinking has a very long tradition, and there are relatively few exceptions to this view (see chapter 13.1 by Jamnik, this handbook).

On the descriptive side, the evidence is inconclusive. Most cognitive experiments were based on the following hypothesis: if visual images are important for rational thinking, then tasks with an easily imaginable content should be easier to solve than tasks with a content that is difficult to imagine. One can already see that this question is primarily concerned with content effects, which, by definition, should not really play a role in logic. However, the empirical findings suggest otherwise. Some experiments have shown that logical thinking is indeed easier with easily imaginable materials. Other experiments did not find any connection at all, or the results even pointed to the opposite direction. These contradictory findings have only been resolved in recent years, when it was shown that the different results were caused by the fact that very different types of imaginal thinking were investigated (Knauff, 2013; Knauff et al., 2003; Knauff & Johnson-Laird, 2002). Summarizing the more recent results, it appears that the assumptions on the normative side are not entirely wrong. Indeed, some studies could show that visual mental images can even be a nuisance to logical thinking. This may, however, not be true for all kinds of rational thinking. But in general, the results allow two conclusions. On the one hand, mental representations, in which people mentally simulate what is or could be the case, may be an important precondition for rational thinking. This is, for instance, the core assumption of mental model theory. Mental simulations help us to think about alternatives to reality,

to think counterfactually, to think about the possible consequences of our own actions, and also to understand the mental states of others. On the other hand, this does not say anything specific about the representational format of such mental simulations or models. It is indeed premature to assume that these kinds of representations are equivalent to pictorial mental images (see chapter 13.3 by Knauff, this handbook). Rather, several experimental findings and computational models show that they can be more abstract and only account for information that is relevant for the given task (Knauff, 2013). These findings underscore what cognitive psychologists repeat—with good reasons—like a mantra: we may sometime have conscious access to the outcome of our thoughts, but it is almost impossible to experience the cognitive processes that have led to this outcome. This is why we need both output-oriented and process-oriented research to understand the true nature of human rationality. Introspection cannot be relied upon; it can and does mislead us and makes us believe that our thinking works in the way we experience it.

### 7.3 Social and Evolutionary Preconditions of Rationality

Our previous thoughts on the preconditions of rationality have been limited to the individual. For instance, visual mental images may be some of the most private things we all have, which may be why they are so interesting and prominent in the psychological literature. In subsection 5.2 of this introductory chapter, however, we have already stressed the social aspects of rationality. We would now like to follow up on these thoughts and argue that not only the biological and the related cognitive development of our brain was important for the evolution of rationality. Equally important was the development of *Homo sapiens* into a social being that cooperates, thinks about the beliefs and actions of other people, and tries to convince them of its own beliefs and plans.

One of the earliest theories in this direction was developed by Leda Cosmides and John Tooby (Barkow, Cosmides, & Tooby, 1992; Cosmides, 1989). In their evolutionary theory of rationality, they argue that human reasoning abilities have developed because in bartering and trading, compliance with contracts must be checked and cheaters must be sanctioned. For instance, to hunt a mammoth, many men must work together. To accomplish this, other men must guard the village. That is why a contract is made: if the hunters share their prey, the others take care of their families in their absence. This is a rule that must be followed by both sides to ensure the common good of the entire village. Violations of the rule must be punished. According to Cosmides and Tooby, that is why, in the course of evolution, a system for detecting cheaters has developed in our brain that helps us to draw logically valid conclusions. The idea of a "cheater-detection module" goes back to the evolutionary biologist Robert Trivers and his theory of *reciprocal altruism* (Trivers, 1971). In the past decades, Cosmides and Tooby have developed their account further into an adaptationist program, according to which human cognitive architecture was shaped by natural selection to solve exact and nonintuitive information-processing problems. The researchers describe this approach in chapter 10.6 (in this handbook).

Other social theories are based on the close connection between thinking, argumentation, and communication. For instance, logic helps us in developing a sound argumentation; it brings structure to our arguments and makes them comprehensible for the interlocutor. We generally find illogical and incoherent arguments unconvincing. Already the father of logic, Aristotle (1985), in his *Rhetoric*, dealt with how to persuade other people of one's opinion with good arguments and correct conclusions in meetings and discussions (see also chapter 5.6 by Woods, this handbook). Recently, Hugo Mercier and Dan Sperber (2017) have developed an evolutionary theory of human reasoning according to which the human species developed its tremendous reasoning competence because logically sound arguments are the best means to persuade others.

Mercier and Sperber (2017) also provide a socio-evolutionary account of why reasoning is so often unreliable. Reason, they argue, is not directed toward solitary use, toward achieving rational beliefs and decisions by our own efforts. Rather, reason helps us to justify our beliefs and actions to others, to persuade them through argumentation, and to evaluate the justifications and arguments that others direct at us. Although we agree that social cooperation may be an important evolutionary foundation of rationality, there are also many objections. The most general reservation applies to all evolutionary theories of cognition: they are often "just-so stories," that is, unverifiable narrative explanations for human behavior (Gould, 1978; R. C. Richardson, 2007).

Mercier and Sperber also try to explain why reason did not evolve in other animals. Their answer is: because no other animal has language. This position is related to theories that place the close interaction between social cognition and language at the center of the evolution of cognitive abilities (Heyes, 2012). In essence, these theories state that tasks such as hunting a mammoth require

a high degree of cooperation, which is only possible in combination with language. Of course, other animals, such as groups of lions or societies of ants and bees, also collaborate, in a more instinctive way, but they do not communicate their goals and plans, which is only possible by means of language.

Although we do not have anything like "cognitive fossils," whereas other disciplines can draw on petrified remains of plants or animals from earlier geological ages, many empirical findings emphasize the importance of sociality and cooperation for human theoretical and practical rationality. In a nutshell, these studies show that many mistakes people make, for example, in conditional reasoning, no longer occur if the task is placed in a social context. A famous study was conducted by Cheng and Holyoak (1985), who presented students with a social version of the Wason task. Imagine you are a customs officer at the airport. You should check the following rule: "If there is 'entry' on one side of the form, then 'cholera' is in the list of diseases on the other side. This is to make sure that entering passengers are protected against the disease." When the participants were presented with the corresponding cards for the cases $P$ (entry), $\neg P$ (no entry), $Q$ (cholera on the list), and $\neg Q$ (no cholera on the list), they performed significantly better than in the original Wason task and also better than in tasks embedded in a nonsocial context. Other studies showed that improved performance in the Wason task involving social agreements only occurs when the subjects put themselves in the role of a "rule supervisor" but not when they are only "observers" (Gigerenzer & Hug, 1992). In subsection 5.2 of this introductory chapter, we already reported several studies showing that groups of people are better at solving epistemic reasoning tasks than the individuals that constitute the group.

In the same section, we also reported some results on the social aspects of practical rationality. Game theory, as we already described, is by far the most far-reaching theoretical and empirical framework in this context. A special version of it is *evolutionary game theory* (Weibull, 1997; see also chapter 9.3 by Alexander, this handbook), where successful cooperative action has been explored with different kinds of games, e.g., signaling games (D. Lewis, 1969), the so-called stag hunt (Skyrms, 2004), and the already mentioned prisoners' dilemma (Axelrod, 1997). These studies showed that, depending on the game, successful cooperative action can, but need not, emerge in groups and need not always be evolutionarily stable. Tit for Tat, for instance, can lead to the escalation of conflict. An alternative is the so-called GRIT ("graduated and reciprocated initiatives in tension reduction") strategy, which can lead to stepwise de-escalation. According to this method, it is rational for the party willing to mediate to approach the opponent by first publicly declaring his willingness to reconcile and then making as many concessions as possible, as long as these do not cause any major damage. If the opposing party reciprocates these steps toward reconciliation, chances are good that the conflict can be resolved in a rational way (Osgood, 1962).

Even Charles Darwin was aware of the importance of cooperation in human evolution. In his groundbreaking book *On the Origin of Species*, he wrote that man's low physical strength and speed and his lack of natural weapons, etc., are more than compensated by man's social cooperation, which has enabled him to help and receive help from others (Darwin, 1859). Apparently, this is miles away from the idea of Thomas Hobbes and his famous dictum "Homo homini lupus"—"Man is a wolf to man." It also far away from many, sometimes intentional, misinterpretations of Darwin's theory as a scientific justification for egoism and selfishness. Today, for many researchers the development of social cooperation has been the main cause for the development of the cognitive abilities of humans, which go so far beyond those of all other species. The cognitive anthropologist Michael Tomasello found that other animals and even great apes do not, or at least to a much lesser extent, have the capacity to understand their conspecifics as intentional beings (Call & Tomasello, 2008). This research is closely related to studies on the *false-belief task* that explores the development of "theory of mind" in human infants. For example, in the unexpected-transfer task (Dennett, 1978; Perner & Wimmer, 1985), a child sees that Mary puts a candy into a red box. Then she goes out of the room and the child sees that another child moves the candy from the red into a blue box. Now the observing child is asked, where will Mary look for the Candy? The correct answer, of course, is that Mary will look in the red box. Such theory-of-mind tasks have received great attention in psychology and philosophy, as they require an awareness in the child that another individual does not necessarily possess the same beliefs or knowledge that they themselves possess. They are also important when we want to understand the evolutionary roots of human rationality. From our view, it is hard to imagine that rationality could have evolved without the ability to infer the mental states of others and oneself. Both go hand in hand, and presumably reflections on the mental states of others even came first, before humans ever started to think about their own mental states. In any case, both seem to be essential preconditions for human rational thinking. More details on the

evolutionary foundations of human rationality can be found in chapter 1.3 by Schurz (this handbook).

## 8. Frontiers of Rationality

In the previous sections, we have provided a systematic overview of the psychological and philosophical frameworks for rationality. Most of these frameworks attempt to characterize rationality as a general concept that can be applied to many domains and fields of human thinking, reasoning, judgment, and decision making. This generality is important, because we do not want to have too many small, highly specific empirical results about, and conceptions of, rationality for every single domain or area of discourse. But, one might object, does this generality sufficiently recognize that there are many fields and applications of accounts of rationality—in everyday life, in public discourse, and in the sciences and humanities—that might require more specific concepts of rationality? In the following, we want to mention at least the more important of these domains and discuss whether they indeed require special concepts of rationality or can be subsumed under the more general theories presented in this handbook. That is the first aim of this final section of this introductory chapter. And then there are so many other topics in philosophy and psychology that overlap with rationality. We could not cover all these neighboring fields, but we should at least briefly explain how these areas are related to the empirical and theoretical conceptions of rationality presented in this handbook. That is the second aim of this section.

### 8.1 Are There Special Forms of Rationality?

When we ask this question, we first think of counterexamples, for example, logic, which is a general discipline par excellence—the canons of logic apply everywhere. Similarly, game theory is intended to apply to all agents in social situations. It has thus invaded large areas of the social sciences (see chapter 10.4 by Raub, this handbook) and is not restricted to particular domains, for example, to games in the ordinary sense or financial bargaining, even though it has to say special things about particular situations. And so on. While these theories are very general and largely domain independent, it has been argued that there are also domain-specific forms of rationality not to be subsumed under the existing general accounts. We want to briefly address three of these areas: Do we need a particular concept of scientific rationality? Is there a special communicative rationality designed for social interaction? Do we need a particular account of rationality for artificial intelligence?

Let us begin with the sciences (both natural and otherwise), which consider themselves to be the heartland of rationality. Does this mean that there is a special *scientific rationality*? It is clear where the self-confidence of scientists comes from: from the rise of the scientific method through figures like Francis Bacon, Galileo Galilei, and René Descartes and the subsequent Age of Enlightenment and the accompanying successes in the explanation, prediction, and manipulation of natural phenomena. This was the beginning of the evolution of our modern rational practices and our conceptions of epistemic rationality. The main goal of science is to acquire new knowledge and justified beliefs by means of empirical methods and systematic theorizing. This epistemic work lies at the heart of science, and scientists seem to pursue their epistemic work in a particularly rational way. At least, this is the common ideology.

But scientists also have problems of a practical nature. For instance, the choice of research issues is a practical decision. Should I try to develop vaccinations against a pandemic virus or rather implants for cosmetic surgery? Who should benefit from my research, the whole society or a privileged group? These are clearly value-laden decisions, for the budgeting institution as well as for the individual researcher. The legal claim of the freedom of science is a practical decision as well, based on general freedom rights and assumptions about how to best motivate and organize research. Many scientists think that such practical problems are not scientific problems, which could be solved by scientific methods, and hence do not lie in their specific competence. Scientific rationality does not refer to such problems.

But then, the so-called value-freedom of science is often cited as a hallmark of scientific rationality. This attitude was famously promoted by Weber (1917/1973) but has also been criticized as being itself a normative and value-laden position. Habermas, as one of the main critical voices, has emphasized the normative and philosophical presuppositions of empirical research (Habermas, 1968). The *positivism controversy* centered on such issues (Adorno et al., 1969). There is also a lot of contemporary discussion about the postulate of the value-freedom of science, which is represented in chapter 14.2 by Bueter (this handbook).

Thus, it seems that it cannot be denied that scientific practice is suffused with norms and values. Can we still defend the idea of the value-freedom of science as an ingredient of scientific rationality? The only way to do so is to limit the idea to the area of epistemic rationality. Only here can value-freedom be a norm for science, and then it means that no value judgment or normative

conclusion can be inferred from scientific methods alone. Doing so would be to commit the naturalistic fallacy, that is, to violate Hume's principle that no *ought* can be inferred from an *is* (see subsection 3.3 of this introductory chapter and chapter 1.1 by Sturm, this handbook). This also implies that scientific knowledge can support normative conclusions only in the presence of normative premises. These premises, however, must be made explicit and transparent to the public. From what we have written in subsection 4.3 of this chapter, it is clear that we want to modify this requirement. There, we concluded that the empirical scientist must also engage in normative theorizing, because she must deal with all the defeasible bridges between descriptive and normative considerations.

So, if science has a special claim on rationality, this can only refer to the epistemic side. Scientists are supposed to be specialists in epistemic rationality. Cognitive psychologists, for instance, have highly sophisticated methodologies of hypothesis testing, mostly of a statistical and an experimental nature. This is what cognitive psychology has in common with the other sciences, although each field has its own challenges and its own responses. Scientific methodologies are sophisticated normative canons of epistemic rationality, and they exceed everyday epistemic rationality by far. This is a competence science has built up over centuries. Of course, this is not to say that scientists always follow these canons. Phenomena like "p-hacking" (performing statistical tests until they yield a significant result) or publication bias (that mostly statistically significant results are published in journals) indicate that science also suffers from certain structural deficits in epistemic rationality.[38]

However, such deficiencies should not distract from the special normative competence of the sciences with regard to epistemic rationality. Still, this is not yet to say that there are special principles of scientific rationality (chapter 14.1 by Andersen & Andersen in this handbook offers some relevant considerations). Ockham's razor (Sober, 2015), classical statistics, and inference to the best explanation (Lipton, 1991) may be claimed to constitute specifically scientific methods. The general trend in philosophy of science, though, is to subsume such scientific methods under general principles of epistemic rationality. For instance, the Popper–Kuhn controversy mentioned above seemed to suggest otherwise. According to Kuhn, scientists follow a special practice, which Lakatos (1978) and others have attempted to rationalize. This rationalization seemed to be tailored for the sciences. However, it later led to investigations into the logic of theory change in general, and this in turn resulted in belief revision theory as displayed in chapter 5.2 by Rott (this handbook) and related accounts. Thus, what seemed special to science has been subsumed under general epistemic rationality. Another example is provided by the dispute between classical and Bayesian statistics, where classical statistics defends special principles of statistical inference, while Bayesian statistics tries to get along with general principles of Bayesian learning. Thus, it is at least open whether the sciences can or must appeal to special principles of epistemic rationality. This question is also important when we consider the connection between scientific rationality and the public understanding of science, which is discussed by Bromme and Gierth in chapter 14.3 of this handbook.

A different topic is *communicative rationality*. In subsection 7.2 of this introductory chapter, we briefly discussed the extent to which rationality presupposes linguistic faculties. A complementary idea is that social linguistic behavior, that is, communication, embodies a particular form of rationality. A prominent example is the approach developed in the *opus magnum* of Habermas (1981), who has a democratic notion of practical rationality that transcends that of instrumental rationality. For him, rationality is rather a form of public justification. Thus, communicative rationality is manifested in a discourse free of domination, in which all arguments are considered without prejudice. Rationality and validity claims in general have to prove themselves in such a discourse. Ultimately, communicative rationality is guided by a very Kantian idea: namely, that in rational discourse, we should treat our fellow humans not as means to our ends but as ends in themselves. If so, communicative rationality would be entailed by one version of Kant's categorical imperative.

Not surprisingly, Habermas enthusiastically received the program of *inferentialist semantics* consistently expounded by Robert Brandom (1994, 2000). There, a linguistic community is represented as a community of reason-givers and reason-takers, and it is explained how this exchange of reasons and inferential relations is constitutive of linguistic meanings. Yet, Brandom's notion of a reason is based on the ordinary notion of an epistemic or practical reason, as it is widely discussed in this handbook (e.g., in chapter 2.1 by Broome and chapter 5.3 by Kern-Isberner, Skovgaard-Olsen, & Spohn, both in this handbook). So Brandom does not support the assumption of a particular communicative rationality. In fact, he only insists that linguistic practice is an essentially rational practice.

A different account of communicative rationality is provided by H. P. Grice's (1975, 1989) theory of *conversational implicatures*. This theory conceives of communication as a rational activity governed by the so-called cooperative principle. Based on this principle, Grice shows why communication works the way it does. However, nothing

more than the rationality of cooperation is assumed in his approach, and again no deeper theory of rationality is involved.

Indeed, in analytic philosophy, there is a general tendency to deny a special form of communicative rationality. This tendency is clearly expressed in chapter 10.3 by Meggle (this handbook). It mainly attempts to develop the *intentionalistic account* of meaning of Grice (1957), which, in contrast to Habermas, does not assume richer forms of rationality transcending instrumental rationality. David Lewis (1969) then builds on Grice's approach and amends it by his account of conventions in general and meaning-constitutive linguistic conventions in particular. This account appeals to game-theoretic rationality, which Meggle conceives as a form of instrumental rationality and not as a more embracive kind of practical rationality, as Habermas sees communicative rationality. However, we can interpret the game-theoretic approach also in a different way and argue that according to game theory, individuals treat each other as persons with their own aims and interests and thus respect the Kantian formula. If so, the analysis of conventions and linguistic communication provided by this approach conforms to Habermas's ideal as well.

These are just three examples that illustrate that the forms of rationality that play a role in communication are already well covered in this handbook. They also explain why we did not further pursue the prospect of a special theory of communicative rationality. This also conforms to our general attitude of not dealing with philosophical and psychological issues of concept formation and linguistic meaning in this handbook.

A very different issue is the relation between rationality and *artificial intelligence* (AI). In section 7 of this introductory chapter, we already learned that rationality should not be identified with intelligence. Hence, we will not discuss here whether AI systems are intelligent. But we should discuss whether such systems are rational in the general sense or whether we need a special account of rationality to evaluate AI systems. To do this, it is important to distinguish two different approaches to artificial intelligence: *symbolic* AI, in which knowledge and reasoning procedures are explicitly represented by symbols, constraints, and predicates, and AI based on *machine learning*, which uses statistical methods to identify patterns and regularities in massive data sets to make predictions and decisions. Interestingly, the two approaches have opposite strengths when it comes to epistemic and practical rationality. In symbolic AI systems, all information is explicitly represented in the "knowledge base" and is more or less readable for humans. Importantly, it is required to keep the knowledge base consistent, and where inconsistencies appear, rational updating and revision strategies are used to reestablish consistency. Consistency is one of the general requirements of epistemic rationality and a research area where AI, philosophy, and cognitive psychology are intimately linked. In fact, many theories of rational belief revision were developed at the intersection of these disciplines (Gärdenfors, 1992) and are already represented in the general concept of epistemic rationality.

The epistemic rationality of symbolic AI systems is not limited to the way knowledge is represented but concerns also the way inferences are performed. Typically, symbolic AI systems use the general principles of classical logic to draw inferences and evaluate conclusions. However, this approach has many difficulties and thus has been complemented by systems of nonmonotonic logic, which mirror different, sometimes weaker, rationality requirements. Still, the goal of symbolic AI is to build systems that resemble or even exceed human epistemic rationality.

The problem, though, is that even if these systems reach a relatively high level of epistemic rationality, they often fail to show satisfactory performance when it comes to decision making. For instance, so-called expert systems, which boomed in the 1980s, were not very good at making decisions in medicine, law, or other domains. To understand the limits of these systems, we did not need a special conception of AI rationality.

The limited practical rationality of symbolic AI systems is probably one reason why, in areas where decisions are concerned, they have been largely ousted by AI systems based on machine learning. These methods are not new; they were already used in the 1980s, but with limited success (Rumelhart, McClelland, & the PDP Research Group, 1986). Today, AI systems based on machine learning perform much better, mainly due to the enormously increased computing power of computers and the huge amounts of data from the Internet and social media. The AI systems are trained with these giant data sets and learn to recognize statistical patterns, which are then applied to new data. The training makes it possible to make decisions, for example, which costumer receives a loan, which stocks are bought or sold, for which patient a certain medical therapy is still worthwhile or not, how new employees are hired, and so on. Many of these systems do not perform as well as Silicon Valley wants us to believe. Still, we can evaluate their outputs with our general understanding of what is a good decision.

The problem, though, is that nobody can understand how such systems reach their decisions. We can only

evaluate the output of the system but not the processes that led to its decision. Moreover, such systems have very limited epistemic rationality. They rely on massive data sets but not on knowledge in the sense of true and justified beliefs. They cannot explain how they acquire a belief, which we usually see as a requirement for rationality in humans. An alternative might be *explainable artificial intelligence*, which combines machine learning and symbolic AI to make its knowledge and decisions more transparent. Such systems might have the potential to exhibit more theoretical and practical rationality—in the general sense of these concepts.

### 8.2 Overlaps with Rationality

There is hardly any situation in which our beliefs and actions are driven, or should be evaluated, solely by the norms of rationality. We are of course aware that rationality issues strongly overlap with many other concepts and facets of human mental life that we have not yet covered in this introductory chapter. So, let us at least briefly address some relations to other topics in the vicinity of rationality.

One such topic is the relation between rationality and *emotions*. It surfaces a little in some chapters (chapter 8.1 by Grüne-Yanoff; chapter 8.6 by Thompson, Elqayam, & Ackerman; and chapter 9.4 by Dhami & al-Nowaihi), but it is nowhere systematically treated in this handbook. This may be surprising for psychologists, as the general relationship between emotion and cognition is an issue with enormous theoretical and practical relevance (Damasio, 1994). The situation looks quite different, however, if we look at the more specific relationship between rationality and emotions. On the one hand, neuropsychological studies with patients show that people who are incapable of experiencing emotions but have retained their cognitive abilities can be seriously impaired in certain decision-making tasks (Dimitrov, Phipps, Zahn, & Grafman, 1999). On the other hand, psychological laypeople and experts tend to assume that a "cool head" thinks better than a "hot head" (Blanchette, 2013). Only recently has this topic been systematically investigated in the field of logical thinking. Several questions have played a role here: one is whether the current affective state of a person influences her ability to think logically, independently of the content of the reasoning task. Another question is whether the emotional content of a reasoning problem itself influences logical reasoning performance. And, of course, we can ask how the emotional content of the problem and the person's affective state interact with each other.

The findings are still equivocal. Some empirical studies show that people who witnessed the terror attack in London on July 7, 2005, at close range and who thus had strong emotions performed better than other people on logical reasoning problems with contents related to terror (Blanchette, Richards, Melnyk, & Lavda, 2007). But laboratory studies show that both positive and negative emotional states of the reasoner can have a devastating effect on reasoning performance (Jung, Wranke, Hamburger, & Knauff, 2014). And interactions between the emotional value of a content and the emotional state of the reasoner have seldom been reported (Blanchette, Caparos, & Trémolière, 2018). We cannot discuss all these interesting findings, although it is likely that inconsistent results might also have to do with the fact that for ethical reasons, strong emotions cannot be induced in laboratory research. Moreover, the many different kinds of long-term and short-term affects, sentiments, feelings, emotions, or moods may interact with rational reasoning in many different ways. They can, for instance, put additional load on the cognitive system, but can also be adaptive and help to avoid hazardous situations (De Jong, Mayer, & Van Den Hout, 1997). Already in the 1960s, Schachter and Singer (1962) argued that the experience of emotion consists of two components: physical arousal and its cognitive interpretation. In other words, when an emotion is felt, the person is in a particular physiological state and must figure out what caused this arousal. Since our bodily states are highly ambiguous, emotion itself requires a cognitive act of interpretation and thus of reasoning (Chater, 2018).

This leads directly to the main question for philosophers: can emotions, too, be normatively assessed with regard to their rationality? In subsection 2.2 of this introductory chapter, we put this question aside and focused on actions and beliefs as the sole objects of rationality assessments. However, one might also have the view that emotions do not only have causes, do not just befall us, but also have reasons and are amenable to reasons. We all are familiar with the fact that one is not automatically exculpated by saying, "Sorry, but this is how I feel." People can have more or less "adequate" or, conversely, exaggerated or even pathogenic feelings and emotional responses. They can have persistent emotions whose cause has long dissolved. There can be an absence (or suppression) of feelings where one would expect them to show, and so on. However, even if we grant that emotions are responsive to reasons, one may doubt that assessing the adequacy of emotions is already a judgment about their rationality. The topic is richly discussed (see, e.g., de Sousa, 2011; D. Evans & Cruse, 2004; Helm, 2001).

Again, we have preferred not to cover this topic in the handbook.

A connected topic is the relation between rationality and *morality*. Here, emotions are involved in two ways. The immediate connection is pursued by *sentimentalist theories* of morality, which we find, for instance, in the work of Hume (1751/1975a) and Schopenhauer (1841). Clearly, we have particularly moral sentiments: positive ones like love, sympathy, and compassion and negative ones like remorse, shame, blame, and outrage. For sentimentalist theories, these moral feelings constitute the origin of morality and lie at the basis for justifying moral judgments.

However, there is also an indirect connection. Even if one does not appeal to particular moral sentiments, one may locate the basis of all our evaluations in how we feel. We ultimately strive for happiness, and we seek pleasure and try to avoid pain—this is the traditional terminology, which had a broader meaning than it does in current everyday language. In this picture, our ultimate values refer to certain positive or negative emotional or hedonic states, and then morality is about the general realization of these values. Ancient *hedonism* refers to pleasures and pains, and *eudaimonism* to happiness (though *eudaimonia* was not conceived as just a psychological state). We find it in Jeremy Bentham's (1789/1970) *hedonic utilitarianism*. And we find it in *contractarianism*, originating with Hobbes (1651/1994), where the social contract enforcing moral behavior guarantees the individual pursuit of happiness and the avoidance of mutual harm. These moral conceptions were driven by the idea that morality can be explained by the rational pursuit of these ultimate values. This would be a strong justification: morality would just be instrumental to those values.

Kant (1788/1908), of course, is famous for being opposed to all of this. For him, the first principle of morality, the *categorical imperative*, is an a priori principle of pure reason and thus derives from rationality alone, without appealing to pains, pleasures, or other feelings. This obviously presupposes a stronger notion of rationality, transcending Hume's instrumental rationality (see also chapter 1.1 by Sturm, this handbook). Again, these issues are much too large to be treated properly in this handbook, but they are reflected in chapter 12.1 by Fehige and Wessels and chapter 12.2 by Smith (both in this handbook).

While philosophers reflect on the nature of moral sentiments and their relation to rationality, cognitive psychology and neuroscience conduct experiments to understand how moral values interact with other cognitive and noncognitive processes. One active research area in this context is *law and legal reasoning*. Do people use moral values or legal norms when thinking about crimes and court decisions? Empirical studies show that legal experts such as judges and lawyers can follow the rational norms of the legal system, whereas laypeople often base their opinions about crimes and how they should be punished on their intuitive feeling of rightness and morality. In particular, if they judge a crime to be exceptionally immoral and evil, they tend to demand very severe punishments. And they do this even if, from a legal point of view, there are mitigating or even exculpatory conditions, just to satisfy their need to punish the transgressor (Gazzo Castañeda & Knauff, 2016). Legal logic and logical models of legal argumentation are the topics of two quite different chapters in this handbook: chapter 11.3 by Hilgendorf and chapter 11.4 by Prakken.

Another line of experimental research actually goes back to a thought experiment by Karl Engisch, a German jurist and philosopher of law who considered in his study (1930, p. 288) what is nowadays called the "trolley problem": imagine you are standing on a footbridge and see a runaway trolley moving toward five people lying on the main track. You are standing next to a lever that controls a switch. If you pull the lever, the trolley will be redirected onto a side track and the five people on the main track will be saved. However, there is a single person lying on the side track. You have two options: do nothing and allow the trolley to kill the five people on the main track, or pull the lever and cause the trolley to kill the single person on the side track. Is one option morally required?

This thought experiment has been much discussed in philosophy to differentiate between deontological and consequentialist, in particular utilitarian, accounts of morality. Utilitarians would recommend pulling the lever because it results in fewer fatalities, while deontologists would object that human life is an incommensurable good, so that counting fatalities is not necessarily an argument. Instead, they would argue that it is better to *refrain* from pulling the lever, and thus to let a greater number die, than to actively *do* something resulting in the death of (fewer) persons.

Psychologists, by contrast, investigate how people actually (say they would) decide. They also compare this scenario to a slightly different situation: imagine that a fat man is standing next to you, and instead of turning a lever, you could push him on the track, thereby again saving the group of five but killing the fat man. How do people decide now?

Most of the empirical results show that persons decide differently in the two situations (Awad, Dsouza, Shariff, Rahwan, & Bonnefon, 2020). Is that rational? A PhD

student of philosophy, Joshua Greene, was the first who used this dilemma in a functional brain-imaging study. Greene, Sommerville, Nystrom, Darley, and Cohen (2001) showed that the "personal" dilemma (pushing the man off the footbridge) activated brain regions associated with emotions, while the "impersonal" dilemma (flipping a lever) engages regions associated with controlled, rational reasoning. These findings (and their interpretations) are problematic for many reasons (Waldmann, Nagel, & Wiegmann, 2012) but are related to many different versions of dual-process accounts of moral decision making, which are still popular in cognitive research (see chapter 2.5 by Klauer, this handbook). Again, the topic of morality is too comprehensive to receive treatment in this handbook, but chapter 12.3 by Wiegmann and Sauer (in this handbook) comments on it from the psychological side.

By the way, it may be unsurprising that moral research today gets much attention from military and engineering fields, for instance, in the development of autonomous weapons or vehicles, for which situations can occur where one or another potentially fatal collision is unavoidable. In such cases, the car's software has to "decide" what to crash into and which casualties to hazard. In 2018, the *moral machine project* began to collect information on such decisions on an Internet platform, where people can choose between two different destructive outcomes (Awad et al., 2018). This looks like a dangerous blurring of the border between empirical decisions and normative theories, which we try so carefully to reflect in this handbook. Since Hume, we know that we better avoid inferences from *is* to *ought*.

We may widen the perspective even further, from the relation between morality and rationality to the general *cultural, social, and political embedding* of rationality. Since Plato, political philosophy has been deeply involved in rationality issues. In the *Republic*, he described how society should be rationally organized: ruled by a leading class that is rational, intelligent, and self-controlled (Plato, 2013). Hobbes, in the *Leviathan* (1651/1994), argued that citizens are only protected from mutual violence, as it occurs in the state of nature, if power is delegated to a reasonable sovereign. Later, Jean-Jacques Rousseau, in *The Social Contract* (1762), challenged this view by arguing that all members of a society have the same rights and duties. Although these issues are superimposed by more important topics of political philosophy, such as liberty, equality, and justice, they also have much to do with rationality, yet could not be covered in this handbook. The lack becomes particularly apparent when we consider John Rawls's (1971) theory of *justice as fairness*, in which the so-called veil of ignorance creates a situation in which rational decision making arguably justifies his principles of justice. But then those principles are at issue, not the theory of rational decision making. Hence, we preferred to abstain from extending the handbook in this direction. It is reflected, to some extent, in chapter 10.5 by Nida-Rümelin, Gutwald, and Zuber (this handbook).

We have also abstained from engaging in intercultural considerations. A review of the sparse literature that is available in languages we understand suggests that there is a wide range of alternatives to the conception of rationality in the European and Northern American intellectual tradition. For instance, a core distinction in this handbook is that between normative and descriptive views on rationality. The distinction is so deeply entrenched in our scientific thinking that we can hardly imagine doing without it. But is this necessary? Perhaps not. For instance, the distinction does not seem to be clearly present in traditional Chinese thought. Instead of this dichotomy, we find a general tendency in traditional China to think of things as interactions of *yin* and *yang* and the more fluid interactions of the "five phases," water, wood, fire, metal, and earth (Marchal & Wenzel, 2017). Comparative ethnologists have found various other examples of fundamentally different ways of thinking: people in Melanesia and Polynesia use numerical systems that significantly differ from the norms of Western mathematics (Beller & Bender, 2008). Islam provides alternatives to the economic rationality of rational choice theory (Tafer, Boussahmine, & Bouanini, 2016), and studies looking at the brains of people playing a fairness game found very different neural activities in Buddhist meditators and Western participants (Kirk, Downar, & Montague, 2011). Such observations challenge our arrogance that rationality everywhere can be judged by a single set of criteria. It is not unlikely that we find diverging conceptions of rationality in different cultures and different socioeconomic systems (Lloyd, 2017). Relatedly, Henrich, Heine, and Norenzayan (2010) have argued that psychologists and other behavioral scientists often make strong claims based on experiments with samples drawn entirely from *Western, Educated, Industrialized, Rich, and Democratic* (WEIRD) societies. However, we confess that the present handbook is committed to the Western intellectual tradition in psychology and philosophy.

This is the right place to finally explain why so many areas of so-called continental philosophy are not represented in this handbook. We do not want to suggest that they have nothing substantial to say about rationality. In fact, reflections on reasons and rationality are ubiquitous also in these parts of philosophy. For instance, consistency is central to most concepts of rationality,

but in *dialectic thinking*, contradictory hypotheses lead to a rational synthesis. Dialectic thinking is important in the work of Georg Friedrich Wilhelm Hegel but not represented in this handbook. Friedrich Nietzsche is famous for his *perspectivism* (see, e.g., R. L. Anderson, 1998; Danto, 1965, chapter 3). His criticism was that scholars are often not aware of their own perspective and thus do not control its influence on their work. Maybe his demand for perspectival deconstruction should be part of epistemic rationality, but it is neither treated in this handbook nor exercised by its authors. Or think of the profound criticism of Western *technical rationality* in Martin Heidegger (1954). It may be misguided, but it springs from deep philosophical sources and is still disconcerting. Michel Foucault has tremendously shaped our intellectual landscape. The power to delimit reason is key to his early analysis of madness (Foucault, 1972), and his later analysis of government shows the historical conditions of the possibility of different types of governmental rationalities (Foucault, 2008). And when Habermas is discussed under the heading of *communicative rationality* in chapter 10.3 by Meggle (this handbook), this is restricted to the narrow perspective of philosophy of language and neglects Habermas's much more general political intentions concerning the theory of democracy.

So why is this handbook so obviously biased against continental philosophy and toward analytic philosophy, and within the latter toward the more formal ways of philosophizing? Before we explain this bias, we should mention that already the alleged opposition between "analytic" and "continental" is infelicitous, as "analytic" vaguely designates certain ways of philosophizing, which, however, still diverge dramatically, and "continental" refers to geographical locations and origins, which again have produced quite different strands and styles of philosophizing.

But even if we follow this traditional terminology, there are several reasons why this handbook is biased toward analytic philosophy. One reason is that the handbook brings together cognitive psychology and philosophy, and that analytic philosophy lies much closer to the intersection of these disciplines than other parts of philosophy. This is why analytic philosophers and cognitive psychologists collaborate so successfully under the roof of cognitive science. For the same reason, we also have a certain bias toward the more formal side of analytic philosophy, which provides more specific and detailed accounts of the norms of rationality and has more natural connections to cognitive theories of the human mind. Another, even more fundamental reason for our bias is that continental philosophy often focuses on the wider cultural and political context of rationality, as our examples above indicate. However, as just said, this handbook is silent on this wider context. Although such perspectives on rationality are interesting and important, they do not fit well with the predominantly individualistic orientation of psychological research and are difficult to investigate with the methods of experimental psychology. Analytic philosophers have similar reservations. For them, such political and cultural associations are often sweeping and hard to seize—none of these accounts could be subject to sophisticated normative discourse, as we find it in analytic philosophy. So, we had to draw a line here, and therefore the contributions of continental philosophy are not represented in this handbook.

## 9. Conclusion and Open Issues

The aim of this chapter was to offer the reader a guide through the complex field of rationality research. We hope that we could make clear why we consider the distinction between theoretical and practical rationality so important. We also emphasized why the relation between normative and descriptive theories of rationality is so essential for our philosophico-psychological enterprise and why research on social rationality should go beyond competitive scenarios in which the actors consider only their individual benefits. We moreover explained why it is important that rationality research should focus not only on the rational assessment of the *results* of a thought process but also on the cognitive processes that led from the input to the output. Furthermore, we introduced the concept of a *double equilibrium*, which requires empirical research to listen to normative theorizing and, conversely, reminds us that normative considerations are always defeasible and must withstand critical, empirical scrutiny. In the final sections, we have discussed some cognitive prerequisites of rationality and argued against the view that many specialized conceptions of rationality are needed for the many areas of everyday life, public discourse, and the sciences and humanities. Our position was that rationality assessments in most of these areas can rely on general concepts of rationality.

Of course, there are many open questions and tasks that rationality research still has to tackle. One of these tasks is to overcome the fixation on logic and probability theory, which is particularly prevalent in the psychology of reasoning. We have shown that philosophy has for decades offered a richer set of options. Psychology has just started to pay more attention to such accounts, and we hope that it will continue to go in that direction.

While this is an issue of epistemic rationality, we see a related challenge in practical rationality. Here, we think that it is important to further develop alternatives to standard decision theory. Again, we believe that psychology and philosophy should go beyond these limitations of present research. There are developments in this direction, but currently it seems that philosophy and psychology are pursuing different lines. More integration would be highly desirable.

A related observation is that most of the current work under the heading of social rationality is in fact about individual rationality in social situations. We think this is a weakness of current research that should be overcome. In so many situations of everyday life, people seek to think and act rationally in concert with others, but we do not know much about the underlying cognitive processes. This is a research field where social psychologists, cognitive psychologists, social scientists, and philosophers should collaborate more intensively. We also do not know enough about the normative standards that we should use to assess collective forms of rationality, which is another area where we think that rationality research must make progress. This is important for many fundamental questions of rationality, but it is obvious that this topic is also socially and politically highly relevant.

Another characteristic of current research is that in the various subfields, the process-oriented perspective is pursued sometimes more and sometimes less intensely. Surely this is partly due to the varying difficulties of doing so. Some theories from philosophy may be really hard to bring into process-oriented terms. Still, we think that this is desirable and would strengthen rationality research where it is yet weak.

We also think that philosophers and psychologists should try to reach more consensus on the nature of concepts and meaning. Currently, there seems to be much dissent on this topic between the two disciplines, which affects many areas of cognitive science. For the same reason, we did not discuss the extremely complex question of what we mean when we talk about truth. Truth may be the ultimate goal of rationality, but there are so many different conceptions of truth in philosophy and psychology that it is hard to see how a more unified view could be reached in the near future. Fortunately, this is not primarily a disagreement about rationality. However, in the long run, more collaboration on the topics of meaning and truth between the different disciplines would certainly also be beneficial for our understanding of rationality.

All these issues for future research need much more interdisciplinary interaction. We think that this handbook can show that by such interdisciplinary collaboration, research into rationality has already made great progress. Today we have much more knowledge about human rationality than each of our disciplines could have achieved on their own. However, there is still room for further progress. For instance, it is essential that all parties in rationality research be clear about the complex relation between descriptive and normative considerations. Unclarity about this relation still seems to hamper research. And we think that other disciplines should be more involved in our interdisciplinary endeavor. Linguists, for instance, have substantial things to say about rationality in language, pedagogues and education researchers should systematically study how students can be trained in rational thinking, and anthropologists and ethnologists can help us better understand how rationality is rooted in different cultures and human lifestyles.

We began this introductory chapter with the Aristotelian definition of humans as rational animals, and we mentioned a few objections to this view. So, are we really as rational and smart as we think? Or is this more a self-delusion that persists even if empirical findings put our optimism into question? The intention of this chapter was not to answer this question. That would be presumptuous. What we wanted to do, rather, was to help our readers to develop their own opinion on this question. This opinion, we hope, might benefit from the interdisciplinary and multiperspectival character of this handbook. In the end, we may agree with how Christof Koch (2016, p. 25) summarized the lessons from Darwin's theory of evolution: we are unique, but so is every other species, each in its own way.

## Notes

1. There are many labels for this and similar distinctions: "is–ought," "ontic–deontic," "descriptive–prescriptive," "empirical–normative." We have decided to use the pair "descriptive–normative" throughout and thus follow the usage of Elqayam and Evans (2011).

2. The Priority Programme "New Frameworks of Rationality" was an important step toward improved collaboration between psychologists, philosophers, and researchers in other disciplines in the field of rationality (see the Preface in this handbook).

3. For all this, see the entries "Ratio," "Rationalität," and "Vernunft, Verstand" in Ritter, Gründer, and Gabriel (1971–2007).

4. Section 2 of chapter 2.1 by Broome (this handbook) also contains detailed remarks about the etymology of the English word "reason."

5. The other two are his famous context principle and his rule to strictly distinguish object and concept.

6. Knauff and Gazzo Castañeda (2021) criticize this terminology and show that the new trend toward subjective degrees of belief cannot be regarded as a "paradigm shift" in the Kuhnian sense.

7. Note, though, that the label "Bayesianism" can take on quite different meanings in psychology (Jones & Love, 2011).

8. A philosopher would say that we should talk of "doxastic" attitudes. Doxastic attitudes are belief attitudes in all shades and grades: opinions, doubts, subjective probabilities, and so on, while epistemic attitudes, strictly speaking, concern only knowledge. Still, "epistemic attitude" is the more familiar term. We shall therefore use it always in the wide sense of "doxastic attitude." Of course, the philosopher would add that this simply conforms to the sloppy talk about knowledge found throughout the sciences, which hardly distinguishes knowledge from true or certain belief. And he would refer to the huge philosophical discussion of knowledge in the past 50 years, which is largely neglected by those sciences (see, e.g., Bernecker & Pritchard, 2011, particularly parts II–IV and VII; or Williamson, 2000), as the main advocate of the opinion that the study of knowledge even precedes the study of belief). Let this remark suffice to indicate that the concept of knowledge is a philosophical snakepit from which we keep a healthy distance despite our talk of epistemic attitudes.

9. After the heyday of so-called logical behaviorists like Ryle, who subsumed action explanations under the special kind of (noncausal) dispositional explanation, the view that actions are caused by intentions in an ordinary way was initiated by Hempel (1961–1962) and Davidson (1963) and is still the dominant view in philosophy (see, e.g., Bratman, 2006).

10. The following considerations are more extensively explained in Spohn (2002).

11. Many philosophers say that beliefs are rational relative to the *facts*, or that only facts can be reasons for beliefs. But then these facts must be perceived and hence believed, and so this relation is again mediated by beliefs.

12. This may be taken as a modern version of Berkeley's dictum that only ideas can be similar to ideas (cf. Berkeley, 1710/1949, section 8).

13. The situation resembles the Agrippa trilemma, often taken to be the basic problem of epistemology (see, e.g., BonJour, 1985), according to which a belief is justified only if it is justified by reasons that are in turn justified. The trilemma has a skeptical force: how, then, could a belief ever be justified? Many versions of foundationalism and coherentism, justification internalism, reliabilism, and contextualism respond to this challenge. We need not discuss this here. Certainly, though, the situation vis-à-vis reasonableness is similar.

14. Spohn (2002) expands on these workings of rationality assessments.

15. Elqayam and Evans (2011, p. 238) similarly speak of the "product" and the "process level."

16. In philosophy, the debate about the nature of those contents, as well as about the nature of the concepts of which they consist, is endless, most sophisticated, and without conclusion (see, e.g., Fodor, 1998; García-Carpintero & Macià, 2006).

17. In fact, the formal structure of probabilities was not firmly fixed from the beginning; see Shafer (1978).

18. A plea for *epistemic dualism* is found in Spohn (2012, chapter 10); one of several attempts for unification is found in Leitgeb (2017). For an attempt at connecting the issue with psychological research, see also Weisberg (2020).

19. The conception of intentional objects goes back to Brentano (1874). The book that alerted philosophers most to the problems surrounding intentional objects, propositions, and the like was Quine (1960). The subsequent discussion is endless, without a clear conclusion. For a recent treatment, see Recanati (2016). See also the plea for intentional objects in Spohn (2009, chapter 16).

20. The basic algorithm is the so-called roll-back analysis as explained in Raiffa (1968).

21. Often, though, the issue is not individual rationality but rather policy: how to design institutions when these are assumed to have an influence on people's preferences?

22. There is no textbook summarizing all these alternative accounts. One may consult the comprehensive handbook by Glanzberg (2018). The richest, though critical, discussion of antirealistic truth theories in analytic philosophy is found in Devitt (1991).

23. Such thoughts are also vigorously discussed in philosophy, based on the brain-in-a-vat scenario of Putnam (1981, chapter 1).

24. The term was introduced by C. I. Lewis (1912).

25. Grice (1975) tried to save this representation with the help of his theory of implicatures; see also Bennett (2003, chapter 2).

26. It was a central problem for logical empiricism to explain how we can understand not directly observable dispositions on the basis of observational language and extensional logic. (Philosophers have a more general understanding of dispositions than psychologists; see also section 6 of this introductory chapter.) This problem turned out to be a Kuhnian anomaly for logical empiricism and eventually led to its breakdown. The problem was precisely how to understand conditionals like "If that piece of sugar were put into water, it would dissolve" (= "That piece of sugar is soluble") in terms of classical logic and material implication. Compare, for example,

Hempel's paradigmatic movement from Hempel (1958) to Hempel (1973).

27. So, this was at the same time when the Wason selection task was designed. A different attempt to cope with these paradoxes is relevance logic, developed by A. R. Anderson and Belnap (1975). It is, however, not represented in this handbook.

28. Cohen (1981) gives a similar description, although he distinguishes between a narrow and a wide reflective equilibrium.

29. It is not only a methodological issue. Philosophers tend to carry it deeply into metaphysics and semantics, usually within metaethics and with respect to morality and moral predicates and properties. However, with respect to normativity, morality and rationality are in the same boat. Thus, these philosophical discussions, as pursued, for example, in Ridge (2019), carry over to rationality.

30. Hume (1975c, p. 469) famously said, referring to the transition from *is* to *ought*, that it "seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it." The label "naturalistic fallacy" was introduced by Moore (1903, §10) for the related fallacy of inferring the goodness of something solely from its descriptive properties.

31. Paradigmatically by Hempel (1961–1962).

32. To our knowledge, this kind of double reflective equilibrium was first described in Spohn (1993, p. 188). See also Spohn (2002, section 4).

33. As is more fully argued in Spohn (2011).

34. Rich textbooks are, for example, Myerson (1991) and Maschler, Solan, and Zamir (2013). See also section 9.1 by Albert and Kliemt (this handbook).

35. The distinction between social rationality and individual rationality we are invoking here is more salient in the practical context. That is why only this context is considered here. However, it pertains just as well to the epistemic context: see the distinction between epistemology *in* groups versus epistemology *of* groups in chapter 10.1 by Dietrich and Spiekermann (this handbook).

36. There are, however, many other paradigmatic and highly interesting social situations modeled by game theory. For brief explanations, see chapter 9.1 by Albert and Kliemt (this handbook).

37. It has been important to logical behaviorists like Ryle (1949). For a general discussion of the philosophical intricacies of dispositions, see, for example, Mumford (1998).

38. One could argue that the so-called replication crisis (that many published findings could not be replicated in subsequent studies) is a drastic example of the limits of epistemic rationality in science (Open Science Collaboration, 2015). However, many researchers have questioned the popular interpretation of the relatively low rates of replicability. For example, a "regression to the mean" can explain why even if the initial results were correct, they need not be replicable in subsequent studies (Fiedler & Prager, 2018). The publication from 2015 also showed that replicability varies strongly over different subdisciplines and that replication rates in cognitive psychology are better than in other areas.

## References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617.

Adorno, T. W., Dahrendorf, R., Pilot, H., Albert, H., Habermas, J., & Popper, K. R. (1969). *Der Positivismusstreit in der deutschen Soziologie* [The positivism dispute in German sociology]. Neuwied, Germany: Luchterhand.

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

Anderson, A. R., & Belnap, N. D., jr. (1975). *Entailment: The logic of relevance and necessity*. Princeton, NJ: Princeton University Press.

Anderson, J. R. (2000). *Cognitive psychology and its implications* (5th ed.). New York, NY: Worth.

Anderson, R. L. (1998). Truth and objectivity in perspectivism. *Synthese*, *115*, 1–32.

Aristotle. (1984). *De anima* [On the soul]. In J. Barnes (Ed.), *Complete works of Aristotle: Vol. 1. The revised Oxford translation* (pp. 641–692). Princeton, NJ: Princeton University Press.

Aristotle. (1985). *Ars rhetorica* [Rhetoric]. In J. Barnes (Ed.), *Complete works of Aristotle: Vol. 2. The revised Oxford translation* (pp. 2152–2269). Princeton, NJ: Princeton University Press.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In L. W. Porter, H. L. Angle, & R. W. Allen (Eds.), *Organizational influence processes* (2nd ed., pp. 295–303). Armonk, NY: Sharpe. (Original work published 1951)

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., . . . Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64.

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, *117*(5), 2332–2337.

Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences*, *6*(1), 47–52.

Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton, NJ: Princeton University Press.

Baddeley, A. D. (1986). *Working memory*. Oxford, England: Oxford University Press.

Bara, B. G., Bucciarelli, M., & Lombardo, V. (2001). Model theory of deduction: A unified computational approach. *Cognitive Science*, *25*(6), 839–901.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.). (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. New York, NY: Oxford University Press.

Bartlett, F. (1995). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press. (Original work published 1932)

Bayes, T. (1970). An essay towards solving a problem in the doctrine of chances. In E. S. Pearson & M. G. Kendall (Eds.), *Studies in the history of statistics and probability*. London, England: Griffin. (Original work published 1764)

Bazerman, M. H., & Moore, D. A. (2013). *Judgment in managerial decision making* (8th ed.). Chichester, England: Wiley. (Original work published 2002)

Beals, K. L., Smith, C. L., Dodd, S. M., Angel, J. L., Armstrong, E., Blumenberg, B., . . . Trinkaus, E. (1984). Brain size, cranial morphology, climate, and time machines. *Current Anthropology*, *25*(3), 301–330.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*, 421–441.

Beller, S. (1999). Wenn Wissen logisches Denken erleichtert bzw. zu verhindern scheint: Inhaltseffekte in Wasons Wahlaufgabe [When knowledge facilitates or apparently prevents logical reasoning: Content effects in the Wason selection task]. In H. Gruber, W. Mack, & A. Ziegler (Eds.), *Wissen und Denken: Beiträge aus Problemlösepsychologie und Wissenspsychologie* (pp. 35–52). Wiesbaden, Germany: Deutscher Universitätsverlag.

Beller, S., & Bender, A. (2008). The limits of counting: Numerical cognition between evolution and culture. *Science*, *319*(5860), 213–215.

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford, England: Oxford University Press.

Bentham, J. (1970). *An introduction to the principles of morals and legislation* (H. Burns & H. L. A. Hart, Eds.). Oxford, England: Oxford University Press. (Original work published 1789)

Berkeley, G. (1949). *Treatise concerning the principles of human knowledge*. In A. A. Luce & T. E. Jessop (Eds.), *The works of George Berkeley, Bishop of Cloyne* (Vol. 2). London, England: Thomas Nelson & Sons. (Original work published 1710)

Bernecker, S., & Pritchard, D. (Eds.). (2011). *The Routledge companion to epistemology*. London, England: Routledge.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, *52*, 1007–1028.

Binkley, R. (1968). The surprise examination in modal logic. *Journal of Philosophy*, *65*, 127–136.

Blanchette, I. (Ed.). (2013). *Emotion and reasoning*. Hove, England: Routledge.

Blanchette, I., Caparos, S., & Trémolière, B. (2018). Emotion and reasoning. In *The Routledge international handbook of thinking and reasoning* (pp. 57–70). New York, NY: Routledge/Taylor & Francis.

Blanchette, I., Richards, A., Melnyk, L., & Lavda, A. (2007). Reasoning about emotional contents following shocking terrorist attacks: A tale of three cities. *Journal of Experimental Psychology: Applied*, *13*(1), 47–56.

BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.

Boole, G. (1951). *An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities*. New York, NY: Dover. (Original work published 1854)

Bossaerts, P., & Murawski, C. (2017). Computational complexity and human decision-making. *Trends in Cognitive Science*, *12*, 917–929.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, *402*(6758), 179–181.

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature*, *36*(1), 75–111.

Braine, M. D. S. (1990). The "natural logic" approach to reasoning. In W. F. Overton (Ed.), *Reasoning, necessity, and logic: Developmental perspectives* (The Jean Piaget Symposium Series 16, pp. 133–157). Hillsdale, NJ: Erlbaum.

Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Mahwah, NJ: Erlbaum.

Brandom, R. B. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Cambridge, MA: Harvard University Press.

Brandom, R. B. (2000). *Articulating reasons: An introduction to inferentialism*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (2006). *Structures of agency: Essays*. Oxford, England: Oxford University Press.

Brentano, F. (1874). *Psychologie vom empirischen Standpunkte* [Psychology from an empirical point of view]. Leipzig, Germany: Duncker & Humblot.

Brodbeck, F. C., Kerschreiter, R., Mojzisch, A., & Schulz-Hardt, S. (2007). Group decision making under conditions of distributed knowledge: The information asymmetries model. *Academy of Management Review*, *32*(2), 459–479.

Bruner, J. S. (1957). Going beyond the information given. In E. H. Gruber, K. R. Hammond, & R. Jessor (Eds.), *Contemporary approaches to cognition: A symposium held at the University of Colorado* (pp. 41–69). Cambridge, MA: Harvard University Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–192.

Carnap, R. (1971). A basic system of inductive logic: Part I. In R. Carnap & R. C. Jeffrey (Eds.), *Studies in inductive logic and probability* (Vol. I, pp. 33–165). Berkeley: University of California Press.

Carnap, R. (1980). A basic system of inductive logic: Part II. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. II, pp. 7–155). Berkeley: University of California Press.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford, England: Oxford University Press.

Chan, T. (Ed.). (2013). *The aim of belief*. Oxford, England: Oxford University Press.

Chater, N. (2018). *The mind is flat: The illusion of mental depth and the improvised mind*. New Haven, CT: Yale University Press.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*(4), 391–416.

Claidière, N., Trouche, E., & Mercier, H. (2017). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General*, *146*(7), 1052–1066.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*(3), 317–331.

Cole, M. W., Yarkoni, T., Repovš, G., Anticevic, A., & Braver, T. S. (2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *Journal of Neuroscience*, *32*(26), 8988–8999.

Colombo, M., & Knauff, M. (2020). Editors' review and introduction: Levels of explanation: From molecules to culture. *Topics in Cognitive Science*, *12*, 1224–1240.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason and the human brain*. New York, NY: Grosset/Putnam.

Danto, A. C. (1965). *Nietzsche as philosopher*. New York, NY: Columbia University Press.

Darwin, C. (1859). *On the origin of species*. London, England: Murray.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, *60*, 685–700.

de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. [Foresight: Its logical laws, its subjective sources]. *Annales de l'Institut Henri Poincaré*, *7*, 1–68.

Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. In S. Dehaene & Y. Christen (Eds.), *Characterizing consciousness: From cognition to the clinic?* (pp. 55–84). Berlin, Germany: Springer.

De Jong, P. J., Mayer, B., & Van Den Hout, M. (1997). Conditional reasoning and phobic fear: Evidence for a fear-confirming reasoning pattern. *Behaviour Research and Therapy*, *35*(6), 507–516.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38.

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.

Descartes, R. (1641). *Meditationes de prima philosophia* [Meditations on first philosophy]. Paris, France: Michel Soly.

de Sousa, R. (2011). *Emotional truth*. Oxford, England: Oxford University Press.

de Vega, M., Glenberg, A. M., & Graesser, A. C. (2008). *Symbols and embodiment: Debates on meaning and cognition*. Oxford, England: Oxford University Press.

Devitt, M. (1991). *Realism and truth* (2nd ed.) Oxford, England: Blackwell.

Dimitrov, M., Phipps, M., Zahn, T. P., & Grafman, J. (1999). A thoroughly modern Gage. *Neurocase*, *5*(4), 345–354.

Dörner, D. (1999). *Bauplan für eine Seele* [Construction plan for a soul]. Reinbek bei Hamburg, Germany: Rowohlt.

Dunlosky, J., & Bjork, R. A. (2008). The integrated nature of metamemory and memory. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 11–28). New York, NY: Psychology Press.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, *21*(4), 419–460.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, *75*, 643–669.

Elqayam, S., & Evans, J. St. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233–248.

Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Thinking & Reasoning*, *19*(3–4), 249–265.

Elqayam, S., & Over, D. E. (2016). From is to ought: The place of normative models in the study of human thought. *Frontiers in Psychology*, *7*, 628.

Engisch, K. (1930). *Untersuchungen über Vorsatz und Fahrlässigkeit im Strafrecht* [Studies on intent and negligence in criminal law]. Berlin, Germany: Liebermann.

Esser, J. K. (1998). Alive and well after 25 years: A review of groupthink research. *Organizational Behavior and Human Decision Processes*, *73*(2–3), 116–141.

Evans, D., & Cruse, P. (Eds.). (2004). *Emotion, evolution, and rationality*. Oxford, England: Oxford University Press.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, England: Erlbaum.

Evans, J. St. B. T. (1993). Bias and rationality. In K. I. Manktelow & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 6–30). London, England: Taylor & Francis/Routledge.

Evans, J. St. B. T. (2018). Dual-process theories. In *The Routledge international handbook of thinking and reasoning* (pp. 151–166). New York, NY: Routledge/Taylor & Francis.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, England: Erlbaum.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.

Fangmeier, T., Knauff, M., Ruff, C. C., & Sloutsky, V. (2006). fMRI evidence for a three-stage model of deductive reasoning. *Journal of Cognitive Neuroscience*, *18*(3), 320–334.

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology*, *40*, 1–10.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function introduction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1594), 1280–1286.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford, England: Clarendon Press.

Fodor, J. A. (2000). *The mind doesn't work that way*. Cambridge, MA: MIT Press.

Foucault, M. (1972). *L'histoire de la folie à l'âge classique* [History of madness]. Paris, France: Gallimard. (First published as *Folie et déraison*, Paris, France: Plon, 1961)

Foucault, M. (2008). *Le gouvernement de soi et des autres: Cours au Collège de France 1982–1983* [The government of self and others]. Paris, France: du Seull/Gallimard.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*, 5–20.

Frege, G. (1884). *Die Grundlagen der Arithmetik: Eine logisch mathematische Untersuchung über den Begriff der Zahl* [The foundations of arithmetic: A logico-mathematical enquiry into the concept of number]. Breslau, Germany: Marcus.

Frege, G. (1918). Der Gedanke. Eine logische Untersuchung [The thought]. *Beiträge zur Philosophie des deutschen Idealismus*, *1*, 58–77.

Frege, G. (1979). *Posthumous writings* (H. Hermes, F. Kambartel, & F. Kaulbach, Eds.; P. Long & R. White, Trans.). Oxford, England: Blackwell.

Frey, B. S., & Meier, S. (2002). *Two concerns about rational choice: Indoctrination and imperialism* (Zurich IEER Working Paper No. 104). Zurich, Switzerland: University of Zurich. Available at SSRN: https://ssrn.com/abstract=301867 or http://dx.doi.org/10.2139/ssrn.301867

Gabbay, D. M., Hogger, C. J., & Robinson, J. A. (Eds.). (1994). *Handbook of logic in artificial intelligence and logic programming: Vol. 3. Nonmonotonic reasoning and uncertain reasoning*. Oxford, England: Oxford University Press.

García-Carpintero, M., & Macià, J. (2006). *Two-dimensional semantics*. Oxford, England: Clarendon Press.

Gärdenfors, P. (1984). Epistemic importance and minimal changes of belief. *Australasian Journal of Philosophy*, *62*(2), 136–157.

Gärdenfors, P. (Ed.). (1992). *Belief revision*. Cambridge, England: Cambridge University Press.

Gazzo Castañeda, L. E., & Knauff, M. (2016). Defeasible reasoning with legal conditionals. *Memory & Cognition*, *44*(3), 499–517.

Gazzo Castañeda, L. E., Sklarek, B., Dal Mas, D., & Knauff, M. (2021). Probabilistic and deductive reasoning in the human brain. Manuscript submitted for publication.

Genç, E., Fraenz, C., Schlüter, C., Friedrich, P., Hossiep, R., Voelkle, M. C., . . . Jung, R. E. (2018). Diffusion markers of dendritic density and arborization in gray matter predict differences in intelligence. *Nature Communications*, *9*(1), 1905.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology*, *20*(6), 733–743.

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. New York, NY: Oxford University Press.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*(2), 127–171.

Gigerenzer, G., & Selten, R. (Eds.). (2002). *Bounded rationality: The adaptive toolbox*. Cambridge, MA: MIT Press.

Gillies, D. (2000). *Philosophical theories of probability*. London, England: Routledge.

Glanzberg, M. (Ed.). (2018). *The Oxford handbook of truth*. Oxford, England: Oxford University Press.

Gläscher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., . . . Tranel, D. (2012). Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proceedings of the National Academy of Sciences*, *109*(36), 14681–14686.

Gödel, K. (1931). Über formal unentscheidbare Sätze der ‚Principia mathematica' und verwandter Systeme I [On formally undecidable propositions of Principia Mathematica and related systems]. *Monatshefte für Mathematik und Physik*, *38*, 173–198.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, *11*(10), 435–441.

Goel, V., Gold, B., Kapur, S., & Houle, S. (1998). Neuroanatomical correlates of human reasoning. *Journal of Cognitive Neuroscience*, *10*(3), 293–302.

Goel, V., Shuren, J., Sheesley, L., & Grafman, J. (2004). Asymmetrical involvement of frontal lobes in social reasoning. *Brain*, *127*(4), 783–790.

Goldman, A. I. (1978). Epistemology and the psychology of belief. *The Monist*, *61*(4), 525–535.

Gomberg, P. (1989). Marxism and rationality. *American Philosophical Quarterly*, *26*(1), 53–62.

Gottinger, H. W. (1982). Computational costs and bounded rationality. In W. Stegmüller, W. Balzer, & W. Spohn (Eds.), *Philosophy of economics* (pp. 223–238). Berlin, Germany: Springer.

Gould, S. J. (1978). Sociobiology: The art of storytelling. *New Scientist*, *80*(1129), 530–533.

Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences*, *15*(2), 56–67.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Grice, H. P. (1957). Meaning. *Philosophical Review*, *66*, 377–388.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). New York, NY: Academic Press.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*, 367–388.

Habermas, J. (1968). *Erkenntnis und Interesse* [Knowledge and human interests]. Frankfurt/Main, Germany: Suhrkamp.

Habermas, J. (1973). *Legitimationsprobleme im Spätkapitalismus* [Legitimation problems in late capitalism]. Frankfurt/Main, Germany: Suhrkamp.

Habermas, J. (1981). *Theorie des kommunikativen Handelns* [The theory of communicative action] (Vols. 1–2). Frankfurt/Main, Germany: Suhrkamp.

Hacking, I. (1975). *The emergence of probability*. Cambridge, England: Cambridge University Press.

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

Hamburger, K., Ragni, M., Karimpur, H., Franzmeier, I., Wedell, F., & Knauff, M. (2018). TMS applied to V1 can facilitate reasoning. *Experimental Brain Research*, *236*(8), 2277–2286.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York, NY: Oxford University Press.

Hardman, D. (2009). *Judgment and decision making: Psychological perspectives*. Hoboken, NJ: Wiley.

Harsanyi, J. C. (1967). Games with incomplete information played by "Bayesian" players, Part I. The basic model. *Management Science*, *14*, 159–182.

Harsanyi, J. C. (1968a). Games with incomplete information played by "Bayesian" players, Part II. Bayesian equilibrium points. *Management Science*, *14*, 320–334.

Harsanyi, J. C. (1968b). Games with incomplete information played by "Bayesian" players, Part III. The basic probability distribution of the game. *Management Science*, *14*, 486–502.

Hart, H. L. A. (1961). *The concept of law*. Oxford, England: Oxford University Press.

Hartmann, S., & Rafiee Rad, S. (2020). Anchoring in deliberations. *Erkenntnis*, *85*, 1041–1069.

Heckhausen, J., & Heckhausen, H. (2018). *Motivation and action*. Berlin, Germany: Springer.

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, *9*, 467–479.

Heidegger, M. (1954). Die Frage nach der Technik [The question concerning technology]. In *Vorträge und Aufsätze* (pp. 9–40). Pfullingen, Germany: Neske.

Helm, B. W. (2001). *Emotional reason: Deliberation, motivation, and the nature of value*. Cambridge, England: Cambridge University Press.

Hempel, C. G. (1961–1962). Rational action. *Proceedings and Addresses of the American Philosophical Association*, *35*, 5–23.

Hempel, C. G. (1958). The theoretician's dilemma: A study in the logic of theory construction. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science* (Vol. II, pp. 37–98). Minneapolis: University of Minnesota Press.

Hempel, C. G. (1973). The meaning of theoretical terms: A critique of the standard empiricist construal. In P. Suppes, L. Henkin, A. Joja, & G. C. Moisil (Eds.), *Logic, methodology and philosophy of science IV* (pp. 367–378). Amsterdam, Netherlands: North-Holland.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Heyes, C. (2012). New thinking: The evolution of human cognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1599), 2091–2096.

Hinterecker, T., Knauff, M., & Johnson-Laird, P. N. (2016). Modality, probability, and mental models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1606–1620.

Hobbes, T. (1994). *Leviathan*. In E. Curley (Ed.), *Leviathan, with selected variants from the Latin edition of 1668*. Indianapolis, IN: Hackett. (Original work published 1651)

Hume, D. (1975a). *An enquiry concerning the principles of morals* (L. A. Selby-Bigge, Ed., 3rd ed., revised by P. H. Nidditch). Oxford, England: Clarendon Press. (Original work published 1751)

Hume, D. (1975b). *An enquiry into human understanding* (L. A. Selby-Bigge, Ed., 3rd ed., revised by P. H. Nidditch). Oxford, England: Clarendon Press. (Original work published 1748)

Hume, D. (1975c). *A treatise of human nature* (L. A. Selby-Bigge, Ed., 2nd ed., revised by P. H. Nidditch). Oxford, England: Clarendon Press. (Original work published 1739–1740)

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. New York, NY: Basic Books. (Original work published 1955)

James, W. (1956). The will to believe. In W. James, *The will to believe and other essays in popular philosophy* (pp. 1–31). New York, NY: Dover. (Original work published 1896)

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*. Oxford, England: Houghton Mifflin.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes* (2nd ed.). Boston, MA: Houghton Mifflin Harcourt.

Jeffrey, R. C. (1965). *The logic of decision*. Chicago, IL: University of Chicago Press.

Jeffrey, R. C. (1992). *Probability and the art of judgment*. Cambridge, England: Cambridge University Press.

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, *50*(1), 109–135.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences*, *5*(10), 434–442.

Johnson-Laird, P. N. (2006). *How we reason*. New York, NY: Oxford University Press.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*(43), 18243–18250.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, England: Erlbaum.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*(3), 640–661.

Johnson-Laird, P. N., & Khemlani, S. (2013). Toward a unified theory of reasoning. *Psychology of Learning and Motivation*, *59*, 1–42.

Johnson-Laird, P. N., Khemlani, S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*(4), 201–214.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition*, *193*, 103950.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–231.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, *142*, 191–204.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber & C.

Schmidt-Petri (Eds.), *Degrees of belief* (pp. 263–297). Dordrecht, Netherlands: Springer.

Jung, N., Wranke, C., Hamburger, K., & Knauff, M. (2014). How emotions affect logical reasoning: Evidence from experiments with mood-manipulated participants, spider phobics, and people with exam anxiety. *Frontiers in Psychology*, *5*, 1–12.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kant, I. (1908). *Kritik der praktischen Vernunft* [Critique of practical reason]. In *Kant's gesammelte Schriften* (Vol. 5, pp. 1–163). Berlin, Germany: Reimer. (Original work published 1788)

Kemmerling, A. (2017). *Glauben: Essay über einen Begriff* [Belief: An essay on a concept]. Frankfurt/Main, Germany: Klostermann.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review*, *103*(4), 687–719.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*(1), 623–655.

Khemlani, S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*(11), 2276–2288.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*, 4–20.

Kirk, U., Downar, J., & Montague, P. R. (2011). Interoception drives increased rational decision-making in meditators playing the ultimatum game. *Frontiers in Neuroscience*, *5*, 49. doi:10.3389/fnins.2011.00049

Kirkham, R. L. (1992). *Theories of truth*. Cambridge, MA: MIT Press.

Klauer, K. C. (1997). Working memory involvement in propositional and spatial reasoning. *Thinking & Reasoning*, *3*(1), 9–47.

Kleene, S. C. (1967). *Mathematical logic*. New York, NY: Wiley.

Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. *Spatial Cognition and Computation*, *1*(3), 261–290.

Knauff, M. (2009). Reasoning. In M. D. Binder, N. Hirokawa, & U. Windhorst (Eds.), *Encyclopedia of neuroscience* (pp. 3377–3382). Berlin, Germany: Springer.

Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. Cambridge, MA: MIT Press.

Knauff, M., & Gazzo Castañeda, L. E. (2021). When nomenclature matters: Is the "new paradigm" really a new paradigm for the psychology of reasoning? *Thinking & Reasoning*.

Knauff, M., Fangmeier, T., Ruff, C. C., & Johnson-Laird, P. N. (2003). Reasoning, models, and images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, *15*(4), 559–573.

Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition*, *30*(3), 363–371.

Knauff, M., Mulack, T., Kassubek, J., Salih, H. R., & Greenlee, M. W. (2002). Spatial imagery in deductive reasoning: A functional MRI study. *Cognitive Brain Research*, *13*(2), 203–212.

Knauff, M., Strube, G., Jola, C., Rauh, R., & Schlieder, C. (2004). The psychological validity of qualitative spatial reasoning in one dimension. *Spatial Cognition & Computation*, *4*(2), 167–188.

Knauff, M., & Wolf, A. G. (2010). Complex cognition: The science of human reasoning, problem-solving, and decision-making. *Cognitive Processing*, *11*(2), 99–102.

Koch, C. (2016). Does brain size matter? *Scientific American*, *27*(1), 22–25.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, *302*(5648), 1181–1185.

Konek, J., & Levinstein, B. A. (2019). The foundations of epistemic decision theory. *Mind*, *128*, 69–107.

Koons, R. (2017). Defeasible reasoning. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/win2017/entries/reasoning-defeasible/

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609–639.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 945–959.

Koriat, A. (2018). When reality is out of focus: Can people tell whether their beliefs and judgments are correct or wrong? *Journal of Experimental Psychology: General*, *147*(5), 613–631.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*(4), 478–492.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36–69.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.

Kosslyn, S. M. (2005). Mental images and the brain. *Cognitive Neuropsychology*, *22*(3–4), 333–347.

Kripke, S. A. (2013). The Church–Turing "Thesis" as a special corollary of Gödel's completeness theorem. In B. J. Copeland,

C. Posy, & O. Shagrir (Eds.), *Computability: Turing, Gödel, Church, and beyond* (pp. 77–104). Cambridge, MA: MIT Press.

Krüger, L., Daston, L. J., & Heidelberger, M. (Eds.). (1987). *The probabilistic revolution: Vol. 1. Ideas in history*. Cambridge, MA: MIT Press.

Krüger, L., Gigerenzer, G., & Morgan, M. S. (Eds.). (1987). *The probabilistic revolution: Vol. 2. Ideas in the sciences*. Cambridge, MA: MIT Press.

Krumnack, A., Bucher, L., Nejasmic, J., Nebel, B., & Knauff, M. (2011). A model for relational reasoning as verbal reasoning. *Cognitive Systems Research*, *12*(3–4), 377–392.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Künne, W. (2003). *Conceptions of truth*. Oxford, England: Oxford University Press.

Kusser, A., & Spohn, W. (1992). The utility of pleasure is a pain for decision theory. *Journal of Philosophy*, *89*, 10–29.

Lakatos, I. (1978). *The methodology of scientific research programmes* (Philosophical Papers, Vol. I; J. Worrall & G. Currie, Eds.). Cambridge, England: Cambridge University Press.

Laughlin, P. R., & Adamopoulos, J. (1980). Social combination processes and individual learning for six-person cooperative groups on an intellective task. *Journal of Personality and Social Psychology*, *38*(6), 941–947.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189.

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

Levi, I. (1967). *Gambling with truth: An essay on induction and the aims of science*. New York, NY: Knopf.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Levinson, S. C., Kita, S., Haun, D. B. M., & Rasch, B. H. (2002). Returning the tables: Language affects spatial reasoning. *Cognition*, *84*(2), 155–188.

Lewis, C. I. (1912). Implication and the algebra of logic. *Mind*, *21*, 522–531.

Lewis, D. (1969). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. II, pp. 263–293). Berkeley: University of California Press.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition*, *83*(3), 265–294.

Liang, D. W., Moreland, R., & Argote, L. (1995). Group versus individual training and group performance: The mediating factor of transactive memory. *Personality and Social Psychology Bulletin*, *21*(4), 384–393.

Lipton, P. (1991). *Inference to the best explanation*. London, England: Routledge.

List, C. (2013). Social choice theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/win2013/entries/social-choice/

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford, England: Oxford University Press.

Lloyd, G. E. R. (2017). *The ambivalences of rationality: Ancient and modern cross-cultural explorations*. Cambridge, MA: Cambridge University Press.

Loewenstein, G., & Elster, J. (Eds.). (1992). *Choice over time*. New York, NY: Russell Sage Foundation.

Lorenz, H. (2009). Ancient theories of soul. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2009/entries/ancient-soul/

Luders, E., Narr, K. L., Thompson, P. M., & Toga, A. W. (2009). Neuroanatomical correlates of intelligence. *Intelligence*, *37*(2), 156–163.

Lyon, A., & Pacuit, E. (2013). The wisdom of crowds: Methods of human judgement aggregation. In P. Michelucci (Ed.), *Handbook of human computation* (pp. 599–614). New York, NY: Springer.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835–1838.

Manktelow, K. (2012). *Thinking and reasoning: An introduction to the psychology of reason, judgment and decision making*. Hove, England: Psychology Press.

Marchal, K., & Wenzel, C. H. (2017). Chinese perspectives on free will. In K. Timpe, M. Griffith, & N. Levy (Eds.), *The Routledge companion to free will* (pp. 374–388). New York, NY: Routledge.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.

Maschler, M., Solan, E., & Zamir, S. (2013). *Game theory*. Cambridge, England: Cambridge University Press.

McAdams, R. H. (2009). Beyond the prisoners' dilemma: Coordination, game theory, and law. *Southern California Law Review*, *82*(209), 209–258.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, *107*(1), 343–352.

McClennen, E. F. (1990). *Rationality and dynamic choice*. Cambridge, England: Cambridge University Press.

Menary, K., Collins, P. F., Porter, J. N., Muetzel, R., Olson, E. A., Kumar, V., . . . Luciana, M. (2013). Associations between cortical thickness and general intelligence in children, adolescents and young adults. *Intelligence*, *41*(5), 597–606.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, MA: Harvard University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Mojzisch, A., & Schulz-Hardt, S. (2006). Information sampling in group decision making: Sampling biases and their consequences. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 299–326). Cambridge, England: Cambridge University Press.

Montague, R. (1974). *Formal philosophy: Selected papers of Richard Montague* (R. H. Thomason, Ed.). New Haven, CT: Yale University Press.

Monti, M. M., Osherson, D. N., Martinez, M. J., & Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: A language-independent distributed network. *Neuroimage*, *37*(3), 1005–1016.

Moore, G. E. (1903). *Principia ethica*. Cambridge, England: Cambridge University Press.

Moorhead, G., Ference, R., & Neck, C. P. (1991). Group decision fiascoes continue: Space shuttle Challenger and a revised groupthink framework. *Human Relations*, *44*(6), 539–550.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, *4*(3), 231–248.

Mumford, S. (1998). *Dispositions*. Oxford, England: Oxford University Press.

Munro, G. D., & Munro, C. A. (2014). "Soft" versus "hard" psychological science: Biased evaluations of scientific evidence that threatens or supports a strongly held political identity. *Basic and Applied Social Psychology*, *36*(6), 533–543.

Myerson, R. B. (1991). *Game theory: Analysis of conflict*. Cambridge, MA: Harvard University Press.

Neubauer, A. C., & Fink, A. (2009). Intelligence and neural efficiency. *Neuroscience & Biobehavioral Reviews*, *33*(7), 1004–1023.

Newell, A. (1973). Production systems: Models of control structures. In W. G. Chase (Ed.), *Visual information processing* (pp. 463–526). New York, NY: Academic Press.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*, 87–127.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Oxford, England: Prentice-Hall.

Nozick, R. (1993). *The nature of rationality*. Princeton, NJ: Princeton University Press.

Oaksford, M. (2015). Imaging deductive reasoning and the new paradigm. *Frontiers in Human Neuroscience*, *9*, 101. doi:10.3389/fnhum.2015.00101

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, England: Oxford University Press.

Oaksford, M., & Chater, N. (2012). Dual processes, probabilities, and cognitive architecture. *Mind & Society*, *11*(1), 15–26.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 305–330.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Osgood, C. E. (1962). *An alternative to war or surrender*. Urbana: University of Illinois Press.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11*(6), 988–1010.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997. *American Political Science Review*, *92*(1), 1–22.

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, *100*(3), 641–672.

Otworowska, M., Blokpoel, M., Sweers, M., Wareham, T., & van Rooij, I. (2018). Demons of ecological rationality. *Cognitive Science*, *42*(3), 1057–1066.

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

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, *52*, 1029–1050.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Peebles, D., & Cooper, R. P. (2015). Thirty years after Marr's *Vision*: Levels of analysis in cognitive science. *Topics in Cognitive Science*, *7*(2), 187–190.

Peirce, C. S. (1877). The fixation of belief. *Popular Science Monthly*, *12*, 1–15.

Perner, J., & Wimmer, H. (1985). "John *thinks* that Mary *thinks* that . . .": Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, *39*(3), 437–471.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford, England: Oxford University Press.

Pietschnig, J., Penke, L., Wicherts, J. M., Zeiler, M., & Voracek, M. (2015). Meta-analysis of associations between human brain volume and intelligence differences: How strong are they and what do they mean? *Neuroscience & Biobehavioral Reviews*, *57*, 411–432.

Pinker, S. (1994). *The language instinct*. London, England: Allen Lane.

Plato. (1921). *Sophist* (H. N. Fowler, Trans.). In *Plato in 12 volumes* (Vol. VII, Loeb Classical Library). Cambridge, MA: Harvard University Press.

Plato. (2013). *Republic* (Vols. I–II, C. Emlyn-Jones & W. Preddy, Eds. & Trans.). Cambridge, MA: Harvard University Press.

Politzer, G., & Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking & Reasoning*, *7*(3), 217–234.

Popper, K. R. (1945). *The open society and its enemies* (Vols. 1–2). London, England: Routledge & Kegan Paul.

Popper, K. R. (1989). *Logik der Forschung* [The logic of scientific discovery] (9th ed.). Tübingen, Germany: Mohr. (Original work published 1934)

Prado, J., Spotorno, N., Koun, E., Hewitt, E., Van der Henst, J.-B., Sperber, D., & Noveck, I. A. (2015). Neural interaction between logical reasoning and pragmatic processing in narrative discourse. *Journal of Cognitive Neuroscience*, *27*(4), 692–704.

Prowse Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking & Reasoning*, *15*(1), 69–100.

Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind: A symposium* (pp. 138–164). New York, NY: Collier Books.

Putnam, H. (1975). The nature of mental states. In *Mind, language and reality* (Philosophical Papers, Vol. 2, pp. 429–440). Cambridge, England: Cambridge University Press. (Original work published 1967)

Putnam, H. (1981). *Reason, truth and history* (Philosophical Papers, Vol. 1). Cambridge, England: Cambridge University Press.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Quine, W. V. O. (1969). Epistemology naturalized. In *Ontological relativity and other essays* (pp. 69–90). New York, NY: Columbia University Press.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*(3), 389–404.

Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., & Tse, A. P. P. (2017). Formal nonmonotonic theories and properties of human defeasible reasoning. *Minds and Machines*, *27*(1), 79–117.

Ragni, M., Franzmeier, I., Maier, S., & Knauff, M. (2016). Uncertain relational reasoning in the parietal cortex. *Brain and Cognition*, *104*, 72–81.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choice under uncertainty*. New York, NY: Random House.

Ramsey, F. P. (1978). *Foundations: Essays in philosophy, logic, mathematics and economics* (D. H. Mellor, Ed.). London, England: Routledge & Kegan Paul.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Belknap Press.

Reason, J. T. (1987). The Chernobyl errors. *Bulletin of the British Psychological Society*, *40*, A46.

Recanati, F. (2016). *Mental files in flux*. Oxford, England: Oxford University Press.

Reverberi, C., Shallice, T., D'Agostini, S., Skrap, M., & Bonatti, L. L. (2009). Cortical bases of elementary deductive reasoning: Inference, memory, and metadeduction. *Neuropsychologia*, *47*(4), 1107–1116.

Revlin, R., Cate, C. L., & Rouss, T. S. (2001). Reasoning counterfactually: Combining and rending. *Memory & Cognition*, *29*(8), 1196–1208.

Richardson, H. S. (1994). *Practical reasoning about final ends*. Cambridge, England: Cambridge University Press.

Richardson, R. C. (2007). *Evolutionary psychology as maladapted psychology*. Cambridge, MA: MIT Press.

Ridge, M. (2019). Moral non-naturalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2019/entries/moral-non-naturalism/

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.

Ritter, J., Gründer, K., & Gabriel, G. (Eds.). (1971–2007). *Historisches Wörterbuch der Philosophie* [Historical dictionary of philosophy] (Vols. 1–13). Basel, Switzerland: Schwabe.

Rousseau, J.-J. (1762). *Du contrat social; ou Principes du droit politique* [On the social contract; or, Principles of political rights]. Amsterdam, Netherlands: Rey.

Rubinstein, A. (1998). *Modeling bounded rationality*. Cambridge, MA: MIT Press.

Ruff, C. C., Knauff, M., Fangmeier, T., & Spreer, J. (2003). Reasoning and working memory: Common and distinct neuronal processes. *Neuropsychologia*, *41*(9), 1241–1253.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vols. 1–2). Cambridge, MA: MIT Press.

Ryle, G. (1949). *The concept of mind*. London, England: Hutchinson.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*, 61–71.

Sanfey, A. G., Loewenstein, G., McClure, S. M., & Cohen, J. D. (2006). Neuroeconomics: Cross-currents in research on decision-making. *Trends in Cognitive Sciences*, *10*(3), 108–116.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626), 1755–1758.

Savage, L. J. (1954). *The foundations of statistics*. New York, NY: Wiley.

Sayre-McCord, G. (Ed.). (1988). *Essays on moral realism*. Ithaca, NY: Cornell University Press.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399.

Schopenhauer, A. (1841). *Die beiden Grundprobleme der Ethik* [The two fundamental problems of ethics]. Frankfurt/Main, Germany: Hermannsche Buchhandlung.

Schroeder, M. (2018). The moral truth. In M. Glanzberg (Ed.), *The Oxford handbook of truth* (pp. 579–601). Oxford, England: Oxford University Press.

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes*, *118*(1), 24–36.

Schurz, G. (1997). *The is–ought problem: A study in philosophical logic*. Dordrecht, Netherlands: Kluwer.

Schurz, G., & Hertwig, R. (2019). Cognitive success: A consequentialist account of rationality in cognition. *Topics in Cognitive Science*, *11*, 1–30.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge, England: Cambridge University Press.

Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Sedgewick, R., & Flajolet, P. (2013). *An introduction to the analysis of algorithms* (2nd ed.). Boston, MA: Addison-Wesley.

Sen, A. K. (1970). *Collective choice and social welfare*. San Francisco, CA: Holden-Day.

Shafer, G. (1978). Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, *19*, 309–370.

Shoenfield, J. R. (1967). *Mathematical logic*. Reading, MA: Addison-Wesley.

Simon, H. A. (1947). *Administrative behavior*. New York, NY: Macmillan.

Simon, H. A. (1957). *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*. New York, NY: Wiley.

Simon, H. A. (1959). Theories of decision making in economics and behavioral science. *American Economic Review*, *49*(3), 253–283.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge, England: Cambridge University Press.

Sober, E. (2015). *Ockham's razor: A user's manual*. Cambridge, England: Cambridge University Press.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford, England: Blackwell.

Spohn, W. (1982). How to make sense of game theory. In W. Stegmüller, W. Balzer, & W. Spohn (Eds.), *Philosophy of economics* (pp. 239–270). Berlin, Germany: Springer.

Spohn, W. (1993). Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein? [How can the theory of rationality be normative and empirical at the same time?] In L. Eckensberger & U. Gähde (Eds.), *Ethik und Empirie: Zum Zusammenspiel von begrifflicher Analyse und erfahrungswissenschaftlicher Forschung in der Ethik* (pp. 151–196). Frankfurt/Main, Germany: Suhrkamp.

Spohn, W. (2002). The many facets of the theory of rationality. *Croatian Journal of Philosophy*, *2*, 247–262.

Spohn, W. (2009). *Causation, coherence, and concepts: A collection of essays*. Dordrecht, Netherlands: Springer.

Spohn, W. (2011). Normativity is the key to the difference between the human and the natural sciences. In D. Dieks, W. J.

Gonzalez, S. Hartmann, T. Uebel, & M. Weber (Eds.), *Explanation, prediction, and confirmation* (pp. 241–251). Dordrecht, Netherlands: Springer.

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

Spohn, W. (2016). Truth and rationality. *Tomsk State University Journal of Philosophy, Sociology, and Political Science*, *36*, 7–19.

Spohn, W. (2017). Knightian uncertainty meets ranking theory. *Homo Oeconomicus*, *34*, 293–311.

Spohn, W. (2020). Defeasible normative reasoning. *Synthese*, *197*, 1391–1428.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford, England: Blackwell.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–695.

Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry*, *14*(3–4), 304–313.

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

Stigler, S. M. (1983). Who discovered Bayes's theorem? *The American Statistician*, *37*, 290–296.

Störring, G. (1908). Experimentelle Untersuchungen über einfache Schlussprozesse [Experimental studies on basic inference processes]. *Archiv für die gesamte Psychologie*, *11*, 1–127.

Stuss, D. T. (2011). Functions of the frontal lobes: Relation to executive functions. *Journal of the International Neuropsychological Society*, *17*(5), 759–765.

Stuss, D. T., & Levine, B. (2002). Adult clinical neuropsychology: Lessons from studies of the frontal lobes. *Annual Review of Psychology*, *53*, 401–433.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York, NY: Doubleday.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*(3), 261–288.

Tafer, Z., Boussahmine, A., & Bouanini, S. (2016). Behavioral economic, rationality and Islamic ethics. *Journal of Business and Economics*, *7*(5), 871–888.

Theiner, G., Allen, C., & Goldstone, R. L. (2010). Recognizing group cognition. *Cognitive Systems Research*, *11*(4), 378–395.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140.

Thompson, V. A., Therriault, N. H., & Newman, I. R. (2016). Meta-reasoning: Monitoring and control of reasoning, decision making, and problem solving. In L. Macchi, M. Bagassi, & R. Viale (Eds.), *Cognitive unconscious and human rationality* (pp. 275–299). Cambridge, MA: MIT Press.

Tomasello, M. (2009a). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2009b). *Why we cooperate*. Cambridge, MA: MIT Press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*(1), 35–57.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York, NY: Academic Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In C. Unkelbach & R. Greifeneder (Eds.), *The experience of thinking: How the fluency of mental processes influences cognition and behavior* (pp. 11–32). Hove, England: Psychology Press.

Vandierendonck, A., & De Vooght, G. (1997). Working memory constraints on linear reasoning with spatial and temporal contents. *Quarterly Journal of Experimental Psychology: Section A*, *50*(4), 803–820.

van Fraassen, B. C. (1984). Belief and the will. *Journal of Philosophy*, *81*, 235–256.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

von Wright, G. H. (1971). *Explanation and understanding*. Ithaca, NY: Cornell University Press.

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). New York, NY: Oxford University Press.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology*. Harmondsworth, England: Penguin.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Weber, M. (1973). Der Sinn der „Wertfreiheit“ der soziologischen und ökonomischen Wissenschaften [The meaning of “value freedom” in the sociological and economic sciences]. In *Gesammelte Aufsätze zur Wissenschaftslehre* (4th ed., pp. 146–214). Tübingen, Germany: Mohr. (Original work published 1917)

Weber, M. (1921–1922). *Wirtschaft und Gesellschaft* [Economy and society] (Vols. I–II). Tübingen, Germany: Mohr (Siebeck).

Weibull, J. W. (1997). *Evolutionary game theory* (Vol. 1). Cambridge, MA: MIT Press.

Weisberg, J. (2020). Belief in psyontology. *Philosophers' Imprint*, *20*(11), 1–27.

Wilkins, M. C. (1928). *The effect of changed material on ability to do formal syllogistic reasoning* (Archives of Psychology, No. 102). New York, NY: Woodworth.

Williamson, T. (2000). *Knowledge and its limits*. Oxford, England: Oxford University Press.

Witte, E. H., & Davis, J. H. (Eds.). (1996). *Understanding group behavior* (Vols. 1–2). Mahwah, NJ: Erlbaum.

Wolf, A. G., Rieger, S., & Knauff, M. (2012). The effects of source trustworthiness and inference type on human belief revision. *Thinking & Reasoning*, *18*(4), 417–440.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451–460.

Wundt, W. M. (1910). Psychologismus und Logizismus [Psychologism and logicism]. In *Kleine Schriften* (Vol. 1, pp. 511–634). Leipzig, Germany: Engelmann.

Wundt, W. M. (2010). *Grundriss der Psychologie* [Outlines of psychology]. Leipzig, Germany: Engelmann. (Original work published 1896)

Zalta, E. N. (1988). *Intensional logic and the metaphysics of intentionality*. Cambridge, MA: MIT Press.

Zhao, J., Crupi, V., Tentori, K., Fitelson, B., & Osherson, D. (2012). Updating: Learning versus supposing. *Cognition*, *124*, 373–378.

Zhu, D. H. (2013). Group polarization on corporate boards: Theory and evidence on board decisions about acquisition premiums. *Strategic Management Journal*, *34*(7), 800–822.

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

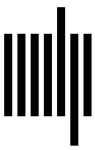
# The Handbook of Rationality

## 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

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