CONTENTS

1

Introduction

What does it mean to be rational? In the middle of the twentieth century, it looked like there was a definitive answer to this question. Mathematicians, economists, and statisticians all agreed that rational agents choose actions that maximize their **expected utility**, weighing the utility of each outcome by its probability (von Neumann and Morgenstern, 1944; Savage, 1954). But when psychologists looked carefully at how people choose actions, it became clear that this is an imperfect description of human behavior: people systematically violate the axioms of rationality and act in ways that are inconsistent with maximizing expected utility (Tversky and Kahneman, 1974; Kahneman et al., 1982).

One response to these findings is simply to decide that people are irrational—that the mathematicians, economists and statisticians had characterized rationality correctly and humans are faulty actors. However, there is another possible response, which is to ask whether we are holding humans to the right standard. Perhaps the fault lies not in human behavior, but in the definition of rational action. One clue that this might be the case is that, despite not following the prescriptions of classical rationality, human beings are the best example we have of a general-purpose intelligent system. If we want to understand rationality not in abstract, but in the context of what produces intelligent behavior, there might be lessons we can learn by starting from the choices that humans make.

The perspective of trying to understand how an intelligent agent should behave led to an alternative notion of rationality, one that came not from mathematics, economics, and statistics, but from computer science. One of the things that is missing from the principle of maximizing expected utility is any notion of how difficult it might be to do so. To make a classically rational decision, we would have to consider all the outcomes that might arise from each action we take and assign each outcome a utility and a probability. In any realistic setting, the number of possible outcomes will likely be very large, and assigning a precise utility and probability to each outcome could be very

difficult. In particular, it could be extremely **computationally costly** to run an algorithm that does this—the algorithm might take so long to execute that the original decision becomes irrelevant.

Computer scientists are used to trading off the quality of an output with the cost of computing it. Formalizing this trade-off is crucial for designing artificial intelligence (AI) systems that need to take informed actions in the real world. This is especially true for designing autonomous robots that interact with unstructured real-world environments in real time and compute all of their decisions on onboard hardware, such as autonomous vehicles. By considering how such AI systems should act, computer scientists developed the theory of **bounded optimality**—an alternative definition of rationality that focuses not on whether an agent selects the best *action*, but on whether the agent follows the best *algorithm* for making that decision taking into account both utility and computational cost (Horvitz, 1987; Russell and Wefald, 1991a; Russell and Subramanian, 1995; Russell, 1997). The bounded-optimal agent settles for a good enough action when time is limited. This poses a new kind of decision problem: deciding what internal computations to execute in order to select an external action (Russell and Wefald, 1991b).

We (the authors, Falk, Fred, and Tom) are interested in considering alternative ways of defining rationality not because we want to know whether or not humans are rational, but because knowing how people *should* be solving a problem is often a key part of understanding how they *do* solve that problem and how they could solve it even better. In cognitive science, **rational analysis** is a method for understanding human behavior in which we consider the optimal solution to the problems that human minds face (Anderson, 1990). Traditionally, this kind of analysis was done within the framework of classical rationality. However, this omits an important part of actually understanding human behavior: understanding the constraints under which human minds operate. To address this omission, we and our colleagues developed the approach of **resource-rational analysis**, which uses bounded optimality and associated ideas to ask how human minds *should make the best use of their limited cognitive resources* to solve the problems they face (Lieder et al., 2012; Griffiths et al., 2015; Lieder and Griffiths, 2020b).

In this book, we present a formal framework for applying resource-rational analysis to understand and improve human behavior, a set of tools we have developed to make this easier, and examples of how we have used this approach to revisit classic questions about human cognition, pose new ones, and enhance human rationality. Our target audience is not limited to the people who have traditionally been concerned about rationality: psychologists studying judgment and decision-making, economists, and philosophers. We

also address the broader group of psychologists, cognitive scientists, and neuroscientists trying to understand human minds and brains, as well as computer scientists interested in reproducing such systems in machines. We see these ideas as being relevant to this broader audience because understanding the rational use of cognitive resources offers not just a new way to think about the evidence for human irrationality in decision-making, but also a new perspective on many long-standing questions in cognitive science. Many theories in the social sciences and humanities are (implicitly or explicitly) built on assumptions about human rationality. This book provides constructive suggestions for how those assumptions can be improved. Moreover, the new approach to improving people's decisions that we present is relevant for anyone who wants certain decisions to be made well, including researchers, educators, and practitioners who care about improving people's lives or the functioning of groups, organizations, or society.

To understand the potential we see in the resource-rational approach, it is worth observing that the fundamental limitation of the classical notion of rationality is not just that it does not engage with how difficult it might be to make a decision, but that it does not engage with cognitive processes at all. Maximizing expected utility is fundamentally a **behaviorist** (Skinner, 1953) notion of rationality: it defines rational action purely in terms of the agent's response to its environment. By contrast, considering the question of how an agent should best deploy its cognitive resources to solve a problem places resource-rational analysis in firm contact with those cognitive processes. For the first time, we can say not just what action an agent should take but also how it should think about what to do. This is a powerful idea that extends across the psychologist's entire toolbox of cognitive processes—attention, planning, and memory all involve the use of limited resources, and resource-rational analysis gives us a new way to understand them.

Even more significantly, resource-rational analysis provides a new way to think about how to support and improve human cognition. When we use classical rationality to assess people's performance, we risk assuming that performance can be improved by teaching people the classically rational solutions to the problems they face. However, because these solutions are computationally costly, they may be impractical for many real-world decisions. Resource-rational analysis suggests an alternative approach: teaching people practical cognitive strategies that make better use of their limited cognitive resources. Thinking about computational costs also highlights another way that we can support human decision-making: building more computation into human environments. Using this approach, we can design **cognitive prostheses** to support longer-term decision-making (Lieder et al., 2019) and understand the

circumstances under which behavioral **nudges** are most likely to be effective (Callaway et al., 2023).

In addition to a framework, tools, and examples, we hope to share with you our excitement about this approach and our vision for the many ways it can be used to understand and improve human minds. The cognitive revolution provided an alternative to behaviorism by showing how mathematics can be used to express rigorous hypotheses about cognitive processes (Miller, 2003). In the same way, we see resource-rational analysis as offering a path to a cognitive revolution in how we use and think about rationality when developing theories to explain, predict, and improve human behavior. The result is a blueprint for a new bridge between ideal solutions and cognitive processes that is directly informed by empirical data and supports interventions that meaningfully improve human lives. We hope this book will help you join us in building that bridge.

1.1 A simple example

Before we dive into any more details, we will give a simple example that illustrates what a resource-rational analysis looks like, how it differs from classical notions of rationality, and how this difference gives us new tools for understanding human behavior. This example takes the simplest kind of decision an agent could make, considers how it becomes more complicated when resources are limited, and shows a surprising connection to something people do that has long been considered irrational. It is also the first resource-rational analysis that any of us worked on (Vul et al., 2009).

Imagine you have a choice between two options. For simplicity, we will call them Option A and Option B. If you select Option A, you receive a reward with probability p. If you select Option B, you receive the same reward with a probability of $1 - p$. Which option should you choose?

From the perspective of classical rationality, this problem is trivial. If p is greater than 0.5, then you should choose Option A, otherwise choose Option B. But it becomes much more challenging if you don't know the value of p. In many real human decisions, assessing the probability of different outcomes is the hardest part. Doing so may require imagining future possibilities, considering different sources of evidence, recalling similar situations from memory, or engaging in various other cognitive processes.

To simplify things, we are going to model all of those different cognitive processes as methods of **sampling** a possible outcome. Whether we imagine, consider, or recall, we are getting a piece of information about the value of p. We will assume that obtaining this piece of information is equivalent to

sampling a possible outcome for this decision—getting a result that favors Option A with probability p or Option B with probability $1 - p$.

Assuming these cognitive processes are costly—they take time and mental effort—making rational use of our cognitive resources in this setting is a matter of deciding how many samples to generate before choosing one of the options. Making this decision requires considering the cost of sampling relative to the cost of making an error. Finding the optimal trade-off between these two costs can be formulated as an optimization problem.

The solution to this optimization problem might be surprising: for a wide range of relative costs, it is optimal to take a single sample. Intuitively, there are few values of p where more samples are informative. If p is close to 1 or 0, then a single sample indicates the right answer with high probability. If p is close to 0.5, then more samples yield little benefit over choosing at random. If the cost of making an error is large relative to the cost of sampling, then the few cases where p falls in an intermediate range may be enough for it to make sense to draw more samples. But in many situations, one is enough.

This resource-rational analysis explains an aspect of human behavior that, on its surface, appears irrational. Given choices like this, where options are rewarded probabilistically, people tend to select those options with the probability that matches the probability of reward. This **probability matching** behavior is inconsistent with the classically rational solution, which is to always choose the option that is best according to all available information (Vulkan, 2000). However, probability matching makes sense if people are using a cognitive mechanism that behaves like sampling and have to pay a cost for each sample they generate.

While this is a simple example—perhaps the simplest we can imagine—it illustrates some themes that will reappear many times in this book. First, while the classically rational solution is easy to define abstractly, computing a concrete rational action is much more cumbersome and often impractical due to limited cognitive resources. Second, by taking people's limited cognitive resources into account, we can understand aspects of human behavior that differ systematically from the prescriptions of classical rationality. Finally, we can reach that understanding while abstracting away from the specific cognitive mechanisms involved. Throughout the book, we are going to define models at a fairly high level of generality, using processes such as sampling that are intended to stand in as abstract proxies for a variety of different cognitive mechanisms.

While we have not specifically explored ways of improving human decision-making in the simple kinds of decisions we are considering in this example, our resource-rational analysis of such decisions suggests some ways

we could improve them. If we can identify situations where the cost of error is high relative to the cost of sampling, these would be good situations in which to signal that people might want to think carefully before responding. Likewise, if we have information about what range the value of p is likely to fall into, we might be able to tell people whether generating more than one sample is worthwhile. This is a way to support people's decisions by putting some of the computation they might otherwise have to perform into their decision environment.

In the rest of the book, we explore how making the move that we made here—going from a task to a decision about how to deploy our cognitive resources to best solve that task—can be used to shed light not just on decision-making, but also on planning, attention, and memory. The optimization problems involved in allocating these resources become increasingly complex, and we will present a framework for formulating and solving those problems. We will also show how the resulting models connect to traditional questions in cognitive psychology, and how they can be used to design effective interventions for agents with bounded cognitive resources.

1.2 Answers to skeptical questions

The work presented in this book is the culmination of a decade-long research program exploring applications of resource-rational analysis in cognitive science. Consequently, we have had the opportunity to present these ideas to various audiences and, hence, to field various skeptical questions from those audiences. To help you decide whether reading further is worth the cognitive resources involved, this section provides our answers to those skeptical questions. Of course, if you are already convinced, you can skip this section.

1.2.1 Why should I care?

The answer to this question depends on the person asking it. Depending on your interests, you may find different aspects of resource-rational analysis appealing.

If you are a **psychologist**, resource-rational analysis provides a different way to think about cognitive processes. Many of the phenomena that interest psychologists, such as problem-solving and deductive reasoning, are trivial from the perspective of classical rationality: the goal is simply to solve the problem. In these cases, the journey—the sequence of cognitive processes we engage in when solving the problem—is more important than the destination. For a long time, psychologists have developed theories about how people perform such tasks by manually assembling different cognitive

mechanisms, strategies, or heuristics guided by empirical results. If you use resource-rational analysis, you can have computers do this tedious work for you: having defined a task and a set of elementary cognitive processes that could be involved, the optimal way to combine these processes can be derived automatically. This provides a way to answer questions about how people should think and what they should be thinking about, in the same way that classical rational analysis answers the question of how people should act. Consequently, resource-rational analysis makes deeper contact with the traditional preoccupations of cognitive psychology than rational analysis. It also provides a different way to design interventions to support human decision-making, focusing on guiding people toward better strategies and offloading some of the computation involved into the environment in which decisions are made.

If you are an **economist**, you are likely to be familiar with the idea of modeling human behavior as the solution to an optimization problem, but possibly also concerned that people systematically deviate from classical rationality. Resource-rational analysis is a way to create models of human behavior that are just as general—and use many of the same tools—as classical rational models, but also capture some systematic deviations from classical economic rationality. As we will discuss in the next chapter, this approach is aligned with recent work in economics, particularly in the **rational inattention** literature (Sims, 2003; Caplin and Dean, 2015), but also provides a way to more explicitly connect those ideas to cognitive processes, and thereby make the models' predictions even more accurate.

If you are a **neuroscientist**, having a theory of rational action that makes closer contact with cognitive mechanisms means that theory also makes closer contact with the brain. The successes of classical rational analysis led to a flurry of research on how the brain might implement probabilistic inference, typically by searching for one kind of computation or neural circuit that might solve this problem. The resource-rational perspective suggests that different instances of probabilistic inference might be solved in different ways, depending on the structure of the problem and the elementary cognitive processes available. For example, we will discuss a variety of sampling and planning algorithms that make sense to use under different circumstances. This provides a more nuanced set of targets for neuroscientists to search for in the brain. In addition, the core problem of selecting which cognitive processes to engage in has a structure that parallels that of **reinforcement learning**, creating the possibility that the extensive work on the neuroscience of reinforcement learning (Dolan and Dayan, 2013; O'Doherty et al., 2015; Lee et al., 2012; Mattar and Lengyel, 2022; Miller and Venditto, 2021) might also help us understand this metacognitive process. Moreover, resource-rational

analysis is a promising vehicle for translating neuroscientific insights into the basic computations neurons and neural circuits perform and how costly those computations are into computational models of cognition and their resulting applications.

If you are a **computer scientist**, resource-rational analysis provides an opportunity to understand how human minds are so efficient in their use of limited cognitive resources to solve a wide range of problems and to define a formal framework for emulating this in machines. Bounded optimality is a notion that originated in the AI literature, and there are still a lot of opportunities to improve on the computational methods that are used to solve the problem of discovering efficient algorithms. Research in machine learning has explored a variety of ad hoc methods for deciding how much computation to perform (Graves, 2016; Banino et al., 2021). Formulating these problems more explicitly in terms of **rational metareasoning** (Russell and Wefald, 1991a) may be a way to improve on these methods. Finally, a formal framework that captures some of the ways that people systematically deviate from classical rationality is a valuable tool for making sense of human behavior. If you are interested in building systems that interact with people, then better understanding the mapping from people's preferences to their behavior is essential for being able to infer people's preferences from their behavior (Ho and Griffiths, 2022).

If you are a **curious person** who is primarily interested in how to make better decisions in your own life, this book offers a different perspective on what it actually means to be rational. Many books about decision-making emphasize human irrationality, focusing on deviations from classical rationality. Our perspective is more nuanced: when considering the resource constraints that human minds operate under, many of the seemingly irrational things people do make more sense. It is reasonable to want to be more rational in your own decisions, but the path to getting there has to respect those constraints. We discuss improving human decisions in detail in Chapter 8, which might be of particular interest to you. You can find more pragmatic advice in Tom's book *Algorithms to Live By* (Christian and Griffiths, 2016), which is not explicitly framed in terms of resource-rational analysis but shares with this book the idea that computer science has an important contribution to make to understanding rational action.

1.2.2 Isn't this just bounded rationality?

Immediately after everyone started to agree on how to define rationality, Herbert Simon pointed out that it was an unreasonable standard for human behavior (Simon, 1955). His concern was one we share: that humans have limited time and cognitive resources, making maximization of expected utility

impractical. Simon wanted researchers to acknowledge the idea of **bounded rationality**—that the constraints under which human minds operate make classical rationality unachievable. This idea is very much a part of resource-rational analysis.

However, Simon viewed bounded rationality as an idea, rather than a formal framework. In a letter to the psychologist Gerd Gigerenzer, he wrote, "I have never thought of either bounded rationality or satisficing as precisely defined technical terms, but rather as signals to economists that they needed to pay attention to reality, and a suggestion of some ways in which they might. But I do agree that I have used bounded rationality as the generic term, to refer to all the limits that make a human being's problem spaces something quite different from the corresponding task environments: knowledge limits, computational limits, compatibility of component goals" (Augier and March, 2004, p. 406). This is where we differ: we view resource-rational analysis as a formal framework that can be useful for understanding the behavior of bounded agents, turning the idea of bounded rationality into a generalizable approach that can be used to derive precise predictions about cognition and behavior given a specification of the problem to be solved and the cognitive processes available for doing so.

In the same letter, Simon objected to the interpretation of bounded rationality in terms of constrained optimization—assuming that people solve an optimization problem similar to the one solved by a classically rational approach but with bounded resources providing constraints on the solution. In fact, he threatened to sue the next person who interpreted bounded rationality in this way. To anticipate any such litigation, we want to state that this is not how we think about resource-rational analysis. Rather, we see this approach as fundamentally changing the optimization problem implicit in the definition of rationality from one that is focused on external actions to one that is focused on internal computations. The simple example given above illustrates this: thinking about cognitive processes in terms of sampling leads to a completely new optimization problem, one that is focused on using those cognitive processes effectively.

It is notable that the one formal model that Simon published in his papers on bounded rationality has a very similar flavor to the models we consider in this book. Simon's model of **satisficing** (Simon, 1955) turns a choice problem into a stopping problem, where options are considered until one is found that is above a derived threshold. This kind of structure arises repeatedly in resource-rational models, where limited cognitive resources require us to consider our options (or gather information, or sample from distributions) sequentially. Many of the models we present later in the book focus on the problem of making intelligent decisions about when to stop thinking in a way that is very similar to satisficing.

1.2.3 How is this different from other perspectives on rationality?

Understanding what it means to be rational, and how the definition of rationality relates to what people actually do, has been a preoccupation of many researchers from many different disciplines. We provide a detailed treatment of these ideas in Chapter 2, but want to briefly sketch how we see the relationship of resource-rational analysis to previous approaches.

The **heuristics and biases** research program carried out by Daniel Kahneman and Amos Tversky in the 1970s provided clear evidence that people's judgments and decisions deviate from classical rationality (Tversky and Kahneman, 1974; Kahneman et al., 1982). This research program was originally framed in terms of the computational costs of rationality: given that probabilistic inference and calculating expected utility are intractable, people must find some way to approximate them. Heuristics are the natural solution, and biases are the clues that we can use to identify those heuristics. We view resource-rational analysis as being very aligned with this original goal: considering how the costs of using cognitive resources trade off with performance gives us a way to characterize what makes a heuristic worth using. As a consequence, we can rederive some of the heuristics that Kahneman and Tversky discovered (see Chapter 4) and derive new heuristics directly from the formulation of the problem people have to solve (see Chapter 6).

The notion of **ecological rationality** was offered by Gerd Gigerenzer and colleagues as an alternative way of understanding heuristics (Gigerenzer and Todd, 1999a; Gigerenzer, 2000). Focusing on the interaction between cognition and the environment, ecological rationality explores the idea that particular heuristics might be adaptive solutions in particular environments. In this way, using heuristics should not be considered a deviation from rationality, but rather a reasonable strategy in itself. More recently, this perspective has been argued for from the perspective of the **bias-variance trade-off**, which shows that simple strategies can be best when information is limited (Gigerenzer and Brighton, 2009). Again, we see this approach as being well aligned with resource-rational analysis, and some of the heuristics identified by this research program fall out of our framework (see Chapter 6). However, one important difference is that we consider computational costs as an important additional factor that pushes people toward simpler heuristics, beyond the effect of the bias-variance trade-off.

Work on **rational inattention** in economics has explored how classical rationality can be brought into closer alignment with human behavior by assuming that gathering information is costly (Sims, 2003; Caplin and Dean, 2015). The resulting models consider how much investment an agent

should make in gathering information—allocating attention to it—in order to find the best compromise between the costs it incurs and the benefits it provides for decision-making. This approach shares with the models we will discuss in this book the idea of introducing resource constraints into the formulation of the decisions that humans face. One difference is that the models we present here include cognitive processes—computations—as well as information gathering. However, we see the class of models developed under the flag of rational inattention as instances of resource-rational analysis.

Other researchers have used the term **computational rationality** to refer to applications of bounded optimality to cognitive science (Lewis et al., 2014; Gershman et al., 2015). We prefer to refer to these as resource-rational models (and have used this term since 2012; Lieder et al., 2012) as a way to emphasize the fact that they are about the rational use of cognitive resources and to avoid making a commitment to those resources being computational (something that allows this approach to make direct contact with rational inattention, for example).

1.2.4 Why are you proposing this now?

While the mathematical ideas behind bounded optimality were first proposed in the 1980s, it has been hard to use this approach to identify concrete cognitive strategies until the last decade or so. It has now become possible to actually define meaningful cognitive models based on bounded optimality because of innovations in the way we formulate and solve the underlying computational problems, and because of significant increases in the speed of the computers on which we solve those problems. Using the resulting models as the basis for designing effective behavioral interventions is potentially even more computationally challenging, which is one reason why it is a topic we have only been able to explore recently. One of the important things we do in this book is to provide you with the details on how we draw on ideas in artificial intelligence and statistics to set these problems up in a way that makes them (at least approximately) solvable.

1.2.5 So aren't you also creating a harder problem for human minds to solve?

Yes, we are defining a harder problem. Deciding how to think is certainly harder than deciding how to act. However, that harder problem does not necessarily need to be solved by human minds, and definitely does not need to be solved on the same timescale.

The problem of optimizing how we use our cognitive resources could be solved in different ways. One way would be simply considering all of the ways we could use our resources and choosing the best strategy. This is extremely computationally challenging—it involves presenting the harder version of the problem directly to human minds. But, importantly, there are other ways to solve this problem. One is to make use of the fact that we have repeated opportunities to perform particular tasks, and even new tasks often share components with tasks we have performed before. This means that we can *learn* better cognitive strategies over our lifetimes. Another is to recognize that variants of those same tasks were also faced by our ancestors, meaning that we can *evolve* better cognitive strategies across generations. In some cases, we do this via cultural evolution, coming up with intuitive strategies or algorithms we share with other people (e.g., Thompson et al., 2022). In other cases, biological evolution can act as the optimizing force—something that seems particularly plausible for making effective use of resources such as attention and memory that are shared with many other species (e.g., Tomlin et al., 2015).

In this way, although deciding how to think is harder than deciding how to act, humanity has had much more time to solve the problem. Individual human decisions are typically made on a timescale ranging from a few hundred milliseconds to a few weeks. Cognitive strategies are learned over a lifetime or tuned over many generations. And the way we use particular cognitive resources may pre-date all human minds.

We also expect that the mind solves the problem of deciding how to allocate its cognitive resources imperfectly. Resource-rational analysis is useful to the extent that people's cognitive strategies approximate the optimal solutions to these problems; it does not require that they be perfect. There is room for slippage, suboptimality, and satisficing in reasoning, learning, and evolution.

1.2.6 Doesn't this lead to an infinite regress?

Yes, we have introduced another level of abstraction into decision-making, and the same move can potentially be iterated. Thus, while our focus here is on deciding how to decide, we can imagine deciding how to decide how to decide, and so on. However, we do not see this kind of regress as a major problem. The key reason is that each level of abstraction has significantly diminishing returns. Deciding how to decide helps us explain a wide range of otherwise puzzling aspects of human behavior. Adding another level of abstraction—deciding how to decide how to decide—moves the focus from cognitive processes to cognitive architectures, which we explore in Chapter 7 (e.g., Milli et al., 2021). At this level, the question is how we should design a mind for it to be most effectively used by an agent that decides how to best

allocate cognitive resources. While we can still make progress in exploring this question, the answers are fairly abstract and harder to relate directly to human behavior. Adding further levels of abstraction gets us even further from behavior and begins to engage us with questions that start to push us beyond the limits of cognitive science (e.g., how should you design an evolutionary process to select a cognitive architecture that allows an agent that optimally allocates cognitive resources to solve problems?). While these questions are theoretically interesting, we see the biggest payoffs of these ideas lying at the lower levels of the potentially infinite regress.[1]

1.2.7 Are you really not interested in whether people are actually rational?

Yes, really. Or, to be more precise, we are not interested in testing the hypothesis that people are actually rational under different criteria for rationality. To us, what matters more is whether a particular definition of rationality is useful for understanding human behavior. We use resource rationality as a **methodological assumption** that allows us to make progress in the broader project of making sense of people's thoughts and actions (Godfrey-Smith, 2001). Classical rationality is arguably useful in this respect, even if people deviate from it in systematic ways. Most Bayesian models of cognition are implicitly formulated within this framework. These models help us understand the aspects of people's behavior that we can explain in terms of the optimal solutions to the problems that human minds face (Anderson, 1990; Griffiths et al., 2010). Resource-rational analysis allows us to go a step beyond that, introducing assumptions about constrained cognitive resources and using them to explain some of the systematic deviations from classical rationality.

Given the vast space of possible Bayesian models and cognitive mechanisms, we anticipate (and our critics have suggested; Jones and Love, 2011; Bowers and Davis, 2012) that it will be hard to falsify the assumption of either of these forms of rationality. For that reason, we do not think it is productive to engage in debates about whether or not people are rational. Rationality, of whatever denomination, provides a *framework* in which we can develop *models* of human behavior (Lakatos, 1970). Individual models are falsifiable—they generate specific predictions that we can test in further experiments.

1. It is worth noting that there are also precedents of not being too concerned about infinite regress in the context of understanding decision-making. The classic example is in game theory, which considers the asymptotic equilibria of processes where players deliberate about the consequences of their actions. Whether similar asymptotic results can be obtained in the case of resource-rational analysis is an interesting research question.

Establishing whether people's behavior is consistent with specific models allows us to evaluate whether the factors considered in those models—hypothesis spaces and prior distributions in Bayesian models, augmented by cognitive mechanisms and constraints in resource-rational models—are candidates for explaining different aspects of that behavior.

Ultimately, we take the success of specific models based on resource-rational analysis as evidence that the framework is *useful*, not as evidence that the assumption that people always act in a resource-rational way is *true*. The approach we present in this book will be successful if it helps us develop theories that help us predict and explain why people act in the way they do and design interventions that are effective in improving that behavior.

1.3 The rest of the book

If you have gotten this far, you are hopefully satisfied with our answers to your skeptical questions—or are at least reserving judgment until you get more details. In the same spirit of letting you make an informed decision about how to use your time, we will use the remainder of this chapter to briefly summarize the rest of the book. Our focus in the book is on presenting the theoretical ideas behind resource-rational analysis and illustrating these ideas through some of the applications and experimental results that demonstrate the value of this approach. However, this is still a new perspective and we anticipate that some of the strongest evidence for its utility is still to come. We aim to equip readers to be able to gather that evidence for themselves.

Chapter 2 provides a brief history of different perspectives on rationality in psychology, economics, and computer science, expanding on the description that appears above and providing more details on the background behind resource-rational analysis and the context in which it arose.

Chapter 3 introduces the key ideas behind the formal framework of resource-rational analysis, drawing on bounded optimality and rational metareasoning to define a set of mathematical tools that we will use to define problems and state their solutions in the rest of the book.

Chapter 4 revisits the evidence for human irrationality, providing an analysis of some of the classic heuristics proposed by Kahneman and Tversky from the perspective of resource-rational analysis. The heuristics include **anchoring-and-adjustment** (Tversky and Kahneman, 1974) and the **availability heuristic** (Tversky and Kahneman, 1973), both of which are related to ideas about how to best use sampling algorithms to estimate probability distributions.

Chapter 5 applies resource-rational analysis to strategy selection. According to explanations of human decision-making postulating heuristics, the

mind is equipped with a toolbox full of different decision strategies. This raises the question of how people know when to use which strategy. This chapter explains how the problem of selecting a strategy can be expressed as a form of rational metareasoning—a problem that has been studied in the artificial intelligence literature (Russell and Wefald, 1991a). We then present a theory according to which people solve this problem by learning to predict how well different strategies work in different situations. We conclude the chapter with some empirical findings demonstrating that people can learn to become more rational.

Chapter 6 turns from strategy selection to strategy discovery. While strategy selection is about choosing between known strategies, strategy discovery is much harder: it requires us to put together new strategies out of elementary cognitive operations. In this chapter, we show how this problem can be solved using **metalevel Markov decision processes** (Hay et al., 2012), with applications to decision-making, planning, and memory.

Chapter 7 considers the implications of resource-rational analysis for two other key questions in cognitive science: how should we represent the world and how should human minds be structured? In answering the first question, we show how thinking about representations in terms of the computational costs they incur can account for how people form abstractions (Ho et al., 2022; Correa et al., 2023). In answering the second question, we show how resource rationality favors minds composed of a small number of cognitive systems (justifying the idea of having a "fast" and a "slow" system) (Milli et al., 2021).

Chapter 8 considers how resource-rational analysis can be used as a guide to constructing effective interventions for improving people's decisions. In this chapter, we show how having a more accurate normative standard for people's behavior can allow us to better identify real opportunities for improvement and to derive effective interventions to achieve those improvements. In particular, we discuss the approaches we have taken to teach people resource-rational decision strategies (Callaway et al., 2022a), construct cognitive prostheses (Lieder et al., 2019), and nudge people toward making better decisions with less effort (Callaway et al., 2023).

Chapter 9 concludes the book, summarizing the results from the previous chapters and considering their implications, as well as highlighting future directions for this research program.

We have constructed the book with the intent that it be read from start to finish, with later chapters building on ideas presented in earlier chapters and assuming some of their content as context. Hence, perhaps appropriately, the main decision you have to make about how to best use your cognitive resources is when to stop reading and start acting on the things you have learned.

INDEX

Page numbers in italics indicate figures and tables.