UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Beyond bounded rationality: Reverse-engineering and enhancing human intelligence

Permalink
https://escholarship.org/uc/item/0mh5z130

Author
Lieder, Falk

Publication Date
2018

Peer reviewed|Thesis/dissertation

Beyond bounded rationality: Reverse-engineering and enhancing human intelligence

By

Falk Lieder

A Dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Neuroscience

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Thomas L. Griffiths, Chair
Professor Silvia Bunge
Professor Friedrich T. Sommer
Professor Stuart J. Russell

Spring 2018

Abstract

Beyond bounded rationality: Reverse-engineering and enhancing human intelligence

by Falk Lieder

Doctor of Philosophy in Neuroscience

University of California, Berkeley

Professor Thomas L. Griffiths, Chair

Bad decisions can have devastating consequences, and there is a vast body of literature suggesting that human judgment and decision-making are riddled with numerous systematic violations of the rules of logic, probability theory, and expected utility theory. The discovery of these *cognitive biases* in the 1970s challenged the concept of Homo sapiens as the rational animal and has profoundly shaken the foundations of economics and rational models in the cognitive, neural, and social sciences. Four decades later, these disciplines still lack a rigorous theoretical foundation that can account for people's cognitive biases. Furthermore, designing effective interventions to remedy cognitive biases and improve human judgment and decision-making is still an art rather than a science. I address these two fundamental problems in the first and the second part of my thesis respectively.

To develop a theoretical framework that can account for cognitive biases, I start from the assumption that human cognition is fundamentally constrained by limited time and the human brain's finite computational resources. Based on this assumption, I redefine human rationality as reasoning and deciding according to cognitive strategies that make the best possible use of the mind's limited resources. I draw on the bounded optimality framework developed in the artificial intelligence literature to translate this definition into a mathematically precise theory of bounded rationality called *resource-rationality* and a new paradigm for cognitive modeling called *resource-rational analysis*. Applying this methodology allowed me to derive resource-rational models of judgment and decision-making that accurately capture a wide range of cognitive biases, including the anchoring bias and the numerous availability biases in memory recall, judgment, and decision-making. By showing that these phenomena and the heuristics that generate them are consistent with the rational use of limited resources, my analysis provides a rational reinterpretation of cognitive biases that were once interpreted as hallmarks of human irrationality. This suggests that it is time to revisit the debate about human rationality with the more realistic normative standard of resource-rationality. To enable a systematic assessment of the extent to which human cognition is resource-rational, I present an automatic method for deriving resource-rational heuristics from a mathematical specification of their function and the mind's computational constraints. Applying this method to multi-alternative risky-choice led to the discovery of a previously unknown heuristic that people appear to use very

frequently. Evaluating human decision-making against resource-rational heuristics suggested that, on average, human decision-making is at most 88% as resource-rational as it could be.

Since people are equipped with multiple heuristics, a complete normative theory of bounded rationality also has to answer the question of when each of these heuristics should be used. I address this question with a rational theory of strategy selection. According to this theory, people gradually learn to select the heuristic with the best possible speed-accuracy trade-off by building a predictive model of its performance. Experiments testing this model confirmed that people gradually learn to make increasingly more rational use of their finite time and bounded cognitive resources through a metacognitive reinforcement learning mechanism.

Overall, these findings suggest that–contrary to the bleak picture painted by previous research on heuristics and biases–human cognition is not fundamentally irrational, and can be understood as making rational use of bounded cognitive resources. By reconciling rationality with cognitive biases and bounded resources, this line of research addresses fundamental problems of previous rational modeling frameworks, such as expected utility theory, logic, and probability theory. Resource-rationality might thus come to replace classical notions of rationality as a theoretical foundation for modeling human judgment and decision-making in economics, psychology, neuroscience, and other cognitive and social sciences.

In the second part of my dissertation, I apply the principle of resource-rationality to develop tools and interventions for improving the human mind. Early interventions educated people about cognitive biases and taught them the normative principles of logic, probability theory, and expected utility theory. The practical benefits of such interventions are limited because the computational demands of applying them to the complex problems people face in everyday life far exceed individuals' cognitive capacities. Instead, the principle of resource-rationality suggests that people should rely on simple, computationally efficient heuristics that are well adapted to the structure of their environments. Building on this idea, I leverage the automatic strategy discovery method and insights into metacognitive learning from the first part of my dissertation to develop intelligent systems that teach people resource-rational cognitive strategies. I illustrate this approach by developing and evaluating a cognitive tutor that trains people to plan resource-rationally. My results show that practicing with the cognitive tutor improves people's planning strategies significantly more than does practicing without feedback. Follow-up experiments demonstrate that this training effect transfers to more difficult planning problems in novel and more complex environments, and that this transfer effect is retained over time. This indicates that discovering and teaching resource-rational heuristics may be a promising approach to improving human judgment and decision-making. While this approach adapts people's heuristics to the structure of their environment, the theory of resource-rationality suggests that human decision-making can also be improved by adapting the structure of the environment to the heuristics people already use. I illustrate this approach by developing a cognitive prosthesis for goal achievement that helps people overcome procrastination, spring into action, and achieve their goals on time.

By virtue of integrating rational principles with cognitive constraints, resource-rationality provides a realistic normative standard for human reasoning and decision-making. My findings about human rationality and metacognitive learning are consistent with the view that evolution and learning adapt the mind to the structure of its environment and the constraints imposed by its limited resources. These adaptive mechanisms appear to optimize for resource-rationality, and the benefits of training with the cognitive tutor demonstrate that this adaptation can be accelerated with the help of artificial intelligence. This makes resource-rationality a promising theoretical framework for modeling and improving human cognition.

i

Contents

Acknowledgments

I am greatly privileged to have had the opportunity to pursue my passion, in developing a theoretical foundation and practical tools for improving the human mind throughout this dissertation. This would not have been possible without the generous support and ingenuity of those who have helped and inspired me.

I am most grateful to my advisor Tom, who gave me the freedom to explore my intellectual interests while providing insightful guidance, great mentorship, thoughtful contributions, and generous support along the way. My gratitude extends to Silvia Bunge for sharing her insights on cognitive training, and for giving me advice on professional development. My thanks also go to those who made it possible for me to pursue this research as part of the neuroscience Ph.D. program, including the members of my dissertation committee and the committee of my qualifying examination, namely Tom Griffiths, Silvia Bunge, Stuart Russell, Fritz Sommer, Ming Hsu; the mentors of my lab rotations, Tom Griffiths, Ming Hsu, and Stuart Russell; as well as Dan Feldman, the former director of the neuroscience Ph.D. program, and current director Michael Silver.

I would also like to acknowledge Fred Callaway, Sayan Gul, Owen Chen, Paul Krueger, and Jim Ruther Nill, each of whom played an important role in the research presented in Chapters 5–6 of Part 1 and Chapters 8–9 of Part 2. Fred Callaway programmed the Mouselab-MDP paradigm and most of the computational infrastructure that made the cognitive tutor possible, and he was crucial to the development of the strategy discovery method presented in Chapter 6. Sayan Gul made significant contributions to the development and evaluation of that very strategy, and helped me to explore its application to multi-alternative risky choice. Paul Krueger's input was vital to the experiments and analyses presented in Chapters 5 and 6, he also helped to develop the automatic strategy discovery method and contributed substantially to the evaluations of the cognitive tutor presented in Chapter 9. Owen Chen and Jim Ruther Nill did most of the work of programming the to-do list gamification app evaluated in Chapter 8. Furthermore, Priyam Das assisted me in running earlier experiments on cognitive tutoring. Without the help of this incredible team, I would not have been able to take this research program as far as I did, nor as quickly as I did. Additionally, I would like to thank Noah Goodman for his contributions to the research presented in Chapters 1 and 2; Quentin Huys for his contributions to the research presented in Chapter 2; Ming Hsu for his contributions to the research presented in Chapter 3; and Dillon Plunkett, Jessica Hamrick, and Stuart Russell for their contributions to research presented in Chapter 4. I would also like to thank the 11 research assistants I supervised during my time as a Ph.D. student, namely Phoebe Lin, Ronald Kwan, Sidharth Goel, Priyam Das, David Lu, Zisu Dong, Owen Chen, Eric Zhang, Andrew Tan, Nicholas Qi Cai,

and Sayan Gul. The footnotes at the beginning of each chapter provide a more detailed acknowledgement of each collaborator's contributions to the work presented in my dissertation.

My greatest intellectual debts are to my advisor; for his ability to draw useful connections between problems in psychology and ideas in machine learning, as well as artificial intelligence. I am also indebted to the inspiring, fundamental work on rational metareasoning and bounded optimality by Stuart Russell and colleagues (Hay, Russell, Tolpin, & Shimony, 2012; Russell, 1997; Russell & Subramanian, 1995; Russell & Wefald, 1991a, 1991b) and related work by Eric Horvitz (Horvitz, Cooper, & Heckerman, 1989). I would like to to thank Thomas Icard and Andreas Stuhlmüller for pointing me towards this work. It has fundamentally shaped the way I think about bounded rationality. Furthermore, the cognitive prosthesis developed in Chapter 8 and the cognitive tutor developed in Chapter 9 are based on the theory of reward shaping developed by Ng, Harada, and Russell (1999) and I am grateful to Tom for recognizing its relevance to improving human decision-making. These influences have been truly instrumental to my Ph.D.

Looking back further, I would like to recognize everyone who made it possible for me to eventually embark on a Ph.D. at UC Berkeley, and give thanks to those who prepared me for it. Thinking back to who I was as a child reminds me that my life could have turned out very differently; and this makes me realize how astonishingly fortunate I have been. Starting from the beginning, these people include my parents; the high school math teacher who noticed my precocity and recommended that I switch to another school with a special focus on mathematics and the sciences; my high school psychology teacher, Frank Meitzner, who referred me to a professor at the local university when I was in 9th grade; and the researchers who generously mentored me and gave me hands-on research experience during my last 4 years in high-school, namely Lars-Eric Petersen, Astrid Milde-Busch, Michael Hanke, and Dieter Heyer. I also benefited from the education and research possibilities I received as an undergraduate student at the university of Osnabrück, including internships in the labs of Peter König and Carsten Konrad. I gained a great deal from the opportunity to study abroad at ETH Zürich. I benefited tremendously from working with Klaas-Enno Stephan, whose vision to revolutionize psychiatry with the help of computational modeling inspired me to formulate a meaningful scientific vision of my own. These ideas crystallized into personal transformation when I read a book that my brother had given me for Christmas;"The 7 Habits of Highly Effective People". My vision was further shaped by discovering the research of Josh Tenenbaum and colleagues; attending the IPAM summer school on probabilistic models of cognition as recommended by Andreas Stuhlmüller; and starting a research project under the supervision of Noah Goodman and Tom Griffiths. Formulating this scientific vision gave me meaning, made me who I am today, and is ultimately responsible for the research presented in this dissertation. This leaves me eternally grateful to everyone whose influence on me inspired and shaped my outlook, and scientific project. I was lucky that Tom and Noah gave me to opportunity to collaborate with them while I was a masters student at ETH Zürich. My work with them kickstarted my own research program on bounded rationality and open many doors for me – including the door to the neuroscience Ph.D. program at UC Berkeley.

0

General Introduction

The decisions we make determine our personal and collective destiny. Technological, scientific, social, and cultural advances have given us tremendous power over our lives, the lives of others, and the future of humanity. The power of our choices comes with the responsibility to choose wisely. Yet making good decisions is much easier said than done. We have all witnessed regrettable decisions, unwarranted conclusions, and questionable arguments more often than we would like.

To address the problem of questionable arguments and unwarranted conclusions, Aristotle set out to characterize what distinguishes valid inferences from fallacies (R. Smith, 2017). His efforts laid the foundation of modern logic. This inspired the creation of artificial intelligence and became a normative standard for human reasoning. While logic is a normative theory of deductive reasoning under certainty, human reasoning often involves uncertainty and inferring unobservable principles from limited data. Bayesian statistics holds that we should draw such inferences by updating our beliefs according to the rules of probability theory (Bayes, 1763; Laplace & Simon, 1951; Savage, 1971). Finally, expected utility theory (von Neumann & Morgenstern, 1944) prescribes that we should always choose the course of action that maximizes our expected utility.

Over the past 50 years a substantial literature on heuristics and biases has documented that people's judgments and decisions often violate these normative principles (Gilovich, Griffin, & Kahneman, 2002; Tversky & Kahneman, 1974; Wason, 1968). These systematic errors are known as *cogni-*

tive biases. As the resulting errors can have severe consequences, developing interventions to remedy these biases has become a prominent research topic. In the following paragraphs, I briefly review the main approaches that have been explored previously and identify their limitations, which motivate the research of this dissertation.

Debiasing, the first approach, aims to eliminate or reduce cognitive biases through motivational, cognitive, or technological interventions (Larrick, 2002). Motivational approaches to debiasing seek to reduce cognitive biases by adding financial incentives for good performance or by holding people accountable. Incentives and accountability generally increase effort, however, this does not necessarily equate to increased performance (Camerer & Hogarth, 1999). The effectiveness of motivational approaches appears to critically depend on whether people already possess effective cognitive strategies (Camerer & Hogarth, 1999; Lerner & Tetlock, 1999).

Cognitive approaches to debiasing teach strategies that are consistent with normative principles or approximate them. It aims for strategies that are simple and memorable. Early cognitive approaches taught people basic statistical principles (e.g., (a) the law of large numbers and (b) the variability of small samples [Fong & Nisbett, 1991]) and simple implications of normative principles (e.g., (a) how to check whether an if-then statement is true [Cheng, Holyoak, Nisbett, & Oliver, 1986] or (b) that sunk costs should be ignored [Larrick, Morgan, & Nisbett, 1990]). People learned to apply those simple rules to simple problems. Moreover, some studies found transfer to simple problems that are superficially different from the examples used during training (Fong & Nisbett, 1991, e.g.,). Another successful example of cognitive debiasing is teaching people to ask themselves why their initial judgment or decision might be wrong. This strategy has been found to reduce overconfidence, the anchoring bias, and the hindsight bias (Arkes, 1991; Mussweiler & Strack, 2000). However, cognitive approaches to debiasing and their evaluation have been restricted to simple rules for simple problems, and Larrick (2002) argued that it would be unsuitable for more complex normative strategies such as Bayes rule.

Technological approaches to debiasing include: (a) replacing human judgments by regression models (Dawes, Faust, & Meehl, 1989), (b) performing decision analysis (Howard, 1988), and (c) decision support systems (Power, Sharda, & Burstein, 2015). Decision analysis guides people to decompose their decision problem, estimate its components, and then combine their estimates according to expected utility theory. The effectiveness of these tools remains to be evaluated (Larrick, 2002). Decision support systems carry out facets of the decision process for the decision-maker, to compensate for their cognitive limitations. They can thus be interpreted as cognitive prostheses. Most decision support systems are highly specific to a particular domain, such as supply chain management

for a particular industry. There are more general decision support systems based on decision-analysis (Edwards & Fasolo, 2001). Unfortunately, they inherit the issues arising from the inaccuracy and biases of people's probability judgments and utility estimates.

Despite their differences, all of these approaches to debiasing are based on two assumptions. First, they assume that all cognitive biases reflect irrational heuristics and suboptimal cognitive performance. Second, since this literature defines cognitive biases as deviations from the rules of logic, probability theory, and expected utility theory, its interventions aim to bring people's cognitive strategies into closer alignment with those normative principles. In Part 1 of my dissertation, I argue that the assumptions of debiasing are flawed because logic, probability theory, and expected utility theory are unrealistically high normative standards that are oblivious to the computational constraints that people have to work with. Consequently, the traditional approach of debiasing might not be the most effective way to improve human judgment and decision-making in complex real-life situations.

Similar to debiasing, boosting (Hertwig & Grüne-Yanoff, 2017) aims to increase people's decision-making competency. But in contrast to debiasing it does not define competency as adhering to the rules of expected utility theory, logic, and probability theory. Instead, it views a competent decision-maker as somebody who uses simple heuristics that are well adapted to the structure of their environment. Boosting therefore aims to teach people simple rules of thumb that differ considerably from the normative rules taught in classic debiasing interventions. Boosting also aims to change how information is presented to match the presentation format that people's heuristics are adapted to. One successful example of this approach is to present conditional probabilities as natural frequencies. This intervention has been shown to significantly improve people's performance at Bayesian reasoning (Gigerenzer & Hoffrage, 1995). For a more lasting effect, people can be taught to translate conditional probabilities into natural frequencies by themselves (Sedlmeier & Gigerenzer, 2001). Despite these successes, the effectiveness of boosting is limited by our ability to discover effective heuristics. It would be a coincidence if the heuristics people are currently taught were already optimal. So there may still be a lot of room for improvement in the curriculum of boosting. But coming up with better heuristics is very difficult. This dissertation addresses this challenge by developing a principled method for deriving optimal heuristics automatically.

While debiasing and boosting target judgment and decision-making directly, *cognitive training* targets the basic underlying cognitive capacities such as working memory (Klingberg, 2010), processing speed (Ball, Edwards, & Ross, 2007; Nouchi et al., 2012, 2013), attention (Slagter et al., 2007; Tang & Posner, 2009), and cognitive control (Anguera et al., 2013; Karbach & Kray, 2009; Nouchi

et al., 2012, 2013). Generally, cognitive training leads to reliable improvements on the trained task. While these improvements frequently transfer to similar tasks, they rarely transfer to performance in everyday life. Whether existing training programs, such as working memory training, achieve meaningful transfer effects is the subject of a heated debate (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Melby-Lervåg & Hulme, 2013; Morrison & Chein, 2011; Owen et al., 2010; Redick et al., 2013; Shipstead, Redick, & Engle, 2012). Furthermore, the learning mechanisms underlying potentially generalizable improvements in high-level cognition remain poorly understood. Furthermore, there is very little theoretical guidance for designing effective training regimens; this might be a serious bottleneck to the development of effective cognitive training programs. To address this problem Chapters 5 and 9 of my dissertation propose and evaluate a theoretical framework and computational tools for developing a new kind of cognitive training program.

In contrast to debiasing, boosting, and cognitive training, *nudging* aims to exploit people's cognitive biases instead of trying to remedy them (Thaler & Sunstein, 2008). Nudging structures decision environments in such a way that people's biases favor a desirable decision without restricting their freedom of choice. This methodology has been successfully used in public policy to promote organ donation and saving for retirement. The most prominent example of nudging is to make the presumably better option (e.g., to save for retirement) the default, while allowing people to opt out. There are many more examples of how the presentation of choices can be tweaked to improve people's decisions (Johnson et al., 2012). To date, nudging is primarily used in public policy. It allows governments and organizations to gently nudge citizens and consumers towards pro-social and responsible behavior. But the widespread use of nudging raises concerns about manipulation and unintended side effects, which remain unaddressed. Furthermore, nudges are typically only available for a very small fraction of the thousands of decisions we have to make every day. In actuality there are very few tools that people can use to nudge themselves to perform a desired task. Chapter 7 of my dissertation addresses this problem by developing a theory-based approach to nudging and a practical tool that individuals can use to nudge themselves towards their goals.

In summary, while there are at least four existing approaches to improving human judgment and decision-making, all of them have serious limitations. Overcoming those limitations will require significantly more research into their theoretical foundations.

My research is driven by the need for a solid theoretical and computational foundation for improving human decision-making. To be useful, a theoretical framework for improving the human mind should be able to answer the following questions:

1. How should we think and decide to make the best possible use of our limited time and cognitive resources?

2. How does human cognition compare to this idea?

3. How do we learn to think more clearly and make better decisions?

4. What can be done to promote cognitive growth?

5. How can we help people overcome their cognitive limitations and achieve their goals?

This dissertation addresses each of these questions within the domains of judgment, decision-making, and planning. The first part of my dissertation addresses questions 1–3 via a combination of computational modeling and behavioral experiments. The second part of my dissertation addresses questions 4 and 5. There, I leverage the theoretical framework developed in Part I to formulate interventions for expanding the bounds of human rationality. These interventions take the form of a cognitive tutor that teaches people optimal planning strategies, and a cognitive prosthesis that lets people nudge themselves towards their goals.

My research on optimal reasoning and decision-making under limited resources sheds new light on the debate about human rationality. It challenges the conclusion that people are fundamentally irrational and provides more realistic normative standards for assessing human rationality. The theory and computational tools I have developed for deriving optimal cognitive strategies provide a useful methodology for cognitive modeling, one that lets researchers leverage the power of normative principles to develop precise mathematical models of cognitive mechanisms. Furthermore, my model of strategy selection closes an important gap in theories of bounded rationality. Those theories postulate that the mind is equipped with a toolbox of heuristics, and the model presented in Chapter 4 completes them by specifying how people should decide when to use which heuristic. My research on metacognitive learning leads to a nuanced, dynamic perspective on what it means to be rational. According to this view, rationality entails gradually learning to make increasingly more effective use of fallible heuristics. This perspective reconciles people's use of fallible heuristics with the normative principles of rational decision-making and rational learning. Furthermore, it shifts the focus from how people think and decide when they are tested to how their reasoning and decision-making improve over time. According to this dynamic view, human rationality should be measured by people's ability to improve their reasoning and decision-making based on their experience.

In addition to these scientific contributions, the research presented in this dissertation is also a step towards several practical applications for helping people make better decisions. First, the theory of resource-rationality makes it possible to derive optimal cognitive strategies that might enable

people to make better decisions and think more clearly. Second, the theory and models of metacognitive learning (Chapter 5) provide guidance for how to promote and accelerate cognitive growth. To support this argument, I show that these principles can be used to design feedback mechanisms that make cognitive training more effective (Chapter 9). Finally, my research on decision-support (Chapter 8) provides a theoretical foundation for developing cognitive prostheses for goal achievement. The approach I have taken leverages artificial intelligence to enable people to effectively nudge themselves towards their goals. As a proof-of-concept, I present a to-do list gamification app that can help people overcome procrastination and achieve their goals on time.

This dissertation is structured into two parts: The six chapters of the Part 1 develop a methodology for deriving realistic normative models of human cognition. By taking into account people's finite time and bounded cognitive resources, these rational models can explain cognitive biases that would otherwise appear irrational. Chapter 1 reviews the literature on rational models of reasoning and decision-making with limited cognitive resources and identifies open problems. Chapter 2 develops a resource-rational model of a ubiquitous systematic error in human judgment: the anchoring bias. Chapter 3 develops a resource-rational model of a wide range of availability biases in human decision-making, judgment, and memory recall. These findings establish that at least some heuristics can be understood as resource-rational cognitive strategies for specific problems. However, no single heuristic is resource-rational for all problems. Thus, achieving resource-rationality requires adaptively choosing between multiple heuristics. Chapter 4 formalizes this idea by a rational model of strategy selection and tests its predictions in behavioral experiments. The findings suggest that people can select heuristics adaptively because they have learned to predict how well each heuristic will perform for different problems. Inspired by these results, Chapter 5 tests the hypothesis that people gradually learn to make increasingly more rational use of their limited cognitive resources. A series of experiments confirmed this prediction. Follow-up experiments suggested that these improvements may be driven by a metacognitive reinforcement learning mechanism. Finally, while I derived the resource-rational models presented in Chapters 1-5 by hand, Chapter 6 presents and evaluates a method for deriving resource-rational models automatically. The primary objective of all of this research is to go beyond fuzzy, verbal theories of bounded rationality. This is achieved by developing mathematically precise normative models of how people should think and decide; and then leveraging those models to revisit the debate about human rationality.

Part 2 of this dissertation applies these advances to develop tools and interventions to push the boundaries of human rationality farther outward. Concretely, Chapter 8 develops a cognitive prosthesis that augments the brain's decision-making systems. Chapter 9 develops a cognitive tutor that

teaches people resource-rational planning strategies. Finally, I conclude by discussing the implications of these findings for understanding and improving human rationality.

Most chapters of my dissertation are based on previously published articles with several co-authors. In these chapters, I will use the pronoun "we" when describing work that I or my collaborators have done as part of this project. When collaborators have performed substantive components of the presented work this is explicitly acknowledged in the footnote at the beginning of the corresponding chapter. I will continue to use the pronoun "I" in the chapters and sections that are not based on published collaborative work and also to express my personal opinion – which does not necessarily reflect the views of my co-authors.

Part I

Bounded rationality revisited

Introduction to Part I

As laid out in the General Introduction, I believe that developing a precise normative theory of bounded rationality will allow us to establish tools and interventions that will help people think more clearly, and make better decisions. Here, I will argue that bounded optimality is a promising theoretical framework for building such a theory. To illustrate the potential of this framework, I apply it to derive bounded-optimal models of judgment and decision-making; which I will then bring to bear on the debate about human rationality.

The first four chapters argue that human rationality should be understood in terms of the optimal use of our finite time and limited cognitive resources. I formalize this idea within the framework of bounded optimality, as established in the artificial intelligence literature (Russell & Subramanian, 1995), in Chapter 1. This leads to a new normative standard for human cognition called *resource-rationality* and a new methodology for cognitive modeling called *resource-rational analysis*. In the subsequent chapters, I apply this methodology to derive resource-rational models of judgment (Chapter 2), decision-making (Chapter 3), and strategy selection (Chapter 4). I illustrate how such models can contribute to a better understanding of human cognition, and I use them to revisit the debate about human rationality. In order to make resource-rational analysis more easily applicable to a wider range of phenomena, Chapter 6 develops an automatic method for deriving rational process models from first principles. As a proof of concept, I apply this method to multi-alternative risky choice and show that it yields new insights into the mechanisms of human decision-making. This enables a quantitative assessment of human rationality against a realistic normative standard. These case studies illustrate that resource-rational analysis is a promising paradigm for modeling human cognition.

The analyses mentioned so far derived the optimal cognitive strategies for fixed cognitive architectures confronting known environments. Humans, however, often confront unknown environments with a changing brain. I believe that a complete theory of human rationality should take these additional challenges into account by specifying bounded-optimal learning mechanisms that render cognitive mechanisms increasingly more resource-rational by adapting them to cognitive constraints and the structure of the environment. As a first step in this direction, I develop a theory of how peo-

ple learn when to use which heuristic, and outline how this approach could be expanded in order to model how those strategies are learned in the first place (Chapter 5). Based on these models, I hypothesize that people generally learn to make increasingly more rational use of their limited cognitive resources over time. I will argue that an understanding of human rationality which is based on learning, and rooted in a realistic normative standard, can capture human performance far more accurately than models derived from the unrealistically high normative standards of logic, probability theory, and expected utility theory.

1

Resource-Rationality*

1.1 INTRODUCTION

What does it mean to be rational? How should we reason about what is true and how should we decide what to do? Do the cognitive strategies that people already use come close to these ideals or are they far off? Which strategies should we teach people to improve the quality of their reasoning and decision-making?

This chapter introduces the theoretical framework used to answer these questions. Its central idea is that people should make rational use of their finite time and limited cognitive resources. The chapter starts by briefly summarizing and discussing previous theories of rationality and the debate about human rationality. The open problems and limitations of previous work are pointed out and the new theoretical framework of resource-rationality is introduced to address them. This framework leads to a new methodology called *resource-rational analysis* that is the foundation for the work presented in Part 1. To illustrate the utility of taking a resource-rational perspective for understanding human cognition, previous work that can be understood within the resource-rational framework is also reviewed. The chapter closes with a preview of how the subsequent chapters will

---

*This chapter reuses material from Griffiths, Lieder, and Goodman (2015).

build on and apply this overarching theoretical principle.

1.2 NOTIONS OF RATIONALITY

Existing definitions of rationality differ along four dimensions: The first distinction is whether rationality is defined in terms of beliefs (*theoretical rationality*) or actions (*practical rationality*; Harman, 2013). The second distinction is whether rationality is judged by the reasoning process (*rule-based*) or its consequences (*consequentialism*; Sosis & Bishop, 2013). Third, some notions of rationality take into account the agent's computational capacity are bounded (*bounded rationality*) whereas others do not (*unbounded rationality*; Lewis, Howes, & Singh, 2014; Russell, 1997). Fourth, rationality may be defined either by the agent's performance on a specific task or by its average performance in the natural environment (*ecological rationality*; Chater & Oaksford, 2000; Gigerenzer, 2008b; Lewis et al., 2014).

The most influential rule-based notion of rationality is logic. A logic is a formal system of inference rules for transforming one set of formula into another set such that the resulting formula (conclusion) will be true if the initial formula (premises) were true (Reichenbach, 1947). The most basic form of logic is *propositional logic* where each formula comprises atomic statements, such as "Aristotle was a human." that can be connected by **AND**, **OR**, **IF ... THEN ...**, and **NOT**. While logic defines rational rules for reasoning from statements that are known to be true or false, probability theory defines rational rules for reasoning under uncertainty. Concretely, Bayesian rationality holds that people should reason according to the laws of Bayesian probability theory (Oaksford & Chater, 2007). This entails maintaining graded beliefs over alternative hypotheses $\theta \in \Theta$. The degree of belief $P(\theta)$ in a hypothesis θ is formalized as a probability such that $P(\theta|K) = 0$ if and only if the hypothesis cannot possibly be true and $P(\theta|K) = 1$, if and only if, the hypothesis cannot possibly be false given the agent's knowledge K. The latter entails that the sum of the probabilities assigned to all possible values a state of the world can take has to be equal to 1. Finally, the last critical element of probability theory is the notion of conditional probability. Concretely, the conditional probability of θ_1 given θ_2, which is written as $P(\theta_1|\theta_2)$, specifies how strongly one should belief in θ_1 being true if θ_2 was true. Formally, the conditional probability of θ_1 given θ_2 is defined as

$$P(\theta_1|\theta_2) = \frac{P(\theta_1 \wedge \theta_2)}{P(\theta_2)}, \tag{1.1}$$

where the proposition $\theta_1 \wedge \theta_2$ defines the set of worlds in which both propositions (θ_1 and θ_2) are

true simultaneously. The definition notion of conditional probability enforces that all of the agent's beliefs are coherent with each other. Furthermore, Bayesian rationality also demands that all beliefs should be consistent with the agent's observations o. Concretely, probability theory entails that when the agent makes a new observation o, then it should update its belief in each of its hypotheses $\theta \in \Theta$ from $p(\theta)$, which is known as the *prior* to $P(\theta|o)$ which is known as the posterior. Concretely, *Bayes theorem* (Bayes, 1763) holds that a rational agent's posterior belief in hypothesis θ after having seen observation o should be

$$P(\theta|o) = \frac{P(o|\theta) \cdot P(\theta)}{P(\theta) \cdot P(o|\theta) + P(\neg\theta) \cdot P(o|\neg\theta) \cdot}, \tag{1.2}$$

where $\neg\theta$ is the negation of hypothesis θ. This means that observing o turns the odds of θ being true versus false from $\frac{P(\theta)}{P(\neg\theta)}$ into

$$\frac{P(\theta|o)}{P(\neg\theta|o)} = \frac{P(\theta)}{P(\neg\theta)} \cdot \frac{P(o|\theta)}{P(o|\neg\theta)}. \tag{1.3}$$

Intuitively, this means that an observation should increase your degree of belief in a hypothesis θ proportionally to how much more likely that observation is to occur if θ than if θ was false.

By contrast to these process-based notions of rationality, consequentialist notions of rationality evaluate human reasoning based on its outcomes (Sosis & Bishop, 2013). There are two main versions of consequentialist theoretical rationality: *reliabilism* and *pragmatism*. Reliabilism evaluates reasoning strategies based on how reliably they yield correct conclusions across a wide range problems. By contrast, pragmatism evaluates the reasoning strategies according to the usefulness of the resulting beliefs regardless of their factual accuracy. For instance, if a reasoning strategy leads us to incorrectly conclude that filing our taxes will be fun that counts against its rationality from the reliabilist perspective. But from the pragmatist perspective a factually incorrect inference about how enjoyable it is to file taxes could count as rational if it helps you avoid the negative consequences of procrastinating on filing your taxes for too long. The most prominent version of consequentialism is expected utility theory (von Neumann & Morgenstern, 1944). According to expected utility theory, a rational decision-maker should always choose the action $a^\star$ that maximizes the expected utility of the resulting outcome O, that is

$$a^\star = \arg\max_a \mathbb{E}_{P(O|s,a)} \left[u(O)|s, a \right], \tag{1.4}$$

where the agent's utility function u defines how good or bad different outcomes are with respect to

the agent's goals and the outcome O includes both the immediate reward and the next state.

While process-based notions of rationality are conceptually very different from consequentialist notions of rationality, and theoretical rationality is distinct from practical rationality, all of them could, in principle, be attained simultaneously. This might be why the standard picture of rationality combines them by demanding that people reason according to the normative rules (Sosis & Bishop, 2013) of logic and probability (which is a process-based notion of theoretical rationality) and acting according to the maxim to maximize expected utility (which is a form of consequentialism). In my view, the critical shortcoming of the standard picture of rationality are that it does not recognize that to think and decide effectively in the world people and machines have to make efficient use of their limited time and bounded computational resources. Incorporating cognitive constraints into theories of rationality began with the foundational work of Herbert Simon who argued that computational limitations place substantial constraints on human reasoning (Simon, 1972, 1982). The following sections briefly summarize Simon's work on bounded rationality and subsequent extensions and refinements.

1.2.1 EARLY EXTENSIONS TO BOUNDED AGENTS

Simon pointed out that our finite computational capacities make it impossible for us to always find the best course of action, because we cannot consider all possible consequences. He illustrated this using the game of chess, where choosing the optimal move would require considering about 10^{120} possible continuations. Thus, Simon concluded, to adequately model human behavior, we need a theory of rationality that takes our minds' limitations into account. Simon called such an approach *bounded rationality*, emphasizing that it depends on the structure of the environment (Simon, 1956) and entails satisficing, that is accepting sub-optimal solutions that are good enough in place of striving for the very best solution possible. While he provided some formal examples of satisficing strategies (Simon, 1955), Simon viewed bounded optimality as a principle rather than a formal framework.

Arguably, the question of what it means to be rational in the face of limited computational resources is also fundamental to the endeavour of creating artificial intelligence (Russell, 1997). So, it might come as no surprise that computer scientists have subsequently expanded Simon's ideas on bounded rationality into formal theories of computational rationality (for a review see Gershman, Horvitz, & Tenenbaum, 2015). Two early notions of computational rationality were *calculative rationality* and *Type II rationality*. Calculative rationality refers to algorithms whose answers would, eventually, converge to the optimal solution within the limits of infinite computation. These algo-

rithms are commonly run for a much shorter length of time than would be necessary to guarantee their convergence to the optimal solution but their asymptotic guarantees are seen as evidence that they are at least approximating the right thing. Good (1983) defined Type II rationality as the maximization of expected utility taking into account the cost of deliberation. Intuitively, this means that rational bounded agents optimally trade off the expected utility of the action that will be chosen with the corresponding deliberation costs. But Good (1983) did not make this notion mathematically precise.

1.2.2   Rational metareasoning

Later work on *rational metareasoning* formalized Good's ideas with mathematical precision (Horvitz, 1987; Russell & Wefald, 1991b). If reasoning seeks an answer to the question "what should I do?", metareasoning seeks an answer to the question "how should I decide what to do?". The theory of rational metareasoning (Russell & Subramanian, 1995; Russell & Wefald, 1991b) frames this problem as selecting computations so as to maximize the sum of the rewards of resulting decisions minus the costs of the computations involved. Concretely, one can formalize reasoning as a meta-level Markov decision process (meta-level MDP) and metareasoning as solving that MDP (Hay et al., 2012). In brief, a meta-level MDP

$$M_{\text{meta}} = (\mathcal{B}, \mathcal{A}, T_{\text{meta}}, r_{\text{meta}}) \tag{1.5}$$

is a Markov decision process (Puterman, 2014) where the actions $\mathcal{A}$ are computations, the states $\mathcal{B}$ encode the agent's beliefs, and the transition function T_{meta} describes how the computations change those beliefs. $\mathcal{A}$ includes computations $\mathcal{C}$ that update the beliefs, as well as, a special meta-level action $\perp$ that terminates deliberation and initiates acting on the current belief. A belief state b encodes a probability distribution over parameters θ of a model in the domain. The meta-level reward function r_{meta} captures the cost of computation and the external reward r the agent expects to receive from the environment.

This formulation makes rational metareasoning amenable to the wide range of methods that have been developed to solve Markov decision processes including dynamic programming (Puterman, 2014) and reinforcement learning (Sutton & Barto, 1998).

1.2.3 Bounded Optimality

Despite its precision and elegance, rational metareasoning does not take into account the deliberation costs of determining the optimal trade-off between the costs and benefits of reasoning about the world. This problem cannot be solved by applying the same thinking that created it because metareasoning about efficient metareasoning would invite an infinite regress. Instead, Stuart Russell and colleagues overcame this limitation by relaxing the standards of rationality from always selecting the optimal computation to running a program that performs as well as or better than any other program that the agent could execute (Russell, 1997; Russell & Subramanian, 1995; Russell & Wefald, 1991a). This standard is attainable by its very definition. It is conceivable that a bounded optimal program for a particular problem would often select sub-optimal computations because the improvement that could be achieved by selecting optimal computations would be lower than the cost of identifying them. This notion of rationality is known as *bounded optimality*.

Bounded optimality is a theoretical principle for designing intelligent programs that run on performance-limited hardware and have to interact with their environment in real time (Russell & Subramanian, 1995). Equation 1.6 defines bounded optimality as running a program program* that when run on the agent's hardware will generate world states through its decisions and whose expected utility is at least as high as those generated by any other program that the agent's hardware can execute, that is

$$\text{program}^\star = \operatorname*{arg\,max}_{\text{program} \in \text{Programs(HW)}} \mathbb{E}_{P(S_1, \cdots, S_T | S_0, A_t = \text{program(history}_t))} \left[u \left(\text{history}_T \right) \right], \qquad (1.6)$$

where Programs(HW) is the set of programs that the agent's hardware can execute, and history$_t = \{S_0, \cdots, S_t\}$ and program(history$_t$) is the action that program would choose when executed on the hardware after having observed history$_t$, and T is the number of time steps in the episode starting with situation S_0. Finally, u is the utility function that the agent is designed to optimize under the constraints of its hardware.

By solving the optimal program problem defined in Equation 1.6 it is sometimes possible to derive optimal algorithms. For instance, Russell and Subramanian (1995) derived an optimal mail sorting program. This suggests the intriguing possibility that it might also be possible to derive optimal cognitive strategies for people.

The theories of rationality summarized above have had a fundamental impact on classic theories in psychology, economics, philosophy, linguistics,neuroscience, and the social sciences (Braine, 1978; Chater, Tenenbaum, & Yuille, 2006; Fodor, 1975; Frank & Goodman, 2012; Friedman & Savage, 1948; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Harman, 2013; Hedström & Stern, 2008; Knill & Pouget, 2004; Lohman, 2008; Mill, 1882; Newell, Shaw, & Simon, 1958; Oaksford & Chater, 2007; von Neumann & Morgenstern, 1944). Whether and to what extent human reasoning satisfies the premises of these rational models and what this entails for human rationality has been intensely debated (Stanovich, 2009). More recently, the debate has shifted from "Are we rational?" to "Is building rational models a useful methodology for understanding human cognition?" and this is an important distinction (Bowers & Davis, 2012; Chater et al., 2011; Griffiths, Chater, Norris, & Pouget, 2012). This section briefly reviews both the debate about human rationality and the merits and challenges of rational modeling as a methodology for understanding human cognition.

The assumption that people are rational became hotly debated when a series of experiments suggested that people's judgments systematically violate the laws of logic (Wason, 1968) and probability theory (Tversky & Kahneman, 1974), and subsequent studies demonstrated that people's decisions systematically deviate from the prescriptions of expected utility theory (Kahneman & Tversky, 1979). These systematic errors are known as *cognitive biases*. For instance, Tversky and Kahneman (1974) demonstrated that when people are asked to compare the number of countries to a low versus high number that was generated by spinning a wheel of fortune before they estimate its value, then their estimates are systematically biased towards an irrelevant random number; this is known as the *anchoring bias*. Furthermore, people tend to dramatically overestimate the frequency of events that come to mind easily; this is known as the *availability bias*. Moreover, when people are asked to judge the probability of a sequence of coin tosses, they assign a higher probability to sequences that are less regular and, hence, more representative of randomness, even though, all sequences are equally probable; this is known as the *representativeness* heuristic. These are just three among dozens of cognitive biases that have been identified in the last four decades (Gilovich et al., 2002).

According to Tversky and Kahneman (1974), cognitive biases result from people's use of fast but fallible cognitive strategies known as *heuristics*. The discovery of cognitive biases was highly influential, because following the rules of logic and probability was assumed to be the essence of rational thinking. Evidence that people deviate from these rules thus called human rationality into question, and this doubt has shaken the foundations of economics, the social sciences, and rational

models of cognition.

The debate about human rationality concerns the interpretation of these findings, and Stanovich (2009) aptly summarized the positions taken in this debate in the following terms: *Meliorists* interpret cognitive biases as evidence that human reasoning is not as good as it could be. Meliorists often paint a bleak picture according to which people are profoundly irrational (Ariely, 2009; Marcus, 2009; Sutherland, 1992) but they are also optimistic that human reasoning can be improved (Nisbett, 1993). By contrast, *Panglosians* reject the interpretation that people are irrational by one of three arguments: The first argument maintains that the principles of logic, probability theory, or expected utility theory that were used as the yard stick of rationality have serious limitations (Gigerenzer & Goldstein, 1996; Sosis & Bishop, 2013). The second argument maintains that the mind's computational limitations are so severe that even rational people cannot be expected to conform to the normative standards of logic, probability theory, and expected utility theory; proponents of this argument are called *Apologists*. The third argument maintains that many of the apparent violations of rationality have been shown to be consistent with the rational solution to a reasonable alternative construal of the task (Austerweil & Griffiths, 2011; Griffiths & Tenenbaum, 2001; Hahn & Oaksford, 2007; Hahn & Warren, 2009; Oaksford & Chater, 1994, 2007; Tenenbaum, Griffiths, et al., 2001). These rational explanations often draw on the methodology of *rational analysis* (Anderson, 1990; Chater & Oaksford, 1999) introduced in the following section. I believe that there is merit in all three of these arguments, and Chapters 2–6 revisit the debate about human rationality with a more appropriate notion of rationality that accounts for people's cognitive constraints.

1.3.1 Rational models of cognition and Marr's levels of analysis

The debate about human rationality also has implications for how we should model the human mind. A long tradition of rational modeling has leveraged normative principles, such as Bayes theorem and expected utility theory, to explain human behavior. As more and more violations of these normative principles surfaced in the research on judgment and decision-making, the interpretation of rational models has become increasingly constrained to what David Marr called the *computational level* of analysis (Marr, 1982) which defines the function of a cognitive system in terms of the problem that it solves and the optimal solution to that problem. Marr distinguishes the computational level of analysis from the *algorithmic level* of analysis that concerns the representations and cognitive strategies that the system uses to approximate the optimal solution, and the implementation level that concerns how those representations and computational mechanisms are bio-physically realized

in the brain. While most rational models of human cognition are formulated at the computational level of analysis, the remainder of this chapter and most of the subsequent chapters are dedicated to showing that we can gainfully push normative principles down to the algorithmic level of analysis to better understand people's cognitive strategies and representations.

1.3.2 Rational analysis

Contrary to the bleak picture painted by research on heuristics and biases, other studies have shown that many aspects of human cognition can be understood as rational adaptations to the environment and the goals people pursue in it (Anderson, 1990; Chater & Oaksford, 1999). *Rational analysis* leverages this assumption to derive models of human behavior from the structure of the environment.

Concretely, Anderson (1990, p. 29) laid out the six-step methodology for developing rational models of cognition summarized in Figure 1.1. Rational analysis derives models of human behavior from the structure of the environment by assuming that cognitive mechanisms are near-optimally adapted to achieving their goals in people's natural environment. While cognitive psychology has traditionally explained human behavior primarily in terms of the structure of the mind and its capacity limits, rational analysis explains human behavior primarily in terms of the structure of the environment and makes only minimal assumptions about cognitive limitations. In the context of the debate on human rationality, rational analysis has been used to provide rational explanations for a wide range of cognitive biases including the confirmation bias (Austerweil & Griffiths, 2011; Oaksford & Chater, 1994), the representativeness heuristic (Griffiths & Tenenbaum, 2001; Tenenbaum et al., 2001), the gambler's fallacy (Hahn & Warren, 2009), and fallacious argumentation (Hahn & Oaksford, 2007).

1.3.3 Rational process models

The computational challenges posed by rationality are not just problems for human minds; they are also faced by computer scientists and statisticians who work with complex probabilistic models. These computer scientists and statisticians have developed a variety of strategies for approximating the resulting computations, and those strategies provide a source of hypotheses about cognitive processes that could be used to produce behavior that approximates a given rational model. The

1. Precisely specify what are the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted.
3. Make the minimal assumptions about computational limitations.
4. Derive the optimal behavioral function given items 1 through 3.
5. Examine the empirical literature to see if the predictions of the behavioral function are confirmed.
6. If the predictions are off, iterate.

Figure 1.1: The methodology of rational analysis.

result is what has been dubbed a "rational process model" (Griffiths, Vul, & Sanborn, 2012; Sanborn, Griffiths, & Navarro, 2010; Shi, Griffiths, Feldman, & Sanborn, 2010).

For example, Sanborn et al. (2010) showed how two algorithms commonly used to perform probabilistic inference – Markov chain Monte Carlo and particle filters – can be be reinterpreted as hypotheses about the psychological mechanisms of categorization. With a large sample, these algorithms give a close approximation to the ideal rational model, but with a small sample, they deviate from this ideal in systematic ways, producing biases (such as order effects) that are easy to compare against human performance. Such comparisons yield clues about the computational constraints that might be relevant for explaining human behavior.

Process models based on approximation algorithms (such as Monte Carlo methods) are rational in that their answers converge to the optimal solution in the limit of infinite computational resources. However, this is a weak form of rationality that corresponds to the notion of calculative rationality introduced above. One of the contributions of this dissertation will be to bolster the rationality of rational process models by grounding them in the theory of bounded optimality. This idea has led to the notion of resource-rationality and the methodology of resource-rational analysis presented below.

1.4 REDEFINING HUMAN RATIONALITY AS THE RATIONAL USE OF FINITE TIME AND LIMITED COGNITIVE RESOURCES

As reviewed above, research on human judgment and decision-making has established that people do not obey to the norms of logic, probability theory, and expected utility theory. The brain's

finite computational power limits how rational people can possibly act and think. As a result of this *bounded rationality*, the ideals of maximizing expected utility, reasoning according to the laws of logic, and handling uncertainty according to the laws of probability are out of reach for people. As Figure 1.2 illustrates, the realization that people's cognitive limitations rule out that people are optimal (according to the standard picture of rationality), is compatible with a large number of ways how the mind might work instead. So how should we think and decide instead? In my view, bounded optimality is the literature's most principled, general answer to the question of what it means to be rational. I will, therefore, instantiate this abstract notion as a concrete theory of human rationality. Concretely, I propose that to be rational a person has to reason and decide according to cognitive strategies that perform as well as or better than any other strategies that they could be using instead. I will refer to this new normative standard as *resource-rationality*.

As illustrated in Figure 1.2, resource-rationality uniquely identifies the best biologically feasible mind(s) out of the infinite set of bounded rational minds. Concretely, I define a resource-rational mind $m^\star$ for a brain B in an environment E with respect to the utility function u as

$$m^\star = \arg\max_{m \in M_B} \mathbb{E}_{P(T, l_T | E, A_t = m(l_t))} \left[u\left(l_T \right) \right], \tag{1.7}$$

where the agent's life history $l_t = (S_0, \cdots, S_t)$ is the sequence of states it has experienced up until time t, $u(l_T)$ measures how good this life was until it ended at time T, M_B is the set of minds that are biologically feasible given the biophysical constraints of its brain B, S_t is the state of the environment at time t, and $A_t = m(l_t)$ is the action that the mind m chooses in state s_t if the previous states were $s_0, \cdots, s_{t-1}$. The cognitive limitations inherent in the biologically feasible minds M_B include a limited set of elementary operations (e.g., counting and memory recall are available but exact Bayesian inference is not), a limited processing speed (each operation takes a certain amount of time), and potentially other constraints, such as limited working memory. Critically, the world state S_t is constantly changing while the mind m deliberates. Thus, to perform well, the bounded optimal mind $m^\star$ does not only have to generate good decisions but it also has to generate them quickly. Since each cognitive operation takes a certain amount of time, this entails that bounded optimality often requires computational frugality.

Unfortunately, it might be intractable to compute the resource-rational mind defined by Equation 1.7 because it requires optimizing over an entire lifetime. To provide a more tractable definition that can be used to derive predictions about which heuristic h a person should use to make a particular decision or inference, it will be assumed that life can be partitioned into a sequence of episodes
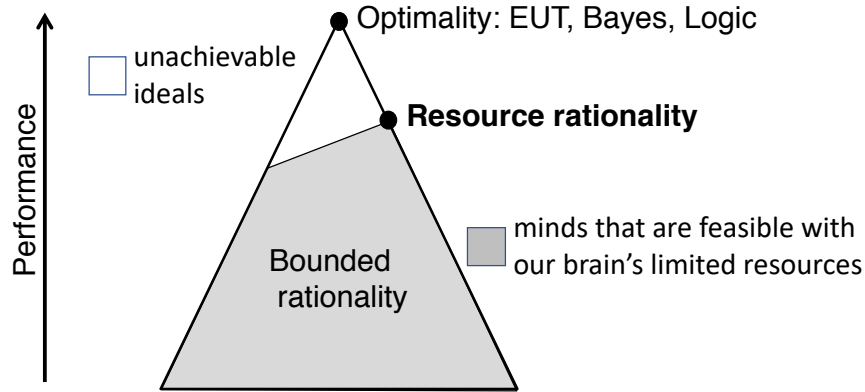
Figure 1.2: Resource rationality and its relationship to optimality and bounded rationality.

each of which starts with a state $s_0 = (w_0, b_0)$ that comprises the unknown state of the external world w_0 and the person's internal belief state b_0. Furthermore, let result(s_0, h) denote the judgment, decision, or belief update that results from applying heuristic h in the initial state s_0. In this setting, we can decompose the value of having applied a particular strategy into the utility of its termination state $u(s_\perp)$ and the computational cost of its execution. The latter is critical because the time and cognitive resources a person expends on any one decision or inference (current episode) take away from their budget for other decisions and inferences (future episodes). To capture this, let the random variable cost(t_h, ρ, λ) denote the total opportunity cost of investing the cognitive resources ρ used or blocked by the heuristic h for the duration t_h of its execution, when the agent's cognitive opportunity cost per quantum of cognitive resources and unit time is λ. In this setting, I define the resource-rational heuristic $h^\star$ for a brain B to use in the belief state b_0 as

$$h^\star(b_0, B, E) = \arg\max_{h \in H_B} \mathbb{E}_{P(\text{result}|s_0, h, E)} \left[u\left(\text{result}\right) \right] - \mathbb{E}_{t_h, \rho, \lambda | h, s_0, B, E} \left[\text{cost}(t_h, \rho, \lambda) \right], \quad (1.8)$$

where H_B is the set of heuristics the brain B can execute. The cost of thinking can be defined by

$$\text{cost}(d, \rho, \lambda) = \int_0^{t_h} \rho(t) \cdot \lambda(t) \, \mathrm{d}t. \quad (1.9)$$

For simplicity, I will assume that the heuristic's cognitive demands ρ and the agent's opportunity cost λ are roughly constant while the heuristic h is being executed. In this case, the cost of thinking can be approximated by cost$(t_h, \rho, \lambda) = t_h \cdot \rho \cdot \lambda$. To further simplify this analysis, $\rho \cdot \lambda$ can be approximated by the agent's reward rate in the environment E; this corresponds to the assumption

that a) the agent cannot multitask and b) the current reward rate is an accurate estimate of the value of the agent's time. In brief, the essence of resource-rationality is that people's cognitive mechanisms should trade off accuracy versus opportunity cost in an adaptive, near-optimal manner.

This definition improves upon what Lewis et al. (2014) call *ecological bounded optimality* in at least three major ways: First, it explicitly captures the opportunity costs of the time and computation that applying a strategy h to the current problem incurs at the expense of the agent's ability to solve other problems concurrently or in the future. Second, it weighs the states the environment might be in according to the person's belief state (b_0) rather than their overall frequency in the environment. This accounts for people's ability to adapt their cognitive strategy to individual problems based on their (imperfect) knowledge about the state of their environment (Payne, Bettman, & Johnson, 1993). Third, the utility function is allowed to depend on the belief state $b_\perp$ that results from reasoning according to h. This captures the potential benefits of belief updates achieved by reflecting in the current episode for decisions made in future episodes.

Another way in which the resource-rational approach advances the methodology of computational rationality is that it leverages ideas from computer science to generate hypotheses about the mind's computational architecture and the space of heuristics that it might support. This gives rise to a methodology for reverse-engineering the mind's cognitive strategies known as *resource-rational analysis* that will be presented in the following section.

Resource-rationality differs from the standard picture of rationality along three of the four dimensions: First, it evaluates reasoning by its utility for subsequent decisions rather than by its formal correctness; this makes it an instance of pragmatism. Second, it agrees with Tversky and Kahneman's approach (Tversky & Kahneman, 1974) in that resource-rationality is an attribute of the process that generates conclusions and decisions. Third, it takes into account the cost of time and the boundedness of people's cognitive resources. Fourth, resource-rationality is defined with respect to the distribution of problems in the environment rather than a set of arbitrary laboratory tasks. Arguably, all three changes are necessary to obtain a normative, yet realistic, theory of human rationality. Unlike the decision theoretic and Bayesian accounts, resource-rationality is not defined by the quality of the people's actions or the truthfulness or coherence of their beliefs, but rather, in terms of their cognitive strategies. Unlike logic and probability theory, it does not measure the quality of these strategies by their adherence to rules that preserve truth or coherence, but rather, by its practical effects on the people's actions and their consequences. Limited time and bounded cognitive resources necessitate tradeoffs. This amplifies the effect of resource-rationality's departures from the standard picture of rationality by its pragmatic perspective on reasoning and its emphasis of per-

1. Start with a computational-level (ie. functional) description of an aspect of cognition, formulated as a problem and its solution.
2. Posit a class of algorithms for approximately solving this problem, a cost to computational resources used by these algorithms, and a utility of more accurately approximating the correct solution.
3. Find the algorithm in this class that optimally trades off resources and approximation accuracy (Equation 1.8).
4. Refine by revising the model, algorithms, or costs (Steps 1, 2, or 3), or by proceeding to the next level down: approximating the algorithms in Step 2 to capture further resource constraints.

Figure 1.3: Recipe of resource-rational analysis.

formance in the real world over performance in the laboratory. The following chapters illustrate that this allows resource-rationality to accommodate cognitive biases that were previously deemed irrational.

1.5 Resource-rational analysis

One of the principles of rational analysis is to make only minimal assumptions about cognitive constraints (see Figure 1.1). But the constraints imposed by people's cognitive limitations are often substantial. Herbert Simon has famously argued that to understand people's cognitive strategies, we have to simultaneously consider people's cognitive constraints and the structure of their environment (Simon, 1956, 1982). To achieve this, *resource-rational analysis* (Griffiths, Lieder, & Goodman, 2015) incorporates cognitive constraints into rational analysis. Concretely, it takes into account which cognitive operations are available to people, how long they take, and how costly they are.

Resource-rational analysis is a four-step methodology (see Figure 1.3) that leverages the theory of resource-rationality introduced above to derive process models of cognitive abilities from formal definitions of their function and assumptions about the mind's computational architecture. This function-first approach starts at the computational level of analysis (Marr, 1982). When the problem solved by the cognitive capacity under study has been formalized, resource-rational analysis postulates an abstract computational architecture, that is a set of elementary operations and their costs, with which the mind might solve this problem. Next, a resource-rational analysis derives the algorithm that is optimal for solving the problem identified at the computational level with the abstract
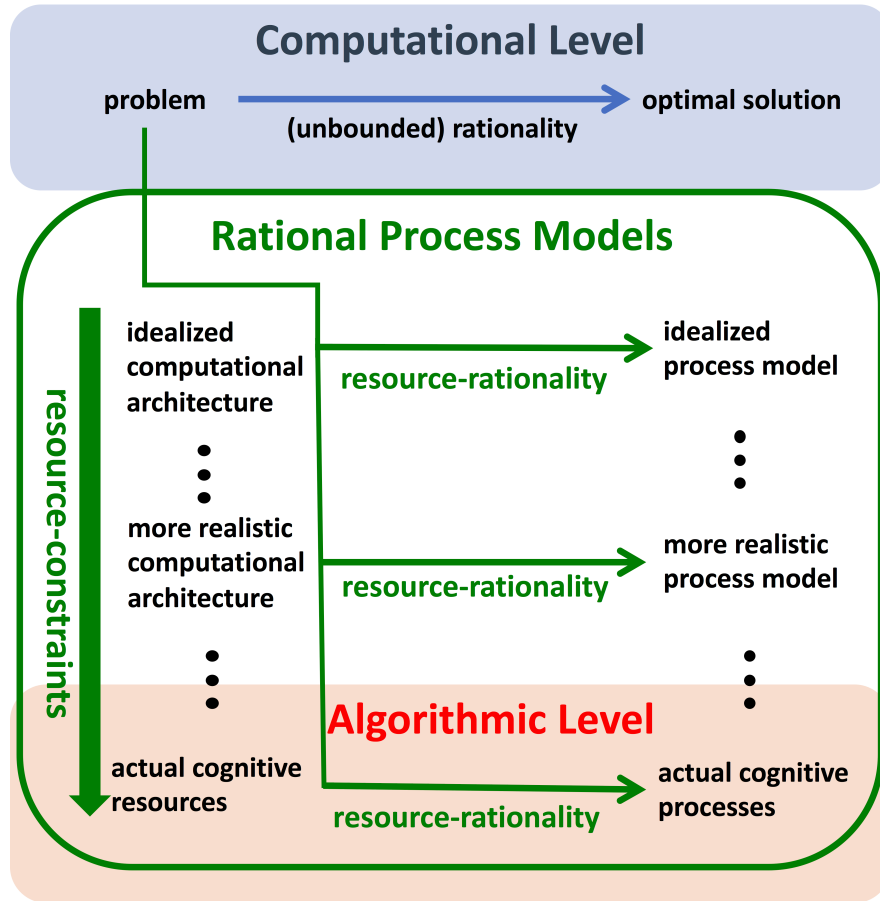
Figure 1.4: Illustration of how resource-rational analysis connects levels of analysis.

computational architecture (Equation 1.8). The resulting process model is used to predict people's responses and reaction times in a given experiment and those predictions are then tested against empirical data. Based on this evaluation, the assumptions about the computational architecture and the problems to be solved are revised and the analysis cycle is repeated (see Figure 1.5). The iterative refinements of the assumed cognitive architecture proceeds from abstract, minimal assumptions to an increasingly more realistic model of the underlying neuro-cognitive architecture (see Figure 1.4). In this way, resource-rational analysis can be used to connect Marr's levels of analysis (Marr, 1982).

By explicitly positing a class of possible algorithms and a cost to the resources used by these algorithms, we can invoke an optimality principle to derive the algorithm that the mind should be using. This makes resource-rational analysis a methodology for analyzing information-processing
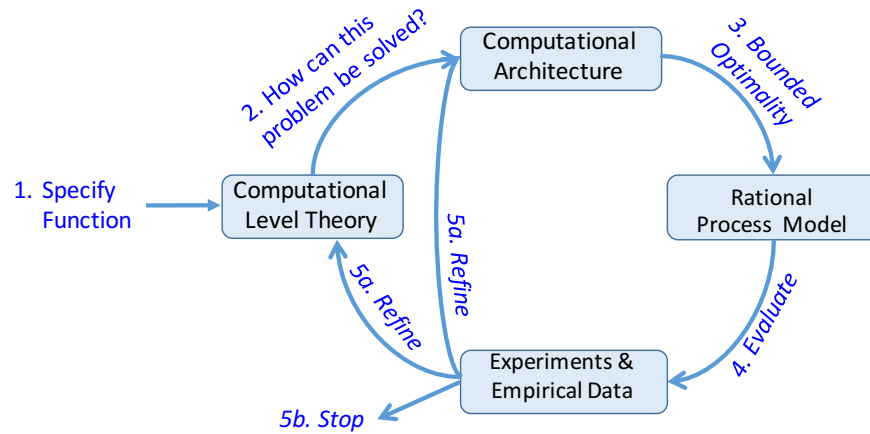
Figure 1.5: Illustration of the resource-rational analysis cycle.

systems at an intermediate level defined by an idealized family of computational mechanisms which corresponds to a particular computational architecture. This method enables us to reverse-engineer not only the problem that a system solves (computational level of analysis), but also, the system's computational architecture.

To identify a family of potential cognitive strategies and the corresponding cognitive architecture (Step 2), resource-rational analysis draws on previous research in the computational sciences. Concretely, having formulated the problem to be solved in precise mathematical terms allows us to mine the literature of artificial intelligence, machine learning, operations research, and other areas of the computer science and statistics for classes of algorithms that have been developed to efficiently solve such problems. Such a literature search generally yields one or more parametric families of algorithms. Different settings of algorithm's parameters often produce qualitatively different behaviors and different speed-accuracy trade-offs. For instance, particle filtering is a general approach that leads to specific algorithms varying in the number of particles, the re-sampling criteria, and so on (Abbott & Griffiths, 2011). This results in an infinite collection of algorithms some of which have qualitatively different properties (e.g., one particle vs. millions of particles). Steps 2 and 3 allow us to find reasonable points within this space of algorithms, which can then be compared to human behavior. To the degree that evolution, development, and learning have adapted the system to make optimal use of its finite computational resources, resource-rational analysis can be used to derive the system's algorithm from assumptions about its computational architecture.

1.6 Resource-rationality as an organizing principle for understanding human cognition

The general principle that human cognition is optimal subject to computational constraints has been successfully instantiated in previous models of decision-making, perception, memory, attention, reasoning, and cognitive control (Gershman et al., 2015; Lewis et al., 2014; Shenhav et al., 2017). In this section, I review this literature to highlight the potential of resource-rational analysis and opportunities for future work.

1.6.1 Resource-rational models of decision-making

Models of bounded-optimal decision-making differ widely in their assumptions about the nature, costs, and limits of the cognitive operations they assume to be available to the decision-maker and in which aspects of decision-making are assumed to be bounded-optimal.

COSTLY INFORMATION ACQUISITION. At a minimum, the models reviewed here assume that it is costly to acquire information while retaining the assumption that acquired information will be processed optimally (Caplin, Dean, & Martin, 2011; Colombo, Femminis, & Pavan, 2014; Gabaix, Laibson, Moloche, & Weinberg, 2006; Lieder, Krueger, & Griffiths, 2017; Reis, 2006; Verrecchia, 1982). By leveraging the principle of bounded-optimality, these studies were able to show that previously proposed heuristics can be resource-rational: Caplin et al. (2011) derived a bounded-optimal version of Herbert Simon's classic satisficing heuristic (Simon, 1956). Similarly, Lieder, Krueger, and Griffiths (2017) found that the Take-The-Best heuristic might be bounded-optimal when the stakes are low and one outcome is much more probable than all the others. Evaluating models of optimal decision-making with information costs in experiments has revealed that human performance is constrained by additional limitations, such as limited working memory (Sanjurjo, 2017) and limited information about the statistics of the decision environment that necessitate exploration (Caplin et al., 2011).

NEURAL NOISE. Other studies have assumed that decision-making is constrained by neural noise corrupting the fidelity of internal representations (Bhui & Gershman, 2017; Howes, Warren, Farmer, El-Deredy, & Lewis, 2016; Khaw, Li, & Woodford, 2017; Summerfield & Tsetsos, 2015). Khaw et al. (2017) that risk aversion follows from Bayesian inference of the gambles' expected values from

a psychophysically plausible noisy representation of their payoffs. Similarly, Howes et al. (2016) showed that contextual preference reversals can be explained as the consequence of an optimal inference of value from a noisy representation of the alternatives' attributes. Bhui and Gershman (2017) point out that the neural noise assumed by these models can be understood as a consequence of bounded-optimal neural coding under metabolic constraints. Assuming that the fidelity of neural representations is constrained, how should value be represented? Bhui and Gershman (2017) present an information theoretic argument for the idea that it is bounded-optimal for the brain to represent utilities and probabilities by their smoothed rank. Their analysis provides a rational justification for the core assumptions of the decision-by-sampling model (N. Stewart, 2009; N. Stewart, Chater, & Brown, 2006) and extends it in a way that explains additional biases in decision-making (range effects and certain context effects).

BOUNDED-OPTIMAL EVIDENCE ACCUMULATION.   A large number of studies have applied bounded-optimality to different components of the drift-diffusion model of decision-making (Gold & Shadlen, 2007): the decision threshold (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Bogacz, Hu, Holmes, & Cohen, 2010; Fudenberg, Strack, & Strzalecki, 2018; Gabaix & Laibson, 2005; Tajima, Drugowitsch, & Pouget, 2016; Vul, Goodman, Griffiths, & Tenenbaum, 2014), evidence generation (Dickhaut, Rustichini, & Smith, 2009; Woodford, 2014, 2016), and evidence accumulation (Bogacz et al., 2006; Tsetsos et al., 2016). While these models were first fully developed for the domain of perceptual decision-making (Bogacz et al., 2006, 2010) they have also been extended to value-based choice (Dickhaut et al., 2009; Fudenberg et al., 2018; Gabaix & Laibson, 2005; Tajima et al., 2016; Tsetsos et al., 2016; Woodford, 2014, 2016). The resulting models offer a resource-rational reinterpretation for phenomena that were previously considered irrational, including intransitive preferences (Tsetsos et al., 2016) and probability matching (Vul et al., 2014). Other studies have found systematic deviations of human performance from the predictions of the optimal drift-diffusion model of perceptual decision-making (Holmes & Cohen, 2014). For instance, Bogacz et al. (2010) found that most people under-perform the optimal speed-accuracy tradeoff by setting their decision threshold too high leading to a heightened accuracy at the expense of responding too slowly. This finding suggests that human performance is constrained by additional bounds that limit the accuracy of their time estimates and that make it costly for them to adjust the decision threshold through cognitive control (Holmes & Cohen, 2014).

LIMITED ATTENTION.   Other studies have applied bounded optimality principles to model the effect of limited attention resources (Caplin & Dean, 2015; Caplin, Dean, & Leahy, 2017; Gabaix, 2014; Lieder, Griffiths, & Hsu, 2017; C. A. Sims, 2003, 2006; Woodford, 2012). The starting point of this line of inquiry was Sim's theory of *rational inattention* (C. A. Sims, 2003, 2006). This theory models people's limited attention in terms of an information theoretic constraint on the mutual information between the state of the environment and the bounded agent's behavior. This model has been able to successfully explain several apparent irrationalities in economic behavior, including the inertia, randomness, and abruptness of the reactions of decision-makers to new financial information.

One limitation of the rational inattention model is that it discounts all information equally whereas people tend to focus their attention on a small number of important pieces of information while neglecting others completely. Gabaix (2014) addressed this limitation by deriving an optimal attention function that selects which variables the decision-maker should attend to and how much attention each of the selected variables should receive. Gabaix (2014) argues that his *sparsemax* model can be used as a more psychologically plausible foundation for elements of micro-economic theory and shows that its predictions deviate from standard micro-economic theory which assumes perfectly rational agents in similar ways as human behavior. A second limitation of the rational inattention model is that it makes very specific assumptions about the cost of attention whose predictions where not borne out by subsequent experiments (Caplin & Dean, 2013; Caplin et al., 2017; Dean & Neligh, 2017). Subsequent models addressed this problem by generalizing the attention cost function in ways that make it possible to reconcile these deviations with bounded-optimal decision-making under limited attention resources (Caplin & Dean, 2015; Caplin et al., 2017). This illustrates that resource-rational analysis can be used to reverse-engineer what the cognitive constraints on human decision-making might be. A third limitation of the rational inattention model is that it abstracts away from the cognitive processes of decision-making. I will address this limitation by developing a resource-rational process model of decision-making with limited attention resources in Chapter 3.

COMPUTATIONAL COMPLEXITY.   The models discussed so far assumed that human decision-making is bounded-optimal, subject to the constraints imposed by incomplete information, neural noise, and limited attention. By contrast, Beck, Ma, Pitkow, Latham, and Pouget (2012) argue that the relatively levels of neural noise measured neuro-physiologically are not nearly high enough to fully explain the variability and suboptimality of human performance. They propose that instead of making optimal use of noisy representations, the brain uses approximations that entail systematic

biases. Such approximations appear to be warranted by the intractable computational complexity of decision-making (Bossaerts & Murawski, 2017). The notions of Type II rationality and rational metareasoning introduced above seek to address this problem by terminating deliberation as soon as the expected improvement of decision quality that can be achieved by performing another computation drops below the cost of computation. Unfortunately, performing this cost-benefit analysis is itself intractable. The *directed cognition* model by Gabaix and Laibson (2005); Gabaix et al. (2006) addresses this concern by selecting computations according to a myopic cost-benefit analysis that looks only a single step ahead. They show that their model predicts people's decisions in certain scenarios more accurately than expected utility theory and predicts qualitative properties of how people choose between multiple complex goods (Gabaix et al., 2006). However, there are many ways in which people violate the prediction of this model (Sanjurjo, 2017; Yang, Toubia, & De Jong, 2015). Resource-rational analysis might be able to address these limitations, and recent study found that a bounded-optimal model of planning explained human performance significantly better than the directed cognition model (Callaway et al., 2018).

Ideas of bounded optimality have been applied to model how people leverage their habits to eschew or reduce the computational challenges of planning. Huys et al. (2015) employed an optimal fragmentation model according to which people decompose sequential decision problems into sub-problems so as to optimally trade-off savings in the cost of planning attained by reusing previous action sequences with the resulting decrease in decision quality. This model helped them to gain insights into how people combine different heuristics to efficiently solve complex planning problems. Recent work has provided additional empirical evidence for the view that people adaptively leverage their habit system to simplify planning (Keramati, Smittenaar, Dolan, & Dayan, 2016). Beyond that, there appears to be very little, if any, research that has applied the principle of bounded-optimality to understand the simple heuristics people use to solve complex problems (Gigerenzer, 2008a; Gigerenzer & Selten, 2002). While it is commonly assumed that a boundedly rational decision-maker would rely on heuristics (Bossaerts & Murawski, 2017; Gigerenzer, 2008a; Gigerenzer & Selten, 2002), there used to be no theory for deriving bounded-optimal heuristics. To address this problem, I developed the theory of resource-rationality presented above (Equation 1.8). Chapters 2-3 apply this framework to derive resource-rational heuristics, and Chapter 6 presents a computational method for discovering rational heuristics automatically and applies it to multi-alternative risky-choice and planning.

1.6.2 RESOURCE-RATIONAL MODELS OF MEMORY

Human memory is fundamentally constrained by the fact that some cognitive resources which are critical for human performance, such as working memory, are very limited (Miller, 1956). The combination of limited resources and adaptive pressures suggests that resource-rational analysis might be particularly useful for understanding the memory mechanisms. In addition to Anderson's famous rational analysis of memory storage and retrieval (Anderson & Milson, 1989), recent work has applied the principles of bounded optimality to understand the working memory mechanisms governing people how many items are committed to working memory (Howes et al., 2016), how those items will be encoded (Orhan, Sims, Jacobs, & Knill, 2014; C. R. Sims, 2015, 2016; C. R. Sims, Jacobs, & Knill, 2012; van den Berg & Ma, 2017), and how their memories will maintained (Suchow, 2014; Suchow & Griffiths, 2016).

Anderson and Milson's rational analysis of memory (Anderson & Milson, 1989) can be interpreted as the first application of the principle of bounded optimality to understanding the human mind. It assumed that memory retrieval is bounded by the computational constraints of time and effort and that retrieving relevant information from memory requires searching through a list of potentially relevant memories until one either finds a relevant one or gives up the search. Given these computational constraints, Anderson and Milson (1989) derived an optimal memory storage mechanism that exploits the statistical structure of the environment to sort the memories in the order of the probability that they will be needed and a stopping rule that terminates the search when its expected gain drops below its cost. The resulting bounded-optimal memory mechanisms correctly predict the effects of frequency, recency, and spacing on the how accurate people were at recalling information and how long it took them to do so. Anderson and Schooler (1991) followed up on this analysis by showing that the frequency with which a previously encountered piece of information will be needed again in people's natural environment does indeed possess the statistical structure that makes the empirically observed memory mechanisms bounded-optimal.

The work by C. R. Sims et al. (2012) and van den Berg and Ma (2017) illustrates that developing and evaluating bounded-optimal models is a promising way to reverse-engineer the constraints that limit people's cognitive performance. The bounded optimality framework allowed C. R. Sims et al. (2012) to derive the effects of different kinds of capacity limits and test the resulting predictions against empirical data. This allowed them to infer that rather than being constrained to a fixed number of items, visual working memory is a more continuous resource that can be flexibly divided to either maintain a small number of items with high fidelity or a larger number of items with lower

fidelity. Furthermore, their bounded optimal model predicts that how information is encoded in working memory should depend on the statistics of the input distribution, the nature of the task, and the relative costs of different kinds of errors (C. R. Sims, 2016; C. R. Sims et al., 2012). This allowed their model to correctly predict how the precision with which items are encoded in working memory depends on task characteristics such as the number of items and the variability of their features. These contingencies challenge the capacity limits inferred from previous working memory studies that have used artificial stimuli and suggest that more naturalistic stimuli might reveal the capacity of human working memory to be less limited than it seems (Orhan et al., 2014). Going one step further, van den Berg and Ma (2017) challenged the ingrained assumption that working memory always distributes a fixed amount of representational resources among the encoded items. Instead, their model expresses the assumption that the total amount of working memory resources invested at any given time is chosen according to a rational cost-benefit analysis that trades-off the expected behavioral performance against the neural/metabolic cost of active memory maintenance. They show that the resulting model provides a better and more principled explanation of how working memory performance depends on the number of items to be remembered. Finally, the work by C. R. Sims (2015) illustrates that resource-rational analysis (see Figure 1.5) can be used to reverse-engineer not only the capacity limitations of working memory, but also, its implicit goals.

1.6.3 Resource-rational models of perception and neural coding

Many previous studies have successfully modeled perception as Bayesian inference (Kersten, Mamassian, & Yuille, 2004; Knill & Pouget, 2004; Knill & Richards, 1996; Lee & Mumford, 2003; Marr, 1982; Yuille & Kersten, 2006).The observation that rational models have been most successful in the domain of perception might reflect that our perceptual systems have been under direct evolutionary pressure for a very long time and have been equipped with considerable neural resources. However, perception is also an intractably difficult problem (Tsotsos, 1988). So, it would be very surprising if it was not also shaped by computational constraints.

Recent work has shown that there are indeed systematic deviations of human perception from Bayesian inference that can be understood in terms of bounded optimality (C. R. Sims, 2016; Stocker, Simoncelli, & Hughes, 2006; Wei & Stocker, 2015, 2017). The principle of bounded optimality has also been invoked to elucidate the underlying neural mechanisms (Lennie, 2003; Levy & Baxter, 1996; Olshausen & Field, 2004; Z. Wang, Wei, Stocker, & Lee, 2016). Stocker et al. (2006) found that the biases and variability of people's judgments of the speed of visual motion were consistent

with the Bayes-optimal use of a noisy internal representation of the sensory evidence. Wei and Stocker (2015,2017) proposed that the limited fidelity of these representations arises from the necessity to distribute finite neural resources across all possible percepts. They proceed to show that the optimal allocation of these limited resources according to natural image statistics can explain why people's orientation judgments are sometimes biased away from their prior expectation rather than towards it. It also correctly predicted a lawful relationship between the perceptual discrimination of a particular stimulus value, say orientation, and the amount of bias in people's perception of it (Wei & Stocker, 2017). Another recent resource-rational analysis of perception (C. R. Sims, 2016) has emphasized that bounded-optimal perceptual representations are also shaped by the fact that certain perceptual errors (e.g., confusing a poisonous mushroom for an edible one) are more costly than others (e.g., confusing two poisonous mushrooms). One payoff of this approach is that it makes it possible to infer people's cost function (which can be interpreted as specifying the goal of perception) from their perceptual performance.

Furthermore, the principle of bounded optimality can also be applied to the neural implementation of perceptual mechanisms. This approach makes it possible to ground the assumed capacity limitations of resource-rational models in biophysical constraints that can be measured independently. One of these constraints is metabolic energy. In fact, action potentials are so metabolically expensive that at most 1% of all neurons in the brain can sustain substantial activity in parallel (Lennie, 2003). This bound imposes serious constraints on neural coding and computation. Indeed, many aspects of morphology, physiology, and wiring of neural circuits can be understood as the adaptation to the evolutionary pressure to achieve a near-optimal trade-off between the computational efficacy and metabolic cost (Levy & Baxter, 1996, 2002; Niven & Laughlin, 2008; Sterling & Laughlin, 2015). This principle can, in turn, be used to mathematically derive optimal neural codes that respect biological constraints (Levy & Baxter, 1996). For instance, Z. Wang et al. (2016) derived some of the visual system's neural codes by maximizing the mutual information between the neural representation and the sensory input subject to metabolic constraints and limited fidelity caused by neural noise. Furthermore, the principle of sparse coding (Olshausen & Field, 1996), which has been highly successful in explaining the receptive fields of sensory neurons (Olshausen & Field, 1997, 2004), can be interpreted as a bounded optimal solution to the problem of accurately representing the environment subject to the constraint that only a very small fraction of all neurons can be active simultaneously. Finally, the effects of metabolic constraints are not restricted to the details of the neural implementation but propagate all the way up to high-level cognition by necessitating cognitive mechanisms like selective attention (Lennie, 2003).

1.7 Conclusion and outlook

The successes of resource-rational analysis summarized above suggest that resource-rationality is a promising theoretical framework for understanding bounded rationality. As these examples illustrate, resource-rational analysis has a number of benefits:

1. The resource-rational perspective provides an overarching principle from which we can derive models of human cognition that are both mathematically precise and psychologically plausible.

2. Resource-rationality can, therefore, be used to develop a theoretical foundation for the economic sciences that is substantially more realistic than expected utility theory.

3. Resource-rationality provides a unifying explanation for a wide range of seemingly unrelated phenomena.

4. Resource-rational models allow us to make sense of cognitive biases.

5. Resource-rationality provides a realistic normative standard against which human behavior can be evaluated to identify genuine sub-optimalities.

6. Resource-rational analysis can be used to reverse-engineer cognitive limitations and to infer a cognitive system's implicit goals from errors in its performance.

The fact that some "cognitive biases" were found to be compatible with the principles of bounded optimality suggests that it is time to re-evaluate human rationality against this more realistic normative standard. Despite the considerable progress summarized above, many questions remain to be answered. In particular, it remains unclear whether classic cognitive biases, such as anchoring and availability, that have been instrumental to the conclusion that people are irrational are compatible with the principles of resource-rational information processing or not. Furthermore, almost all of the resource-rational analyses reported above were restricted to optimizing single parameters of cognitive mechanisms, resource-allocation, or abstractly characterizing representations. But there is still no principled way to derive resource-rational cognitive strategies, and it remains unclear whether and under which conditions the heuristics advocated by proponents of ecological rationality are resource-rational. Furthermore, while bounded optimality has already been applied to answer descriptive and normative questions about the human mind, it has yet to be translated into useful prescriptive theories that can be used to improve human cognition through training, instruction, or technology.

To address these gaps in our knowledge, the remaining chapters of Part 1 revisit the debate about human rationality by applying resource-rational analysis to two cognitive biases that have been instrumental to the conclusion that people are fundamentally irrational: the anchoring bias (Chapter 2) and the availability bias (Chapter 3). Furthermore, my dissertation illustrates how resource-rationality can be leveraged to elucidate the cognitive mechanisms of judgment, decision-making, planning, and strategy selection. To facilitate these efforts, Chapter 6 introduces a computational method for deriving bounded-optimal strategies automatically (Step 3 of resource-rational analysis; see Figure 1.3). In Part 2 I apply this method and insights into people's bounded rationality towards overcoming cognitive limitations and improving decision-making. As part of these efforts, I develop a cognitive tutor for boosting people's decision-making competence (Hertwig & Grüne-Yanoff, 2017) with automatically discovered resource-rational heuristics.

2

A resource-rational perspective on anchoring-and-adjustment[*]

Achieving demanding goals in limited time requires balancing being quick and being accurate. We regret the opportunities we miss when we fail to make up our mind on time, but we also regret the errors we commit by jumping to conclusions. When we think too little our judgments can be skewed by irrelevant information that we happened to see, hear, or think about a moment ago. This phenomenon is known as *anchoring*. Anchoring is one of the cognitive biases discovered by Tversky and Kahneman (1974) and played an important role in the debate about human rationality. It impacts many important aspects of our lives including the outcome of salary negotiations (Galinsky & Mussweiler, 2001), economic decisions (e.g., Simonson & Drolet, 2004), criminal sentences (Englich, Mussweiler, & Strack, 2006), and even our ability to understand other people (Epley, Keysar, Van Boven, & Gilovich, 2004).

In their classic paper, Tversky and Kahneman (1974) showed that people's judgments could be systematically skewed by providing them with an arbitrary number before their judgment: The experimenter generated a random number by spinning a wheel of fortune, and then asked participants

---

[*]This chapter is based on Lieder, Griffiths, Huys, and Goodman (2018a), Lieder, Griffiths, Huys, and Goodman (2018b), and Lieder, Griffiths, Huys, and Goodman (2017).

to judge whether the percentage of African countries in the United Nations was smaller or larger than that number. Participants were then asked to estimate this unknown quantity. Strikingly, the participants' estimates were biased towards the random number: their median estimate was larger when the random number was high than when it was low. This appears to be a clear violation of rationality. According to Tversky and Kahneman (1974) this violation occurs because people use a two-stage process called *anchoring-and-adjustment* (see also Nisbett & Ross, 1980). In the first stage, people generate a preliminary judgment called their *anchor*. In the second stage, they adjust that judgment to incorporate additional information, but the adjustment is usually insufficient.

In Tversky and Kahneman's experiment people appear to have anchored on the random number provided by the experimenter and adjusted it insufficiently. Consequently, when the anchor was low people's judgments were too low, and when the anchor was high then their judgments were too high.

At first sight, anchoring appears to be irrational, because it deviates from the standards of logic and probability which are typically used to assess rationality. But it could also be a reasonable compromise between error in judgment and the cost of computation, and hence be resource-rational. Anchoring-and-adjustment has two components that could be irrational: the generation of the anchor and the process by which it is adjusted. Previous research found that when no anchor is provided, the anchors that people generate for themselves are relevant quantities that are reasonably close to the correct value and can be generated quickly (Epley & Gilovich, 2006). Furthermore, research on human communication suggests that in everyday life it is reasonable to assume that other people are cooperative and provide relevant information (N. Schwarz, 2014). Applied to anchoring, this means that if somebody asks you in real life whether a quantity you know very little about is larger or smaller than a certain value, it would be rational to treat that question as a clue to its value (Zhang & Schwarz, 2013). Thus, having the queried value in mind might make it rational to reuse it as your anchor for estimating the unknown quantity. This suggests that the mechanism by which people generate their anchors could be rational in the real world.[†] If this is true, then the rationality of anchoring-and-adjustment hinges on the question of whether adjustment is a rational process. To answer this question, we investigate whether insufficient adjustment can be understood as a rational tradeoff between time and accuracy. If so, then how much people adjust their initial estimate should adapt rationally to the relative utility of being fast versus being accurate. To formalize this hypothesis, we present a resource-rational analysis of numerical estimation. We then leverage the predictions of this analysis to experimentally test our hypothesis that adjustment is rational. Our analysis

---

[†]We will revisit this issue in more depth in the general discussion.

suggested that the rational use of finite resources correctly predicts the anchoring bias and how it changes with various experimental manipulations (see Table 2.1). Our rational account makes the novel prediction that opportunity cost increases the anchoring bias and decreases reaction time regardless of whether the anchor is provided or self-generated. We tested these predictions in two controlled experiments where participants estimate numerical quantities under four different combinations of time cost and error cost. The experiments confirmed our theory's predictions and provided strong support for our rational process model of adjustment over alternative, less rational models of anchoring. All of these results support the conclusion that adjustment is resource-rational.

The remainder of this chapter begins with a brief survey of empirical findings on anchoring and discusses the challenges that they pose to existing accounts of anchoring-and-adjustment. We then present our resource-rational analysis of numerical estimation, a rational process model that can be interpreted in terms of anchoring-and-adjustment, and a series of simulations demonstrating that this model is sufficient to explain the reviewed phenomena. This motivates our two experiments, which we present in turn. We close by discussing our findings and their implications.

2.1 Empirical findings on the anchoring bias

Anchoring is typically studied in numerical estimation tasks. Numerical estimation involves making an informed guess of the value of an unknown numerical quantity. Since the first anchoring experiment by Tversky and Kahneman (1974) a substantial number of studies have investigated when anchoring occurs and what determines the magnitude of the anchoring bias (see Table 2.1).

The anchors that people use when forming estimates can be relevant to the quantity they are estimating. For instance, Tversky and Kahneman (1974) found that people sometimes anchor on the result of calculating $1 \times 2 \times 3 \times 4$ when the task is estimating $1 \times 2 \times 3 \times 4 \times \cdots \times 8$. However, people can also be misled, anchoring on numbers that are irrelevant to the subsequent judgment. For instance, many anchoring experiments first ask their participants whether an unknown quantity is larger or smaller than a given value and then proceed to have them estimate that quantity. Having compared the unknown quantity to the value provided by the experimenter makes people re-use that value as their anchor in the subsequent estimation task. Those numbers are therefore known as *provided anchors*. Importantly this procedure works with irrelevant numbers such as the random number that Tversky and Kahneman (1974) generated for their participants or one's own social security number (Ariely, Loewenstein, & Prelec, 2003).

Although asking people to compare the quantity to a given number is particularly effective, the anchoring bias also occurs when anchors are presented incidentally (Wilson, Houston, Etling, & Brekke, 1996), although this effect is smaller and depends on particulars of the anchor and its presentation (Brewer & Chapman, 2002). Furthermore, anchoring-and-adjustment can also occur without an externally provided anchor: At least in some cases people appear to generate their own anchor and adjust from it (Epley & Gilovich, 2004). For instance, when Americans are asked to estimate the boiling point of water on Mount Everest they often recall 212°F (100°C) and adjust downwards to accommodate the lower air pressure in higher altitudes.

Although people's adjustments are usually insufficient, various factors influence their size and consequently the magnitude of the anchoring bias. For instance, the anchoring bias is larger the more uncertain people are about the quantity to be estimated (Jacowitz & Kahneman, 1995). Indeed, Wilson et al. (1996) found that people knowledgeable about the quantity to be estimated were immune to the anchoring bias whereas less knowledgeable people were susceptible to it. While familiarity (Wright & Anderson, 1989) and expertise (Northcraft & Neale, 1987) do not abolish anchoring, expertise appears to at least reduce it (Northcraft & Neale, 1987). Other experiments have systematically varied the distance from the anchor to the correct value. Their results suggested that the magnitude of the anchoring bias initially increases with the distance from the anchor to the correct value (Russo & Schoemaker, 1989). Yet this linear increase of the anchoring bias does not continue indefinitely. Chapman and Johnson (1994) found that increasing an already unrealistically large anchor increases the anchoring bias less than increasing a realistic anchor by the same amount.

Critically for the resource-rational account proposed here, the computational resources available to people also seem to influence their answers. Time pressure, cognitive load, and alcohol decrease the size of people's adjustments and inter-individual differences in how much people adjust their initial estimate correlate with relevant personality traits such as the need for cognition (Epley & Gilovich, 2006). In addition to effects related to cognitive resources, adjustment also depends on incentives. Intuitively, accuracy motivation should increase the size of people's adjustments and therefore decrease the anchoring bias. Interestingly, experiments have found that accuracy motivation decreases the anchoring bias only in some cases, but not in others (Epley & Gilovich, 2006; Simmons, LeBoeuf, & Nelson, 2010). On questions where people generated their own anchors, financial incentives increased adjustment and reduced the anchoring bias (Epley & Gilovich, 2006; Simmons et al., 2010). But on questions with provided anchors, financial incentives have typically failed to eliminate or reduce the anchoring bias (Ariely et al., 2003; Tversky & Kahneman, 1974) with some exceptions (Wright & Anderson, 1989). A recent set of experiments by Simmons et al. (2010) suggested that

accuracy motivation increases adjustment from provided and self-generated anchors if and only if people know in which direction to adjust. Taken together, these findings suggests that the anchoring bias depends on how much cognitive resources people are able to and willing to invest.

Before the experiments by Simmons et al. (2010) demonstrated that accuracy motivation can increase adjustment from provided anchors, the bias towards provided anchors appeared immutable by financial incentives (Chapman & Johnson, 2002; Tversky & Kahneman, 1974; Wilson et al., 1996), forewarnings and time pressure (Mussweiler and Strack, 1999; but see Wright and Anderson, 1989). Since incentives were assumed to increase adjustment and increased adjustment should reduce the anchoring bias, the ineffectiveness of incentives led to the conclusion that the anchoring bias results from a mechanism other than anchoring-and-adjustment, such as selective accessibility (Chapman & Johnson, 2002; Epley, 2004; Mussweiler & Strack, 1999). Later experiments found that when people generate the anchor themselves accuracy motivation and time pressure are effective (Epley & Gilovich, 2005, 2006; Epley et al., 2004). This led Epley and Gilovich (2006) to conclude that people use the anchoring-and-adjustment strategy only when they generated the anchor themselves whereas provided anchors bias judgments through a different mechanism.

The wide range of empirical phenomena summarized in Table 2.1 have suggested a correspondingly wide range of explanations, including the idea that anchoring and adjustment is not a simple, unitary process. The remainder of the chapter explores an alternative account, showing that these disparate and seemingly inconsistent phenomena can all be explained by a unifying principle: the rational use of finite time and cognitive resources. From this principle we derive a resource-rational anchoring-and-adjustment model and show that it is sufficient to explain the anchoring bias regardless of whether the anchor was provided or self-generated.

2.2   Anchoring and Adjustment as Resource-Rational Inference

In this section we formalize the problem people solve in anchoring experiments – numerical estimation – and analyze how it can be efficiently solved in finite time with bounded cognitive resources. We thereby derive a resource-rational model of anchoring-and-adjustment. We then use this model to explain a wide range of anchoring phenomena.

Conceptually, our model assumes that adjustment proceeds by repeatedly considering small changes to the current estimate. The proposed change is accepted or rejected probabilistically such that the change is more likely to be made the more probable the new value is and the less probable

the current one is (see Figure 2.1). After sufficiently many adjustments the estimate becomes correct on average and independent of the initial guess. However, each small adjustment costs a certain amount of time. According to our model, the number of steps is chosen to minimize the expected value of the time cost of adjustment plus the error cost of the resulting estimate. In the remainder of this section, we derive our model from first principles, specify it in detail, and show that the optimal number of adjustments is very small. As Figure 2.1 illustrates, this causes the final estimates to be biased towards their respective anchors.

In contrast to previous theories of anchoring (Epley & Gilovich, 2006; Simmons et al., 2010), our model precisely specifies the number, size, and direction of adjustments as a function of the task's incentives and the participant's knowledge. In contrast, to the proposal by Epley and Gilovich (2006) our model covers adjustments from provided anchors *and* self-generated anchors. Furthermore, while Epley and Gilovich (2006) assumed that the correct direction of adjustment is known, our model does not make this assumption and allows the direction of adjustment to change from one step to the next. The model by Simmons et al. (2010) also makes these conceptual assumptions. However, it does not specify precisely how the direction and size of each adjustment are determined. While their model predicts a deterministic back-and-forth in the face of uncertainty, our model assumes that adjustments that improve the estimate are probabilistically preferred to adjustments that do not. This enables our model to capture streaks of adjustments in the correct direction interrupted by small steps in the wrong direction, whereas the model by Simmons et al. (2010) appears to predict that the direction of adjustment should constantly alternate. Finally, while both previous models assumed that adjustment stops as soon as the current estimate is sufficiently plausible (Epley & Gilovich, 2006; Simmons et al., 2010), we propose that the number of adjustments is predetermined adaptively to achieve an optimal speed-accuracy tradeoff on average. In the subsequent section we apply the resulting model to explain the various anchoring phenomena summarized in Table 2.1, and after that we will empirically test its predictions against the predictions of alternative models of adjustment including the stopping rule assumed by Epley and Gilovich (2006) and Simmons et al. (2010).

2.2.1 Resource-Rational Analysis of Numerical Estimation

Resource-rational analysis is a new approach to answering a classic question: how should we think and decide given that our time and our minds are finite?

Having introduced the basic concepts of resource rationality in Chapter 1, we now apply resource-

rational analysis to numerical estimation: We start by formalizing the problem solved by numerical estimation. Next, we specify an abstract computational architecture. We then derive the optimal solution to the numerical estimation problem afforded by the computational architecture. This resource-rational strategy will then be evaluated against empirical data in the remainder of this chapter.

FUNCTION

In numerical estimation people have to make an informed guess about an unknown quantity X based on their knowledge K. In general, people's relevant knowledge K is incomplete and insufficient to determine the quantity X with certainty. For instance, people asked to estimate the boiling point of water on Mount Everest typically do not know its exact value, but they do know related information, such as the boiling point of water at normal altitude, the freezing point of water, the qualitative relationship between altitude, air pressure, and boiling point, and so on. We formalize people's uncertain belief about X by the probability distribution $P(X|K)$ which assigns a plausibility $p(X = x|K)$ to each potential value x. According to Bayesian decision theory, the goal is to report the estimate $\hat{x}$ with the highest expected utility $\mathbb{E}_{P(X|K)}[u(\hat{x}, x)]$. This is equivalent to finding the estimate with the lowest expected error cost

$$x^{\star} = \arg\min_{\hat{x}} \mathbb{E}_{P(X|K)}[\text{cost}(\hat{x}, x)], \tag{2.1}$$

where $x^{\star}$ is the optimal estimate, and $\text{cost}(\hat{x}, x)$ is the error cost of the estimate $\hat{x}$ when the true value is x.

MODEL OF MENTAL COMPUTATION

How the mind should solve the problem of numerical estimation (see Equation 2.1) depends on its computational architecture. Thus, to derive predictions from the assumption of resource-rationality we have to specify the mind's elementary operations and their cost. To do so, we build on the resource-rational analysis by Vul et al. (2014) which assumed that the mind's elementary computation is *sampling*. Sampling is widely used to solve inference problems in statistics, machine learning, and artificial intelligence (Gilks, Richardson, & Spiegelhalter, 1996). Several behavioral and neuroscientific experiments suggest that the brain uses computational mechanisms similar to sampling for a wide range of inference problems ranging from vision to causal learning (Bonawitz, Denison, Gopnik, &

Griffiths, 2014; Bonawitz, Denison, Griffiths, & Gopnik, 2014; Denison, Bonawitz, Gopnik, & Griffiths, 2013; Fiser, Berkes, Orbán, & Lengyel, 2010; Griffiths & Tenenbaum, 2006; N. Stewart et al., 2006; Vul et al., 2014). One piece of evidence is that people's estimates of everyday events are highly variable even though the average of their predictions tends to be very close to the optimal estimate prescribed by Bayesian decision theory (see Equation 2.1, Griffiths & Tenenbaum, 2006; 2011). Furthermore, Vul et al. (2014) found that the relative frequency with which people report a certain value as their estimate is roughly equal to its posterior probability, as if the mind was drawing one sample from the posterior distribution.

Sampling stochastically simulates the outcome of an event or the value of a quantity such that, on average, the relative frequency with which each value occurs is equal to its probability. According to Vul et al. (2014), people may estimate the value of an unknown quantity X using only a single sample from the subjective probability distribution $P(X|K)$ that expresses their beliefs. If the expected error cost (Eq. 2.1) is approximated using a single sample $\tilde{x}$, then that sample becomes the optimal estimate. Thus, the observation that people report estimates with frequency proportional to their probability is consistent with them approximating the optimal estimate using only a single sample.

However, for the complex inference problems that people face in everyday life generating even a single perfect sample can be computationally intractable. Thus, while sampling is a first step from computational level theories based on probabilistic inference towards cognitive mechanisms, a more detailed process model is needed to explain how simple cognitive mechanisms can solve the complex inference problems of everyday cognition. Here, we therefore explore a more fine-grained model of mental computation whose elementary operations serve to approximate sampling. In statistics, machine learning, and artificial intelligence sampling is often approximated by Markov chain Monte Carlo (MCMC) methods (Gilks et al., 1996). MCMC algorithms allow the drawing of samples from arbitrarily complex distributions using a stochastic sequence of approximate samples, each of which depends only on the previous one. Such stochastic sequences are called Markov chains; hence the name Markov chain Monte Carlo.

The remainder of the chapter explores the consequences of assuming that people answer numerical estimation questions by engaging in a thought process similar to MCMC. We assume that the mind's computational architecture supports MCMC by two basic operations: The first operation takes in the current estimate and stochastically modifies it to generate a new one. The second operation compares the posterior probability of the new estimate to that of the old one and accepts or rejects the modification stochastically. Furthermore, we assume that the cost of computation is proportional to how many such operations have been performed. These two basic operations are suf-

ficient to execute an effective MCMC strategy for probabilistic inference known as the Metropolis-Hastings algorithm (Hastings, 1970). This algorithm is the basis for our anchoring-and-adjustment models as illustrated in Figure 2.1.
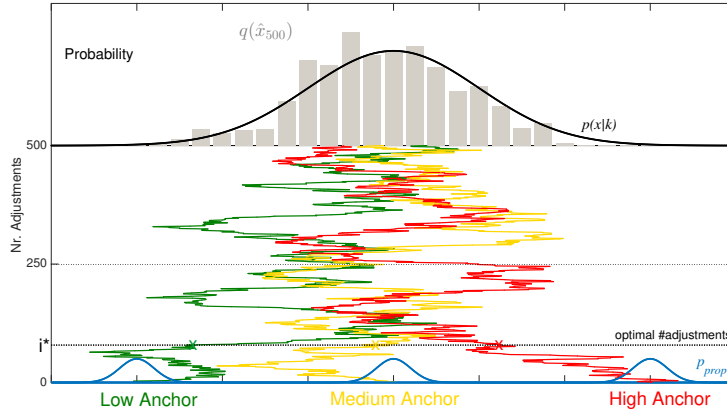
Figure 2.1: The figure illustrates the resource-rational anchoring-and-adjustment. The three jagged lines are examples of the stochastic sequences of estimates the adjustment process might generate starting from a low, medium, and high anchor respectively. In each iteration a potential adjustment is sampled from a proposal distribution p_{prop} illustrated by the bell curves. Each proposed adjustment is stochastically accepted or rejected such that over time the relative frequency with which different estimates are considered $q(\hat{x}_t)$ becomes the target distribution $p(x|k)$. The top of the figure compares the empirical distribution of the samples collected over the second half of the adjustments with the target distribution $p(x|k)$. Importantly, this distribution is the same for each of the three sequences. In fact, it is independent of the anchor, because the influence of the anchor vanishes as the number of adjustments increases. Yet, when the number of adjustments (iterations) is low (e.g., 25), the estimates are still biased towards their initial values. The optimal number of iterations $i^\star$ is very low as illustrated by the dotted line. Consequently, the resulting estimates indicated by the red, yellow, and red cross are still biased towards their respective anchors.

To be concrete, given an initial guess $\hat{x}_0$, which we will assume to be the anchor a ($\hat{x}_0 = a$), this algorithm performs a series of adjustments. In each step a potential adjustment δ is proposed by sampling from a symmetric probability distribution P_{prop} ($\delta \sim P_{\text{prop}}, P_{\text{prop}}(-\delta) = P_{\text{prop}}(\delta)$). The adjustment will either be accepted, that is $\hat{x}_{t+1} = \hat{x}_t + \delta$, or rejected, that is $x_{t+1} = \hat{x}_t$. If a proposed adjustment makes the estimate more probable ($P(X = \hat{x}_t + \delta|K) > P(X = \hat{x}_t|K)$), then it will always be accepted. Otherwise the adjustment will be made with probability $\alpha = \frac{P(X=\hat{x}_t+\delta|K)}{P(X=\hat{x}_t|K)}$, that is according to the posterior probability of the adjusted relative to the unadjusted estimate. This strategy ensures that regardless of which initial value you start from, the

frequency with which each value x has been considered will eventually equal to its subjective probability of being correct, that is $P(X = x|K)$. This is necessary to capture the finding that the distribution of people's estimates is very similar to the posterior distribution $P(X = x|K)$ (Griffiths & Tenenbaum, 2006; Vul et al., 2014). More formally, we can say that as the number of adjustments t increases, the distribution of estimates $Q(\hat{x}_t)$ converges to the posterior distribution $P(X|K)$. This model of computation has the property that each adjustment decreases an upper bound on the expected error by a constant multiple (Mengersen & Tweedie, 1996). This property is known as geometric convergence and illustrated in Figure 2.2.

There are several good reasons to consider this computational architecture as a model of mental computation in the domain of numerical estimation: First, the success of MCMC methods in statistics, machine learning, and artificial intelligence suggests they are well suited for the complex inference problems people face in everyday life. Second, MCMC can explain important aspects of cognitive phenomena ranging from category learning (Sanborn et al., 2010) to the temporal dynamics of multistable perception (Gershman, Vul, & Tenenbaum, 2012; Moreno-Bote, Knill, & Pouget, 2011), causal reasoning in children (Bonawitz, Denison, Gopnik, & Griffiths, 2014), and developmental changes in cognition (Bonawitz, Denison, Griffiths, & Gopnik, 2014). Third, MCMC is biologically plausible in that it can be efficiently implemented in recurrent networks of biologically plausible spiking neurons (Buesing, Bill, Nessler, & Maass, 2011). Last but not least, process models based on MCMC might be able to explain why people's estimates are both highly variable (Vul et al., 2014) and systematically biased (Tversky & Kahneman, 1974).

Optimal resource-allocation

Resource-rational anchoring-and-adjustment makes three critical assumptions: First, the estimation process is a sequence of adjustments such that after sufficiently many steps the estimate will be a representative sample from the belief $P(X|K)$ about the unknown quantity X given the knowledge K. Second, each adjustment costs a fixed amount of time. Third, the number of adjustments is chosen to achieve an optimal speed-accuracy tradeoff. It follows, that people should perform the optimal number of adjustments, that is

$$t^{\star} = \arg\min_{t} \left[\mathbb{E}_{Q(\hat{X}_t)} \left[\text{cost}(x, \hat{x}_t) + \gamma \cdot t \right] \right], \tag{2.2}$$
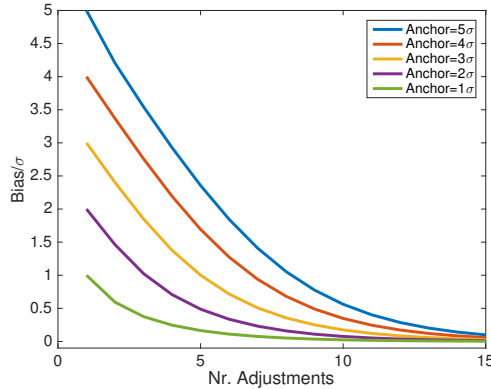
Figure 2.2: In resource-rational anchoring-and-adjustment the bias of the estimate is bounded by a geometrically decaying function of the number of adjustments. The plots shows the bias of resource-rational anchoring-and-adjustment as a function of the number of adjustments for five different initial values located $1, \cdots, 5$ posterior standard deviations (i.e., σ) away from the posterior mean. The standard normal distribution was used as both the posterior $P(X|K)$ and the proposal distribution $P_{\text{prop}}(\delta)$.

where $Q(\hat{X}_t)$ is the distribution of the estimate after t adjustments, x is its unknown true value, $\hat{x}_t$ is the estimate after performing t adjustments, $\text{cost}(x, \hat{x}_t)$ is its error cost, and γ is the time cost per adjustment.

Figure 2.3 illustrates this equation showing how the expected error cost – which decays geometrically with the number of adjustments–and the time cost – which increases linearly – determine the optimal speed-accuracy tradeoff. We inspected the solution to Equation 2.2 when the belief and the proposal distribution are standard normal distributions (i.e. $P(X|K) = P(X^{\text{prop}}) = \mathcal{N}(0, 1)$) for different anchors. We found that for a wide range of realistic time costs the optimal number of adjustments (see Figure 2.4, top panel) is much smaller than the number of adjustments that would be required to eliminate the bias towards the anchor. Consequently, the estimate obtained after the optimal number of adjustments is still biased towards the anchor as shown in the bottom panel of Figure 2.4. This is a consequence of the geometric convergence of the error (see Figure 2.2) which leads to quickly diminishing returns for additional adjustments. This is a general property of this rational model of adjustment that can be derived mathematically (Lieder, Griffiths, & Goodman, 2012).

The optimal speed-accuracy tradeoff weights the costs in different estimation problems according to their prevalence in the agent's environment; for more information please see Appendix A.
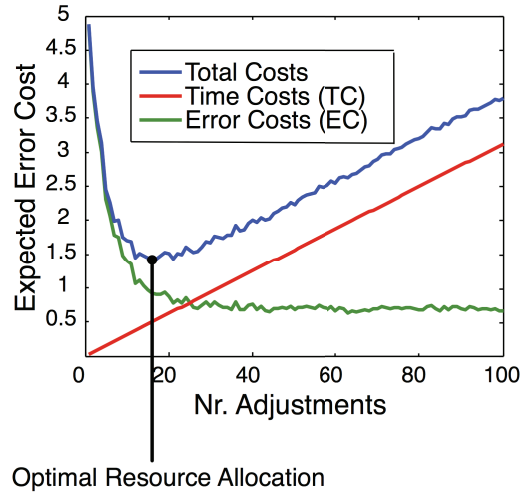
Figure 2.3: The expected value of the error cost $\text{cost}(x, \hat{x}_n)$ shown in green decays nearly geometrically with the number of adjustments n. While the decrease of the error cost diminishes with the number of adjustments, the time cost $\gamma \cdot t$ shown in red continues to increase at the same rate. Consequently, there is a point when further decreasing the expected error cost by additional adjustments no longer offsets their time cost so that the total cost shown in blue starts to increase. That point is the optimal number of adjustments $t^\star$.
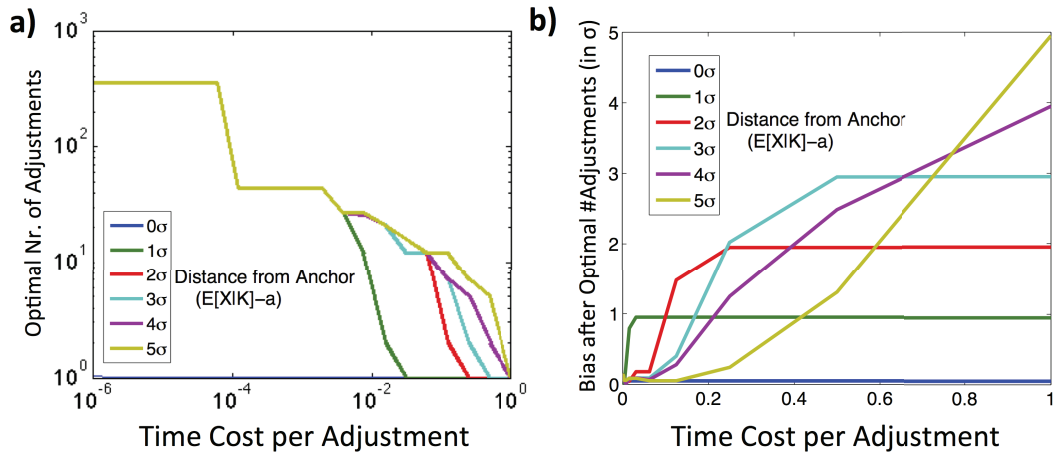
Figure 2.4: Optimal number of adjustments (a) and the bias after optimal number of adjustments (b) as a function of relative time cost and distance from the anchor.

2.2.2 RESOURCE-RATIONAL EXPLANATIONS OF ANCHORING PHENOMENA

Following the definition of the bias of an estimator in mathematical statistics, we quantify the anchoring bias by $B_t(x, a) = \mathbb{E}[\hat{x}_t | x, a] - x$, where $\hat{x}_t$ is a participant's estimate of a quantity x

after i adjustments, and a denotes the anchor. Figure 2.5 illustrates this definition and four basic ideas: First, the average estimate generated by anchoring-and-adjustment equals the anchor plus the adjustment. Second, the adjustment equals the relative adjustment times the total distance from the anchor to the posterior expectation. Third, adjustments tend to be insufficient, because the relative adjustment size is less than one. Therefore, the average estimate usually lies between the anchor and the correct value. Fourth, because the relative adjustment is less than one, the anchoring bias increases linearly with the distance from the anchor to the correct value.

More formally, the bias of resource-rational anchoring-and-adjustment cannot exceed a geometrically decaying function of the number of adjustments as illustrated in Figure 2.2:

$$B_t(x, a) = \mathbb{E}[\hat{x}_t | x, a] - x \leq B_0(x, a) \cdot r^t = (a - x) \cdot r^t, \tag{2.3}$$

where r is the rate of convergence to the distribution $P(X|K)$ that formalizes people's beliefs. Consequently, assuming that the bound is tight, resource-rational anchoring-and-adjustment predicts that, on average, people's predictions $\hat{x}$ are a linear function of the correct value x and the anchor a:

$$\mathbb{E}[\hat{x}_t | x, a] \approx a \cdot r^t + (1 - r^t) \cdot x. \tag{2.4}$$

Therefore the anchoring bias remaining after a fixed number of adjustments increases linearly with the distance from the anchor to the correct value as illustrated in Figure 2.5.

The hypothesis that the mind performs probabilistic inference by sequential adjustment makes the interesting, empirically testable prediction that the less time and computation a person invests into generating an estimate, the more biased her estimate will be towards the anchor. As illustrated in Figure 2.6a, the relative adjustment (see Figure 2.5) increases with the number of adjustments. When the number of adjustments is zero, then the relative adjustment is zero and the prediction is the anchor regardless of how far it is away from the correct value. However, as the number of adjustments increases, the relative adjustment increases and the predictions become more informed by the correct value. As the number of adjustments tends to infinity, the average guess generated by anchoring-and-adjustment converges to the expected value of the posterior distribution.

Our analysis of optimal resource-allocation shows that, for a wide range of plausible costs of computation, the resource-rational number of adjustments is much smaller than the number of adjustments required for convergence to the posterior distribution. This might explain why people's estimates of unknown quantities are biased towards their anchor across a wide range of circumstances.
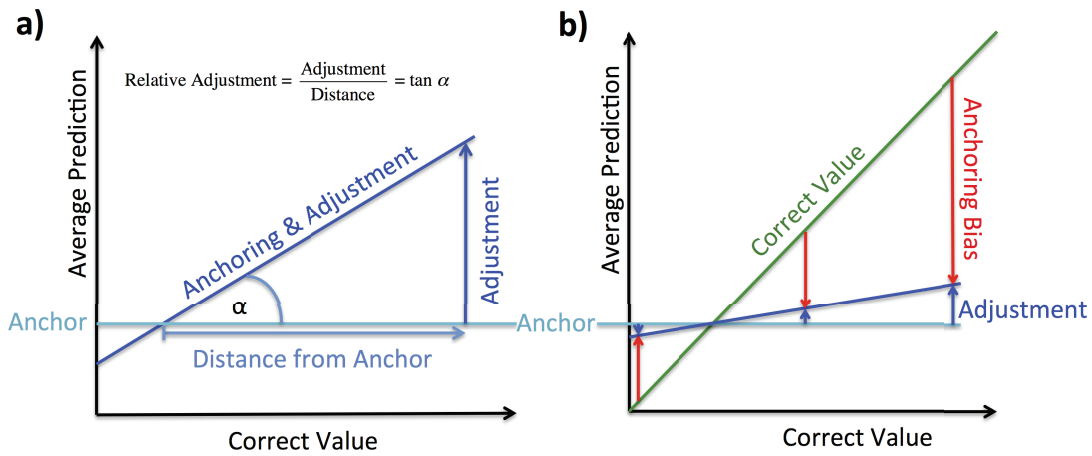
Figure 2.5: If the relative adjustment is less than 100%, then the adjustment is less than the distance from the anchor and the prediction is biased (Panel a) and the magnitude of the anchoring bias increases with the distance of the correct value from the anchor (Panel b).

Yet, optimal resource allocation also entails that the number of adjustments increases with the relative cost of error and decreases with the relative cost of time. Hence, our theory predicts that the anchoring bias is smaller when errors are costly and larger when time is costly; Figure 2.6b illustrates this prediction.

Although we derived the implications of making rational use of finite cognitive resources for a specific computational mechanism based on sampling, the crucial property of diminishing returns per additional computation is a universal feature of iterative inference mechanisms including (stochastic) gradient descent, variational Bayes, predictive coding (Friston, 2009; Friston & Kiebel, 2009), and probabilistic computation in cortical microcircuits (Habenschuss, Jonke, & Maass, 2013). Therefore the qualitative predictions shown in Figures 2.3–2.6 are not specific to the abstract computational architecture that we chose to analyze but characterize bounded rationality for a more general class of cognitive architectures.

In the following sections, we assess these and other predictions of our model through computer simulation and behavioral experiments.

49

Table 2.1: Anchoring phenomena and resource-rational explanations

Anchoring Effect	Simulated Results	Resource-Rational Explanation
Insufficient adjustment from provided anchors	Tversky and Kahneman (1974), Jacowitz and Kahneman (1995)	Rational speed-accuracy tradeoff.
Insufficient adjustment from self-generated anchors	Epley and Gilovich (2006), Study 1	Rational speed-accuracy tradeoff.
Cognitive load, time pressure, and alcohol reduce adjustment.	Epley, & Gilovich (2006), Study 2	Increased cost of adjustment reduces the resource-rational number of adjustments.
Anchoring bias increases with anchor extremity.	Russo and Schoemaker (1989)	Each adjustment reduces the bias by a constant factor (Equation 2.3). Since the resource-rational number of adjustments is insufficient, the bias is proportional to the distance from anchor to correct value.
Uncertainty increases anchoring.	Jacowitz and Kahneman (1995)	The expected change per adjustment is small when nearby values have similar plausibility.
Knowledge can reduce the anchoring bias.	Wilson et al. (1996), Study 1	High knowledge means low uncertainty. Low uncertainty leads to high adjustment (see above).
Accuracy motivation reduces anchoring bias when the anchor is self-generated but not when it is provided.	Tversky and Kahneman (1974), Epley and Gilovich (2005)	1. People are less uncertain about the quantities for which they generate their own anchors. 2. Accuracy motivation increases the number of adjustments but change per adjustment is lower when people are uncertain.
Telling people whether the correct value is larger or smaller than the anchor makes financial incentives more effective.	Simmons et al. (2010), Study 2	Being told the direction of adjustments makes adjustments more effective, because adjustments in the wrong direction will almost always be rejected.
Financial incentives are more effective when the anchor is extreme.	Simmons et al. (2010), Study 3	Values on the wrong side of an extreme anchor are much less plausible than values on the correct side. Therefore proposed adjustments in the wrong direction will almost always be rejected.
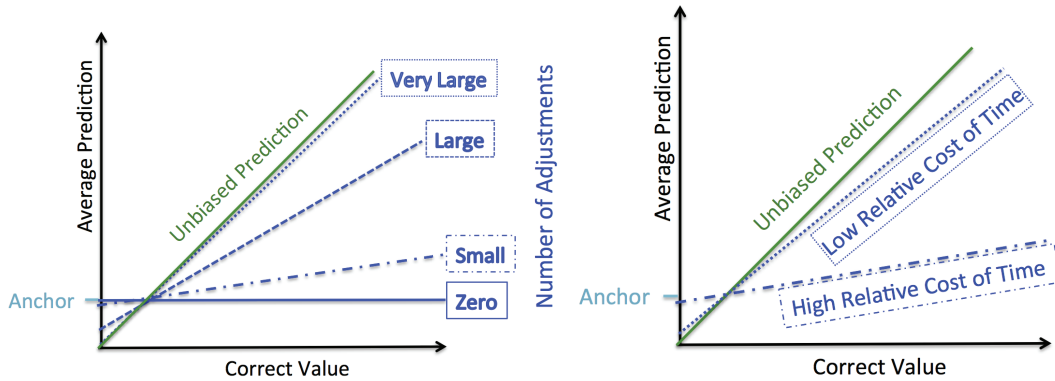
Figure 2.6: The number of adjustments increases the relative size of adjustments (left panel). As the relative cost of time increases, the number of adjustments decreases and so does the relative size of the adjustment (right panel).

2.3   SIMULATION OF ANCHORING EFFECTS

Having derived a resource-rational model of anchoring-and-adjustment we performed computer simulations to test whether this model is sufficient to explain the plethora of anchoring effects reviewed above. To capture our assumption that people make adjustments in discrete steps, we model the size of adjustments using the Poisson distribution $P(\delta) = \text{Poisson}(|\delta|; \mu_{\text{prop}})$. The simulated effects cover a wide range of different phenomena, and our goal is to account for all of these phenomena with a single model.

2.3.1   SIMULATION METHODOLOGY

We simulated the anchoring experiments listed in Table 2.1 with the resource-rational anchoring-and-adjustment model described above. The participants in each of these experiments were asked to estimate the value of one or more quantities X; for instance Tversky and Kahneman (1974) asked their participant to estimate the percentage of African countries in the United Nations. Our model's prediction of people's estimates of a quantity X depends on their probabilistic belief $P(X|K)$ based on their knowledge K, the number of adjustments, the anchor, and the adjustment step-size. Thus, before we could apply our model to simulate anchoring experiments, we had to measure people's probabilistic beliefs $P(X|K)$ about the quantities used on the simulated experiments. Appendix A describes our methodology and reports the estimates with obtained.

To accommodate differences in the order of magnitude of the quantities to be estimated and the effect of incentives for accuracy, we estimated two parameters for each experiment: the expected step-size μ_{prop} of the proposal distribution $P(\delta) = \text{Poisson}(|\delta|; \mu_{\text{prop}})$ and the relative iteration cost γ. These parameters were estimated by the ordinary least-squares method applied to the summary statistics reported in the literature. For experiments comprising multiple conditions using the same questions with different incentives for accuracy we estimated a single step-size parameter that is expected to apply across all conditions and a distinct relative time cost parameter for each incentive condition.

2.3.2 Insufficient adjustment from provided and self-generated anchors

Resource-rational anchoring-and-adjustment provides a theoretical explanation for insufficient adjustment from provided and self-generated anchors in terms of a rational speed-accuracy tradeoff, but how accurately does this describe empirical data? To answer this question, we fit our model to two well-known anchoring experiments: one with provided and one with self-generated anchors.

Provided anchors

As an example of adjustment from provided anchors, we chose the study by Jacowitz and Kahneman (1995), because it rigorously quantifies the anchoring bias. Jacowitz and Kahneman (1995) asked their participants two questions about each of several unknown quantities: First they asked whether the quantity is larger or smaller than a certain value–the *provided anchor*. Next they asked the participant to estimate that quantity. For the first half of the participants the anchor was a low value (i.e. the 15th percentile of estimates people make when no anchor is provided), and for the second half of the participants the anchor was a high value (i.e. the 85th percentile). People's estimates were significantly higher when the anchor was high than when it was low. Jacowitz and Kahneman (1995) quantified this effect by the anchoring index (AI), which is the percentage of the distance from the low to the high anchor that is retained in people's estimates:

$$\text{AI} = \frac{\text{Median}(\hat{X}_{\text{high anchor}}) - \text{Median}(\hat{X}_{\text{low anchor}})}{\text{high anchor} - \text{low anchor}} \cdot 100\% \qquad (2.5)$$

Jacowitz and Kahneman (1995) found that the average anchoring index was about 50%. This means that the difference between people's estimates in the high versus the low anchor condition retained

about half of the distance between the two anchors.

We determined the uncertainty σ for each of the 15 quantities by the elicitation method described above. Since Jacowitz and Kahneman (1995) measured people's median estimates in the absence of any anchor, we used those values as our estimates of the expected values μ, because their sample and its median estimates were significantly different from ours.

Next, we estimated the adjustment step-size parameter and the relative time cost parameter by minimizing the sum of squared errors between the predicted and the observed anchoring indices. According to the estimated parameters, people performed 29 adjustments with an average step-size of 22.4 units. With these two estimated parameters the model accurately captures the insufficient adjustment from provided anchors reported by Jacowitz and Kahneman (1995): The model's adjustments are insufficient (i.e. anchoring index $>$ 0; see Equation 2.5) on all questions for which this had been observed empirically but not for the question on which it had not been observed; see Figure 2.7. Our model also captured the magnitude of the anchoring bias: the model's average anchoring index of 53.22% was very close to its empirical counterpart of 48.48%. Furthermore, our model also captured for which questions the anchoring bias was high and for which it was low: the correlation between the predicted and the empirical anchoring indices ($r(13) = 0.62, p = 0.0135$). The simulated and empirical anchoring effects are shown in Figure 2.7.

Self-generated anchors

As an example of adjustment from self-generated anchors we chose the studies reported in Epley and Gilovich (2006). In each of these studies participants were asked to estimate one or more unknown quantities such as the boiling point of water on Mount Everest for which many participants readily retrieved a well-known related quantity such as $272°$F ($100°$C). Afterwards participants were asked whether they knew and had thought of each intended anchor while answering the corresponding question. For each question, Epley and Gilovich (2006) computed the mean estimate of those participants who had thought of the intended anchor while answering it. We combined the data from all self-generated anchor questions without additional experimental manipulations for which Epley and Gilovich (2006) reported people's mean estimate, i.e. the first five question from Study 1a, the first five questions from Study 1b, and the control conditions of Study 2b (2 questions) and the first seven questions from Study 2c.[‡] We determined the means and uncertainties of the model's beliefs

---

[‡]The quantities were the year in which Washington was elected president, the boiling point on Mt. Everest, the freezing point of vodka, the lowest body temperature, the highest body temperature, and the duration
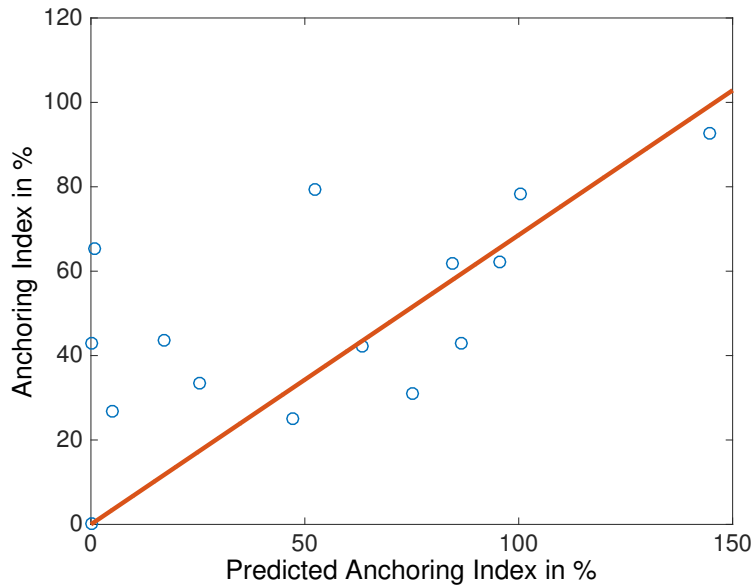
Figure 2.7: Simulation of the provided anchor experiment by Jacowitz and Kahneman (1995).

.

about all quantities used in Epley and Gilovich's studies by the elicitation method described above. The anchors were set to the intended self-generated anchors reported by Epley and Gilovich (2006). We estimated the model's time cost and adjustment step-size parameters by fitting the relative adjustments reported for these studies using the ordinary least-squares method.

The estimated parameters suggest that people performed 8 adjustments with an average step-size of 10.06 units. With these parameters the model adjusts its initial estimate by 80.62% of the distance to the correct value; this is very close to the 80.95% relative adjustment that Epley and Gilovich (2006) observed on average across the simulated studies. Our model captures that for the majority of quantities (13 out of 19) people's adjustments were insufficient. It also captures for which questions people adjust more and for which questions they adjust less from their uncertainties and anchors: as shown in Figure 2.8 our model's predictions of the relative adjustments were significantly correlated with the relative adjustments that Epley and Gilovich (2006) observed across different questions $(r(17) = 0.61, p = 0.0056)$. Comparing the parameter estimates between the experiments with provided versus self-generated anchors suggests that people adjusted less when they had generated the anchor themselves. This makes sense because self-generated anchors are typi-

---

of pregnancy in elephants. Some of these quantities were used in multiple studies.

cally much closer to the correct value than provided anchors.
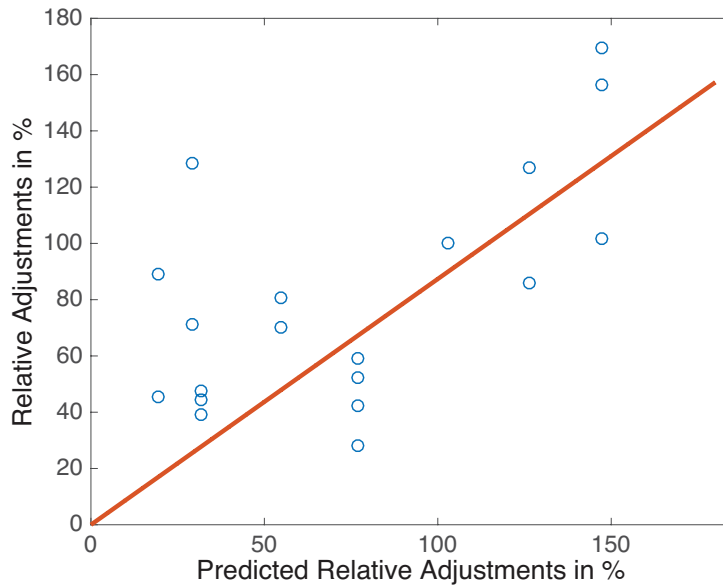
Figure 2.8: Simulation of self-generated anchors experiment by Epley, & Gilovich (2006).

.

In an experiment with self-generated anchors Epley and Gilovich (2006) found that people adjust their estimate less when required to simultaneously memorize an eight-letter string. To investigate whether resource-rational anchoring-and-adjustment can capture this effect, we fit our model simultaneously to participants' relative adjustment with versus without cognitive load. Concretely, we estimated a common step-size parameter and separate time cost parameters for each condition by the least squares method. We included all items for which Epley and Gilovich (2006) reported people's estimates. The resulting parameter estimates captured the effect of cognitive load: when people were cognitively busy, the estimated cost per adjustment was 4.58% of the error cost, but when people were not cognitively busy then it was only 0.003% of the error cost. The estimated average step-size per adjustment was $\mu = 11.69$. According to these parameters participants performed only 14 adjustments when they were under cognitive load but 60 adjustments when they are not. With these parameters our model captures the effect of cognitive load on relative adjustment: cognitive

load reduced the simulated adjustments by 18.61% (83.45% under load and 102.06% without load). These simulated effects are close to their empirical counterparts: people adjusted their estimate 72.2% when under load and 101.4% without cognitive load (Epley & Gilovich, 2006). Furthermore, the model accurately captured for which questions the effect of cognitive load was high and for which it was low; see Figure 2.9. Concretely, our model explained 93.03% of the variance in the effect of cognitive load on relative adjustments ($r(5) = 0.9645, p < 0.001$).

2.3.4   THE ANCHORING BIAS INCREASES WITH ANCHOR EXTREMITY

Next we simulated the anchoring experiment by Russo and Schoemaker (1989). In this experiment business students were first asked about the last three digits of their telephone number. Upon hearing the number the experimenter announced he would add 400 to this number (providing an anchor) and proceeded to ask the participant whether the year in which Attila the Hun was defeated in Europe was smaller or larger than that sum. When the participant indicated her judgment, she was prompted to estimate the year in which Attila had actually been defeated. Russo and Schoemaker (1989) then compared the mean estimate between participants whose anchor had been 500 ± 100, $700 \pm 100, \cdots, 1300 \pm 100$. They found that their participants' mean estimates increased linearly with the provided anchor even though the correct value was A.D. 451.

To simulate this experiment, we determined the values of μ and σ by the elicitation method described above. Since the variability of people's estimates and confidence intervals was very high, we increased the sample size of this one experiment to 200. We set the model parameters to the values estimated from the provided anchor experiments by Jacowitz and Kahneman (1995) (see above). As Figure 2.10 shows, our model correctly predicted that people's estimates increase linearly with the provided anchor (Russo & Schoemaker, 1989). To determine whether the quantitative differences between the model predictions and the data reported by Russo and Schoemaker (1989) were due to differences between business students in 1989 and people working on Mechanical Turk in 2014, we ran an online replication of their experiment on Mechanical Turk with 300 participants. There appeared to be no significant difference between the estimates of the two populations. However, people's estimates were highly variable. Consequently, the error bars on the mean estimates are very large.

Taking into account the high variance in people's judgments, our simulation results are largely consistent with the empirical data. In particular, both Russo and Shoemaker's data and our replication confirm our model's qualitative prediction that the magnitude of the anchoring bias increases
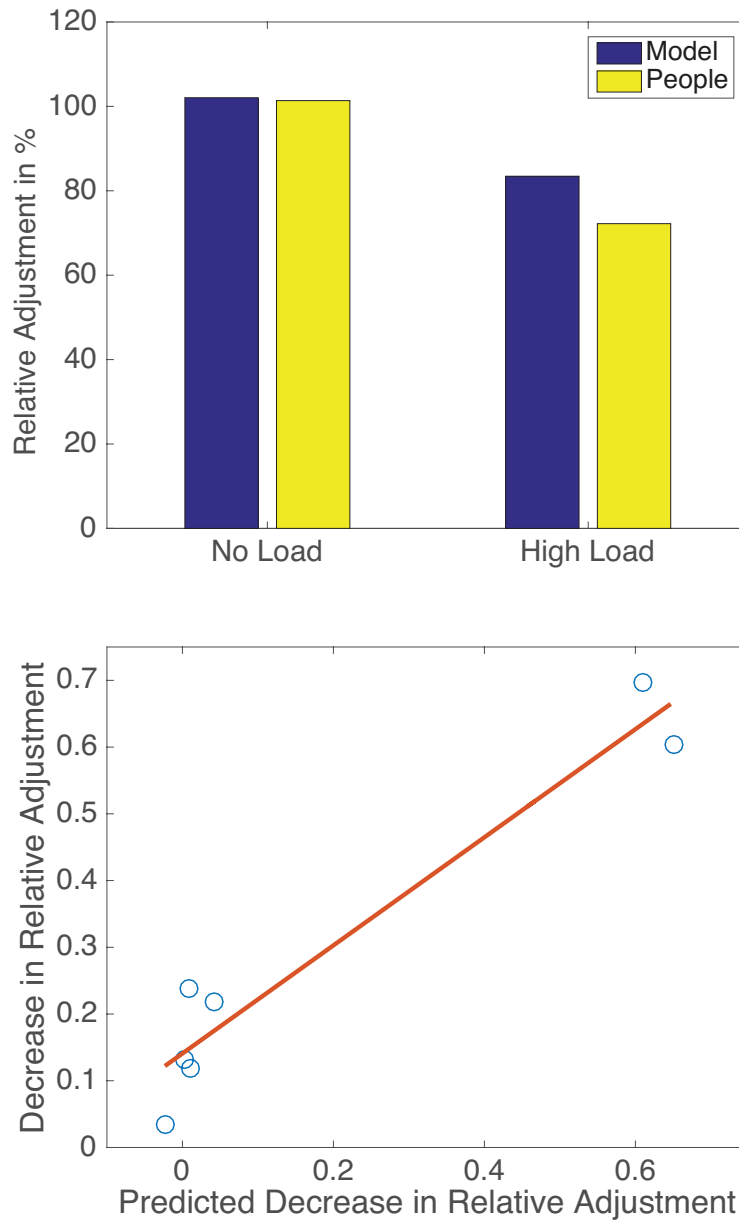
Figure 2.9: Simulated versus observed effect of cognitive load on the size of people's adjustments.

.

linearly with the anchor, although our model's prediction for the highest anchor was more extreme than the average judgment.
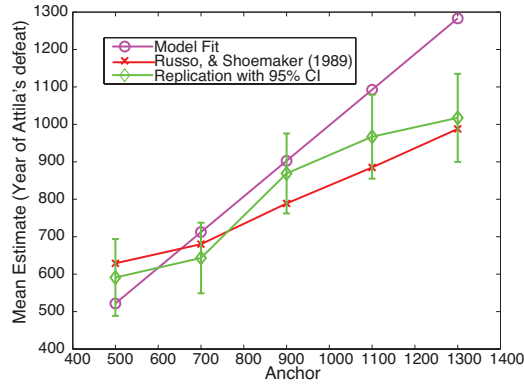
Figure 2.10: Simulated effect of the anchor on people's estimates of the year of Atilla's defeat and empirical data from Russo & Shoemaker (1989).

.

2.3.5 THE EFFECTS OF UNCERTAINTY AND KNOWLEDGE

Several experiments have found that the anchoring bias is larger the more uncertain people are about the quantity to be estimated (Jacowitz & Kahneman, 1995; Wilson et al., 1996). To assess whether and how well our theory can explain this effect, we re-analyzed our simulation of the experiment by Jacowitz and Kahneman (1995) reported above. Concretely, we computed the correlation between the uncertainties σ of the modeled beliefs about the 15 quantities and the predicted anchoring indices. We found that resource-rational anchoring-and-adjustment predicted that adjustments decrease with uncertainty. Concretely, the anchoring index that our model predicted for each quantity X was significantly correlated with the assumed uncertainty (standard deviation σ) about it (Spearman's $\rho = 0.5857, p = 0.0243$). This is a direct consequence of our model's probabilistic acceptance or rejection of proposed adjustments on a flat (high uncertainty) versus sloped (low uncertainty) belief distribution $P(X|K) = \mathcal{N}(\mu, \sigma)$. Our model thereby explains the negative correlation ($r(13) = -0.68$) that Jacowitz and Kahneman (1995) observed between confidence ratings and anchoring indices.

Uncertainty reflects the lack of relevant knowledge. Thus people who are knowledgeable about a quantity should be less uncertain and consequently less susceptible to anchoring. Wilson et al. (1996) conducted an anchoring experiment in which people first compared the number of countries in the United Nations (UN) to an anchor, then estimated how many countries there are in the UN,

58

and finally rated how much they know about this quantity. They found that people who perceived themselves as more knowledgeable were resistant to the anchoring bias whereas people who perceived themselves as less knowledgeable were susceptible to it. Here, we asked whether our model can explain this effect by smaller adjustments due to higher uncertainty. To answer this question, we recruited 60 participants on Mechanical Turk, asked them how much they knew about the number of nations in the UN on a scale from 0 ("nothing") to 9 ("everything") and elicited their beliefs by the method described in Appendix A. We then partitioned our participants into a more knowledgeable and a less knowledgeable group by a median split as in Wilson et al. (1996). We model the beliefs elicited from the two groups by two separate normal distributions (Appendix A).

We found that the high-knowledge participants were less uncertain than the low-knowledgeable participants ($\sigma_{\text{high}} = 35.1$ vs. $\sigma_{\text{low}} = 45.18$). Furthermore, their median estimate was much closer to the true value of 193 ($\mu_{\text{high}} = 185$ vs. $\mu_{\text{low}} = 46.25$). We fit the relative adjustments from the anchor provided in Wilson et al.'s experiment (1930) by the least-squares method as above. With the estimated parameters (17 adjustments, step-size 488.2) the model's predictions captured the effect of knowledge: For the low-knowledge group the model predicted that providing the high anchor would raise their average estimate from 45.18 to 252.1. By contrast, for the high-knowledgeable group our model predicted that providing a high anchor would fail to increase people's estimates (185 without anchor, 163 with high anchor).

2.3.6 Differential effects of accuracy motivation

People tend to invest more mental effort when they are motivated to be accurate. To motivate participants to be accurate some experiments employ financial incentives for accuracy, while others warn their participants about potential errors that should be avoided (forewarnings). Consistent with the effect of motivation, resource-rational anchoring-and-adjustment predicts that the number of adjustments increases with the relative cost of error. Yet, financial incentives for accuracy reduce the anchoring bias only in some circumstances but not in others: First, the effect of incentives appeared to be absent when anchors were provided but present when they were self-generated (Epley & Gilovich, 2005; Tversky & Kahneman, 1974). Second, the effect of incentives was found to be larger when people were told rather than asked whether the correct value is smaller or larger than the anchor (Simmons et al., 2010). Here, we explore whether and how these interaction effects can be reconciled with resource-rational anchoring-and-adjustment.

Epley and Gilovich (2005) found that financial incentives and forewarnings decreased the anchoring bias when the anchor was self-generated but not when it was provided by the experimenter. From this finding Epley and Gilovich (2005) concluded that people use anchoring-and-adjustment only when the anchor is self-generated but not when it is provided. By contrast, Simmons et al. (2010) suggested that this difference may be mediated by people's uncertainty about whether the correct answer is larger or smaller than the anchor. They found that people are often uncertain in which direction they should adjust in questions used in experiments with provided anchors; so this may be why incentives for accuracy failed to reduce the anchoring bias in those experiments. Here we show that resource-rational anchoring-and-adjustment can capture the differential effectiveness of financial incentives in experiments with provided versus self-generated anchors. First, we show through simulation that given the amount of uncertainty that people have about the quantities to be estimated our model predicts a larger effect of accuracy motivation for the self-generated anchor experiments by Epley and Gilovich (2005) than for the provided anchor experiments by Tversky and Kahneman (1974) and Epley and Gilovich (2005).

First, we analyze people's beliefs about the quantities used in experiments with provided versus self-generated anchors with respect to their uncertainty. We estimated the mean μ and standard deviation σ of people's beliefs about each quantity X by the elicitation method described above. Because the quantities' values differ by several orders of magnitude, it would be misleading to compare the standard deviations directly. For example, for the population of Chicago (about $2,700,000$ people) a standard deviation of $1,000$ would express near-certainty, whereas for the percentage of countries in the UN the same standard deviation would express complete ignorance. To overcome this problem, the standard deviation has to be evaluated relative to the mean. We therefore compare uncertainties in terms of the signal-to-noise ratio (SNR). We estimated the SNR by the median of the signal-to-noise ratios of our participants' beliefs $(\text{SNR}_s = \mu_s^2/\sigma_s^2)$. We found that people tended to be much more certain about the quantities Epley and Gilovich (2005) used in their self-generated anchors experiments (median SNR: 21.03) than about those for which they provided anchors (median SNR: 4.58). A Mann-Whitney U-test confirmed that the SNR was significantly higher for self-generated anchoring questions than for questions with provided anchors $(U(18) = 74.0, p = 0.0341)$.

Given that people were more uncertain about the quantities used in the experiments with provided anchors, we investigated how this difference in uncertainty affects the effect of financial incen-

tives on the anchoring bias predicted by our resource-rational model. To do so, we simulated Study 1 from Epley and Gilovich (2005), in which they compared the effects of financial incentives between questions with self-generated versus provided anchors, and the provided anchors experiment by Tversky and Kahneman (1974). To assess whether our model can explain why the effect of motivation differs between questions with provided versus self-generated anchors, we evaluated the effects of motivation as follows: First, we fit our model to the data from the condition with self-generated anchors. Second, we use the estimated numbers of adjustments to simulate responses in the condition with provided anchors. Third, for each question, we measured the effect of motivation by the relative adjustment with incentives minus the relative adjustment without incentives. Fourth, we averaged the effects of motivation separately for all questions with self-generated versus provided anchors and compared the results.

We fit the relative adjustments on the questions with self-generated anchors with one step-size parameter and two relative time-cost parameters: The estimated step-size was 17.97. The estimated number of adjustments was 5 for the condition without incentives and 9 for the condition with incentives. According to these parameters, motivation increased the relative adjustment from self-generated anchors by 12.74% from 65.62% to 78.35%. This is consistent with the significant effect of 33.01% more adjustment that Epley and Gilovich (2005) observed for questions with self-generated anchors. For the condition with provided anchors Epley and Gilovich (2005) used four questions from the experiment by Jacowitz and Kahneman (1995) simulated above and the same incentives as in the questions with self-generated anchors. We therefore simulated people's responses to questions with provided anchors using the step-size estimated from the data by Jacowitz and Kahneman (1995) and the number of adjustments estimated from questions with self-generated anchors. Our simulation correctly predicted that incentives for accuracy fail to increase adjustment from provided anchors. Concretely, our simulation predicted 44.09% adjustment with incentives and 44.48% without. Thus, as illustrated in Figure 2.11, our model captures that financial incentives increased adjustment from self-generated anchors but not from provided anchors. According to our model, this difference is just an artifact of the confound that people know more about the quantities used in experiments with self-generated anchors than about the quantities used in experiments with provided anchors.

Finally, we simulated Study 2 from Epley and Gilovich (2005) in which they compared the effect of warning participants about the anchoring bias between questions with provided versus self-generated anchors. This study had 2 (self-generated anchors vs. provided anchors) × 2 (forewarnings vs. no forewarnings) conditions. Epley and Gilovich (2005) found that in the conditions

with self-generated anchors forewarnings increased adjustment, but in the conditions with provided anchors they did not. As before, we set the model's beliefs about the quantities used in this experiment using the elicitation method described above. We fit our model to the relative adjustments in the conditions with self-generated anchors. Concretely, we used the least-squares method to fit one step-size parameter and two time cost parameters: one for the condition with forewarnings and one for the condition without forewarnings. With these parameters, we simulated people's estimates in the conditions with self-generated anchors (to which the parameters were fit) and predicted the responses in the provided anchor conditions that we had *not* used for parameter estimation.
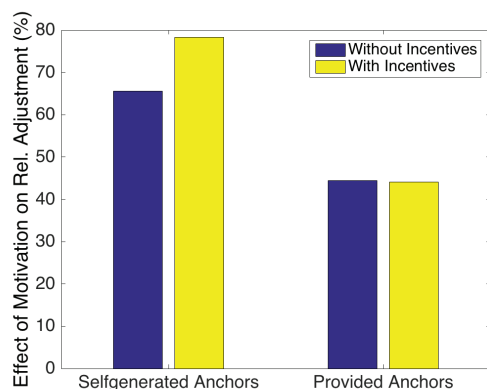
Figure 2.11: Simulation of Study 1 from Epley and Gilovich (2005): Predicted effects of financial incentives on the adjustment from provided versus self-generated anchors.

According to the estimated parameters, forewarnings increased the number of adjustments from 8 to 28. We therefore simulated the responses in both conditions with forewarnings (provided and self-generated anchor questions) with 8 adjustments and all responses in the two conditions without forewarnings (provided and self-generated anchor questions) with 28 adjustments. For the questions with self-generated anchors, forewarnings increased the simulated adjustments by 30% from insufficient 81% to overshooting 111% of the total distance from the anchor to the correct value.[§] By contrast, for questions with provided anchors forewarnings increased the simulated adjustments by only 12.5% from 6.9% to 19.4%. Thus, assuming that forewarnings increase the number of adjustments from provided anchors by the same number as they increase adjustments from self-generated anchors our model predicts that their effect on people's estimates would be less than one third of the effect for self-generated anchors; see Figure 2.12. According to our model, the reason is

---

[§]Overshooting is possible, because the expected value of the estimated belief $P(X|K) = \mathcal{N}(\mu, \sigma)$ can be farther away from the anchor than the correct value.

that people's uncertainty about the quantities for which anchors were provided is so high that the effect of additional adjustments is much smaller than in the questions for which people can readily generate their own anchors. Our results are consistent with the interpretation that the absence of a statistically significant effect of forewarnings on the bias towards the provided anchors in the small sample of Epley and Gilovich (2005) does not imply that the number of adjustments did not increase. Therefore adjustment from provided anchors cannot be ruled out.
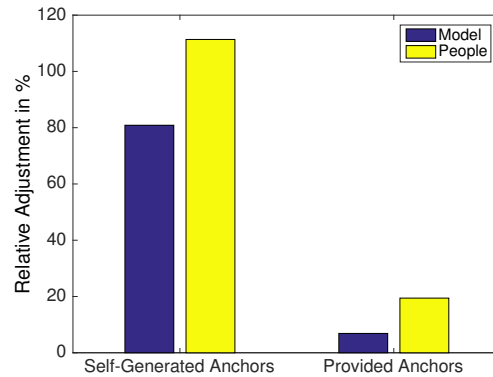
Figure 2.12: Simulation of Study 2 from Epley and Gilovich (2005): Predicted effects of forewarnings for questions from experiments with provided versus self-generated anchors.

DIRECTION UNCERTAINTY MASKS THE EFFECT OF INCENTIVES

Simmons et al. (2010) found that accuracy motivation decreases anchoring if people are confident about whether the quantity is larger or smaller than the anchor but not when they are very uncertain. Simmons et al. (2010) showed that even when the anchor is provided, incentives for accuracy can reduce the anchoring bias provided that people are confident about the correct direction of adjustment. Concretely, Simmons et al.'s second study unmasked the effect of incentives on adjustment from provided anchors by telling instead of asking their participants whether the true value is larger or smaller than the anchor. Similarly, in their third study Simmons et al. (2010) found that the effect of incentives is larger when the provided anchor is implausibly extreme than when it is plausible. Here we report simulations of both of these effects.

First, we show that our model can capture that the effect of incentives increases when people are told the correct direction of adjustment. Simmons et al.'s second study measured the effect of accuracy motivation on the anchoring index as a function of whether people were asked or told if

the correct value is larger or smaller than the anchor. We modeled the effect of being told that the quantity X is smaller or larger than the anchor a by Bayesian updating of the model's belief about X from $P(X|K)$ to $P(X|K, X < a)$ and $P(X|K, X > a)$ respectively. The original beliefs $P(X|K)$ were determined by the elicitation method described in Appendix A. We fit the model simultaneously to all anchoring indices by ordinary least squares to estimate one step-size parameter and one number of adjustments for each incentive condition. According to the estimated parameters, incentives increased the number of adjustments from 5 to 1000 and the average adjustment step-size was 11.6 units. For both incentive conditions, our model captured the variability of adjustments across trials: For trials with incentives for accuracy the correlation between simulated and measured anchoring indices was $r(18) = 0.77\,(p = 0.0001)$, and for trials without incentives this correlation was $r(18) = 0.61\,(p = 0.004)$. Our model also captured the overall reduction of anchoring with incentives for accuracy observed by Simmons et al. (2010), although the predicted 42% reduction of anchoring with incentives for accuracy was quantitatively larger than the empirical effect of 8%. Most importantly, our model predicted the effects of direction uncertainty on adjustment and its interaction with accuracy motivation: First, our model predicted that adjustments are larger if people are told whether the correct value is larger or smaller than the anchor. The predicted 13.7% reduction in the anchoring index was close to the empirically observed reduction by 18.8%. Second, our model predicted that the effect of accuracy motivation will be 6.3% larger when people are told the direction of adjustment. The predicted effect of direction uncertainty is smaller than the 21% increase reported by Simmons et al. (2010) but qualitatively consistent. Therefore, our model can explain why telling people whether the correct value is larger or smaller than the anchor increases the effect of accuracy motivation. According to our model, financial incentives increase the number adjustments in both cases, but knowing the correct direction makes adjustment more effective by eliminating adjustments in the wrong direction.

Second, we simulated Study 3b of Simmons et al. (2010) in which they showed that financial incentives increase adjustments away from implausible anchors. Concretely, this study compared the effect of accuracy motivation on adjustments between plausible versus implausible provided anchors. As before, we determined the model's beliefs by the procedure described above and estimated the number of adjustments with and without incentives (781 and 188) and the adjustment step-size (0.01) by fitting the reported relative adjustments by ordinary-least squares.¶ With this single set of parameters we simulated adjustments from plausible versus implausible provided anchors. The

<hr>

¶The reason that the estimated step-size is so small appears to be that all quantities and distances in this experiment are small compared to those in other experiments such as Study 2 by the same authors. The increase in the number of adjustments appears to compensate for the reduced step-size.

predicted adjustments captured a statistically significant proportion of the effects of anchor type, motivation, and quantity on the size of people's adjustments: $\rho(22) = 0.72, p < 0.0001$. Most importantly, our simulations predicted no statistically significant effect of accuracy motivation on absolute adjustment (mean effect: 0.76 units; 95% CI: $[-0.42; 1.94]$) when the anchor was plausible but a substantially larger and statistically significant effect when the anchor was implausible (17.8 units; 95% CI: $[9.76; 25.91]$); see Figure 2.13. This prediction results from the fact that large adjustments away from plausible anchors will often be rejected because they decrease the estimate's plausibility and small adjustments in the wrong direction are almost as likely to be accepted as adjustment in the correction direction because values on either side of the plausible anchor are almost equally plausible if the distribution is symmetric around its mode. Thus the expected change per adjustment is rather small.
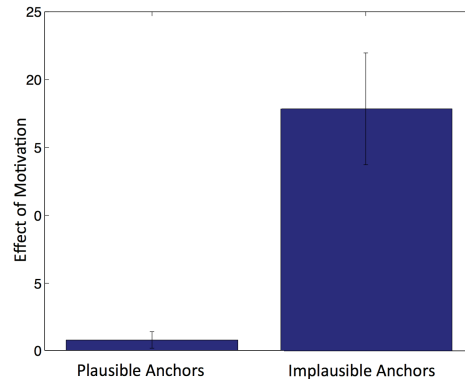
Figure 2.13: Simulation of Experiment 3 from Simmons et al. (2010): Predicted effect of accuracy motivation on adjustments from plausible versus implausible provided anchors.

In conclusion, resource-rational anchoring-and-adjustment can explain why motivating participants to be accurate reduces the anchoring bias in some circumstances but not in others. In a nutshell, our model predicts that incentives for accuracy have little effect when adjustments in either direction hardly change the estimate's plausibility. The simulations reported above demonstrate that this principle is sufficient to explain the differential effect of accuracy motivation on adjustments from provided versus self-generated anchors. Therefore, a single process – resource-rational anchoring-and-adjustment – may be sufficient to explain anchoring on provided and self-generated anchors.

2.3.7 SUMMARY

Our resource-rational analysis of numerical estimation showed that under-adjusting an initial estimate can be a rational use of computational resources. The resulting model can explain ten different anchoring phenomena: insufficient adjustments from both provided and self-generated anchors, the effects of cognitive load, anchor extremity, uncertainty, and knowledge, as well as the differential effects of forewarnings and financial incentives depending on anchor type (provided vs. self-generated), anchor plausibility, and being asked versus being told whether the quantity is smaller or larger than the anchor (see Table 2.1). None of the previous models (Epley & Gilovich, 2006; Simmons et al., 2010) was precise enough to make quantitative predictions about any of these phenomena let alone precisely predict all of them simultaneously. The close match between our simulation results and human behavior suggests that resource-rational anchoring-and-adjustment provides a unifying explanation for a wide range of disparate and apparently incompatible phenomena in the anchoring literature. Our model was able to reconcile these effects by capturing how the effect of adjustment depends on the location and shape of the posterior distribution describing the participants' belief about the quantity to be estimated. For instance, our model reconciles the apparent ineffectiveness of financial incentives at reducing the bias towards provided anchors (Tversky & Kahneman, 1974) with their apparent effectiveness at reducing bias when the anchor is self-generated (Epley & Gilovich, 2005). To resolve this apparent contradiction, we did not have to postulate additional processes that operate only when the anchor is provided–unlike Epley and Gilovich (2006). Instead, our computational model directly predicted this difference from people's higher uncertainty about the quantities used in experiments with provided anchors, because when the uncertainty is high then adjustments in the wrong direction are more likely to be accepted. Our model thereby provides a more parsimonious explanation of these effects than the proposal by Epley and Gilovich (2006). While Simmons et al. (2010) offered a conceptual explanation along similar lines, our model predicted the exact sizes of these effects a priori.

The parameter estimates we obtained differed significantly across the simulated phenomena. This is partly due differences in the incentives and other experimental manipulations. Additional reasons for the variability in the parameter estimates are somewhat arbitrary differences in the resolution of the hypothesis spaces across different quantities and the interdependence between the average change per adjustment and the number of adjustments: the same amount of adjustment can be explained either by a small number of large steps or a large number of small steps. For some experiments maximum likelihood estimation chose the former interpretation and for others it chose the latter. But because a larger step size can compensate for a smaller number of adjustments, it is

quite possible that the model could have explained all of the findings with a very similar step size and number of adjustment parameters if we knew the structure and resolution of people's hypothesis spaces for the quantities used in each experiment. Although the model's parameters were unknown and had to be estimated to make quantitative predictions, all of the qualitative phenomena we simulated logically follow from the structure of the model itself. In this sense, our model did not just capture the simulated phenomena but predicted them. Most importantly, our theory reconciles the apparently irrational effects of potentially irrelevant numbers with people's impressive capacity to efficiently handle a large number of complex problems full of uncertainty in a short amount of time. To further test the proposed cognitive mechanism, the following section test its novel empirical predictions against the predictions of alternative mechanisms including the stopping rule assumed by Epley and Gilovich (2006) and Simmons et al. (2010).

2.4   EXPERIMENTAL TESTS OF THE MODEL'S NOVEL PREDICTIONS

Having established that resource-rational anchoring-and-adjustment can explain a wide range of anchoring phenomena, we will now test its assumption that the number of adjustments is chosen to rationally tradeoff speed versus accuracy and test its novel predictions in two experiments. Here, we derive empirical predictions from this assumption that will be tested in the following two sections.

Recall that the number of adjustments determines how rapidly the anchoring bias increases with the distance of the correct value from the anchor, because the the slope of the anchoring bias is one minus the relative adjustment (Figure 2.5). We can therefore test our theory's predictions about the number of adjustments by measuring the slope of the anchoring bias in people's predictions. In the theory section, we derived an upper bound on the anchoring bias (Equation 2.3). This bound decays geometrically with the number of adjustments. If the bound is tight, then people's average prediction after a fixed number of adjustments should be a linear function of the distance from the anchor to the correct value (Equation 2.4). We can therefore rearrange Equation 2.4 into a linear regression model that allows us to estimate people's anchor a, their relative adjustments ($\frac{\mathbb{E}[\hat{X}|x]-a}{x-a}$), and the resulting anchoring bias $\text{Bias}_t(x,a)$ by regressing their estimates $\hat{X}$ on the correct value x:

$$\hat{X} = \alpha + \beta \cdot x + \varepsilon, \ \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{2.6}$$

$$\frac{\mathbb{E}[\hat{X}|x] - a}{x - a} = \beta, \ a = \frac{\alpha}{1 - \beta} \tag{2.7}$$

$$\text{Bias}_t(x, a) = \alpha - (1 - \beta) \cdot x \tag{2.8}$$

67

Optimal resource allocation implies that the relative adjustment decreases with the relative cost of time (Figure 2.6). Therefore the slope of the anchoring bias should be highest when time cost is high and error cost is low; see Figure 2.6. Conversely, the slope of the anchoring bias should be the shallowest when error cost is high and time cost is low. Lastly, when time cost and error cost are both high or both low, then the slope should be intermediate. Figure 2.14 illustrates these predictions. The following two sections report two experiments testing these predictions for self-generated and provided anchors respectively. Contrary to Epley and Gilovich (2006) our model assumes that people adjust not only with self-generated anchors but also from provided anchors. If this assumption is correct, then error cost should decrease and time cost should increase the anchoring bias regardless of whether anchors are self-generated (Experiment 1) or provided (Experiment 2). While previous studies have investigated the effect of financial incentives or deadlines (Epley & Gilovich, 2006), we are not aware of any study that has explicitly manipulated people's opportunity cost. Our opportunity cost manipulation is a more realistic model of the time constraint on judgment in the real world than imposing a deadline, because it allows participants to invest as much or as little of their valuable time as they like. This difference is critical because it allows us to study whether people rationally allocate their time and limited cognitive resources as predicted by our model. To measure time allocation we recorded our participants' reaction times. Another innovation of our experiments is to measure potential interactions between opportunity cost and error cost and to control people's uncertainty about the quantities to be estimated.
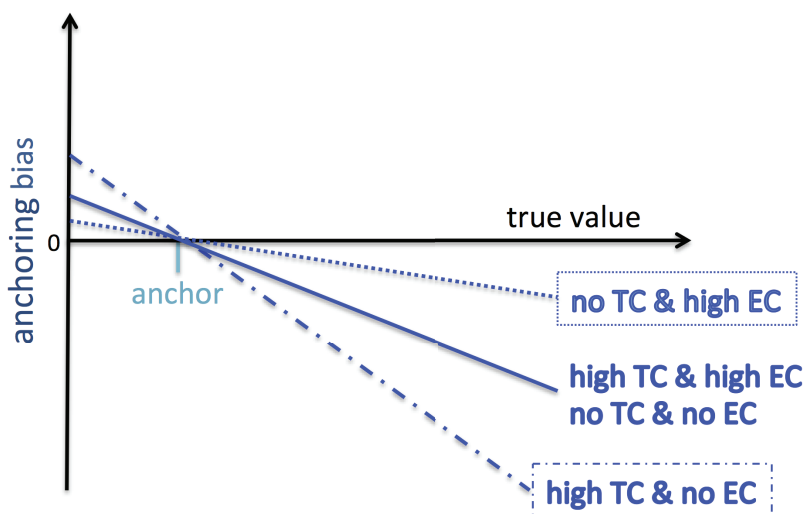
Figure 2.14: Resource-rational anchoring-and-adjustment predicts that the negative anchoring bias increases linearly with the distance from the anchor to the true value.

2.5 Experiment 1: Self-generated Anchors

In the experiments simulated above the biases in people's judgments result not only from anchoring but also from the discrepancy between the truth and what people actually know. To avoid this confound we designed a prediction task in which we can control both the prior and the likelihood function. To test if people adapt the number of adjustments to the relative cost of time we manipulated both the cost of time and the cost of error within subjects.

2.5.1 Method

Participants

We recruited 30 participants (14 male, 15 female, 1 unreported) on Amazon Mechanical Turk. Our participants were between 19 and 65 years old, and their level of education ranged from high school to graduate degrees. Participants were paid $1.05 for participation and could earn a bonus of up to $0.80 for points earned in the experiment. Six participants were excluded because they incorrectly answered questions designed to test their understanding of the task (see Procedure).

Materials

The experiment was presented as a website programmed in HTML and JavaScript. Participants predicted when a person would get on a bus given when he had arrived at the bus stop based on the bus's timetable and examples of previous departure times. Figure 2.15 shows a screenshot from one of the trials. The timeline at the top of the screen was used to present the relevant information and record our participants' predictions. At the beginning of each trial the bus's timetable (orange bars) and the person's arrival at the bus stop (blue bars) were highlighted on the timeline. Participants indicated their prediction by clicking on the corresponding point on the timeline. When participants were incentivized to respond quickly, a falling red bar indicated the passage of time and its cost in the bottom right corner of the screen, and the costs of error and time were conveyed in the bottom left corner; see Figure 2.15. Feedback was provided by highlighting the actual departure time on the number line (green bar) and a pop-up window informed participants about how many points they had earned. The complete experiment can be inspected online at http://cocosci.berkeley.edu/mturk/falk/PredictionExperiment1/experiment.html. We chose this

task to induce a bimodal posterior distribution (bus missed vs. not missed) because this might amplify the difference between sufficient versus insufficient adjustment.

Day 11

Today Will arrived at the bus stop at **7:26:40** AM. I predict that he will board a bus at

You have predicted that Will will wait 12.7 minutes.
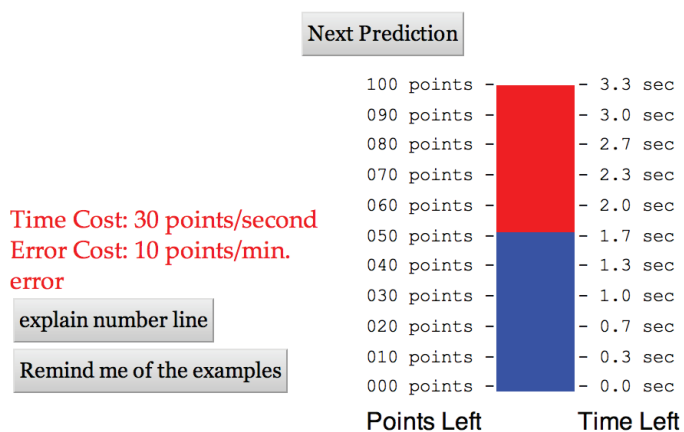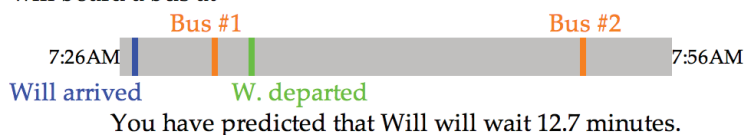
Figure 2.15: Screenshot of a prediction trial from Experiment 1 with time cost and error cost. The number line on the top conveys the bus schedule and when the person arrived at the bus stop. The cost of error and time are shown in the bottom left corner, and the red bar in the bottom right corner shows the passage of time and the cost associated with it.

Procedure

After completing the consent form, each person participated in four scenarios corresponding to the four conditions of a 2×2 within-subject design. The independent variables were time cost (0 vs. 30 points/sec) and error cost (0 vs. 10 points/unit error). The order of the four conditions was randomized between subjects. At the end of the experiment participants received a bonus payment proportional to the number of points they had earned in the experiment. The conversion rate was 1 cent per 100 points, and participants could earn up to 100 points per trial.

Each scenario comprised a cover story, instructions, 10 examples, 5 practice trials, 5 attention check questions, 20 prediction trials, 3 test questions, and one demographic question. Each cover story was about a person repeatedly taking the same bus route in the morning, for example "Jacob commutes to work with bus #22. On average, the first bus departs at 8:01 AM, and the second bus departs at 8:21 AM but departure times vary. On some days Jacob misses the first bus and takes the second bus." In each scenario both the person and the bus route were different. The task instructions informed participants about the cost of time and error and encouraged them to attentively study the examples and practice trials so that they would learn to make accurate predictions. After the cover story, participants were shown when the bus had arrived on the ten workdays of the two preceding weeks (10 examples); see Figure 2.16. Next participants made 5 practice predictions with feedback.

Week 1

If you study the following examples closely, you will earn more points by guessing more accurately. Each example is presented as points on a timeline: the scheduled departure time (i.e. when the bus is supposed to arrived), and when the first bus actually arrived.
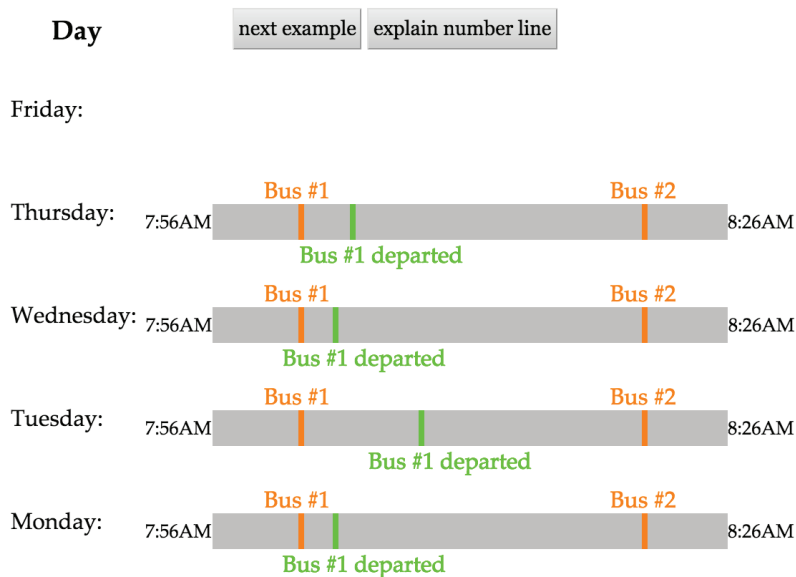
Figure 2.16: Screenshot of the first examples screen of Experiment 1.

The ensuing attention check questions verified the participants' understanding of the time line and the costs of time and error. Participants were allowed to go back and look up this information if necessary. Participants who made at least one error were required to retake this test until they got

all questions correct. Once they had answered all questions correctly, participants proceeded to 20 predictions trials with feedback. In both the practice trials and the prediction trials the feedback comprised the correct departure time, the incurred error cost, the incurred time cost, and the resulting number of points for the trial. The times at which the fictitious person arrived at the bus stop were chosen such that the probability that he had missed the first bus approximately covered the full range from 0 to 1 in equal increments. In the $1^{\text{st}}, 3^{\text{rd}}, \cdots, 2^{\text{nd}}$-last prediction trial the person arrived early and the bus was on time. The purpose of these odd-numbered trials was to set the anchor on the even-numbered trials to a low value. After each scenario's prediction trials we tested our participants' understanding of the number line, the cost of time, and the cost of error once again. We excluded six participants, because their answers to these questions revealed that they had misunderstood the number line, the cost of time, or the cost of error in at least one condition. After this they reported one piece of demographic information: age, gender, level of education, and employment status respectively. On the last page of each block, participants were informed about the bonus they had earned in the scenario.

To pose a different prediction problem on every trial of each block despite the limited number of meaningfully different arrival times, we varied the distribution of the bus's delays between blocks. There were four delay distributions in total. All of them were Pearson distributions that differed only in their variance. Their mean, skewness, and kurtosis were based on the bus lateness statistics from Great Britain.[‖] The order of the delay distributions was randomized between participants independently of the incentives. The 10 examples of bus departure times were chosen such that their mean, variance, and skewness reflected the block's delay distribution as accurately as possible. For each trial, a "correct" departure time x was sampled from the conditional distribution of departure times given that the fictitious person departs after his arrival at the bus stop. Our participants' responses were scored according the condition's cost of time c_t and cost of error c_e according to

$$\text{points} = \max\{0,\ 100 - c_e \cdot \text{PE} - c_t \cdot \text{RT}\}, \tag{2.9}$$

$$\text{PE} = |\hat{x} - x|, \tag{2.10}$$

where PE is the absolute prediction error between the estimate $\hat{x}$ and the true value x, and RT is the response time. The bottom part of Figure 2.15 shows how time cost and error cost were conveyed to the participants during the trials. The red bar on the right moved downward and its position indicates how much time has passed and how many points have consequently been lost.

[‖] Bus Punctuality Statistics GB 2007 report; http://estebanmoro.org/2009/01/waiting-for-the-bus/

The aim of this experiment was to test our theory's novel predictions. Before assessing these predictions, we verified our assumptions that a) people's predictions are biased, and b) the negative anchoring bias increases approximately *linearly* with the distance from the anchor to the correct value (Equation 2.3).

DATA ANALYSIS   Statistical analyses were performed using the Matlab statistics toolbox. Analysis of variance (ANOVA), regression, and t-tests were performed using the functions *anovan*, *regress*, and *ttest* respectively. Repeated measures ANOVAs were performed by including the participant number as a random effects factor.

ANCHORING BIAS AND LINEAR EFFECT OF DISTANCE   To assess whether our participants' predictions were systematically biased, we inspected their average prediction for a range of true bus delays. The true bus delays were sampled from a distribution, of which subjects had seen 10 samples. We binned participants' average predictions when the true bus delay was 0.5 ± 2.5min, 5.5 ± 2.5min, ..., or 35.5 ± 2.5min. Participants showed a systematic bias, overestimating the delay when its true value was less than 3 minutes ($t(815) = 16.0, p < 10^{-15}$), but underestimating it when its true value was larger than 7 minutes (all $p \leq 0.0011$; see Figure 2.17).

Visual inspection suggested that the bias was approximately proportional to the correct value (cf. Equations 2.3-2.4). Fitting the linear regression model derived from our theory (Equations 2.6-2.8)confirmed that the linear correlation between correct value and bias was significantly different from zero ($P(\text{slope} \in [-0.6148, -0.5596]) = 0.95$). This replicates the finding by (Russo & Schoemaker, 1989) predicted by our theory (Equation 2.3) and simulations (Figure 2.10). As shown in Figure 2.17, the bias was positive when the delay was greater than 7.5min and negative for greater delays. Our participants thus appeared to anchor around 7.5min and adjust their initial estimate by about 41.3% of the total distance to the true value (95%-CI: $[38.52\%, 44.04\%]$). Another, and perhaps more rational, strategy for choosing the anchor would be to re-use the estimate from the previous trial as the initial guess on the current trial. If so, then the estimate $\hat{X}_t$ on trial t might be generated according to

$$\hat{X}_t = \hat{x}_{t-1} + \beta \cdot (x_t - \hat{x}_{t-1}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{2.11}$$

where $\hat{x}_{t-1}$ was the participant's estimate on the previous trial and x_t is the true value on the current trial. To determine which of the two regression models better explains our data, we performed a model comparison using the Bayesian Information Criterion (BIC; Kass & Raftery, 1995). Our data provided very strong evidence for our original model with a fixed unknown anchor (BIC: 12 394) over the alternative model (BIC: 12 770). Hence, our participants did not appear to anchor on their previous estimate.[**] Critically, the anchoring effect we observed is more than a simple regression to a mean because its magnitude increased with the cost of time and decreased with the cost of error as shown in the following section.
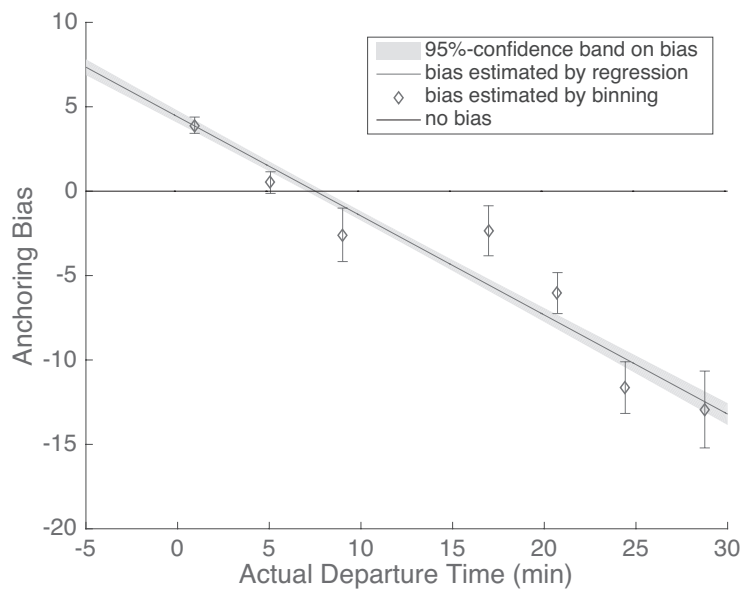
Figure 2.17: In Experiment 1 the magnitude of the anchoring bias grew linearly with the correct value. The error bars indicate 95% confidence intervals on the average bias, that is ± 1.96 standard errors of the mean.

EFFECTS OF TIME AND ERROR COST   Since the data showed standard anchoring effects, we can now proceed to testing its novel predictions. First, we investigated whether people adjust their prediction strategy to the incentives for speed and accuracy. To get a first impression we performed two repeated measures ANOVAs of the absolute error and the log-transformed reaction time in terms of

---

[**]According to the slope estimated using the alternative model, participants adjusted their estimate 65.76% of the distance to the correct value (95% CI: [63.49%; 68.03%]). Thus, regardless of which model is used to analyze our data, the results suggest that people's adjustments were insufficient.

time cost and error cost. The ANOVA models included the main effects of time cost and error cost and their interaction (fixed effects) as well as the main effect of participant number (random effect). The results suggest that participants traded accuracy for speed according to the experiment's incentives (see Figure ??): When errors were costly people took more time ($F(1, 1894) = 28.73, p < 0.0001$) and were more accurate ($F(1, 1824) = 15.52, p < 0.0003$) than when there was no error cost. Conversely, when time was costly people took less time ($F(1, 1824) = 73.51, p < 10^{-8}$) and were less accurate ($F(1, 1824) = 12.07, p = 0.0011$) than when there was no time cost. The interaction between time cost and error cost was significant for log reaction time ($F(1, 1824) = 7.17, p = 0.0075$) but not for accuracy ($F(1, 1824) = 0.13, p = 0.72$).
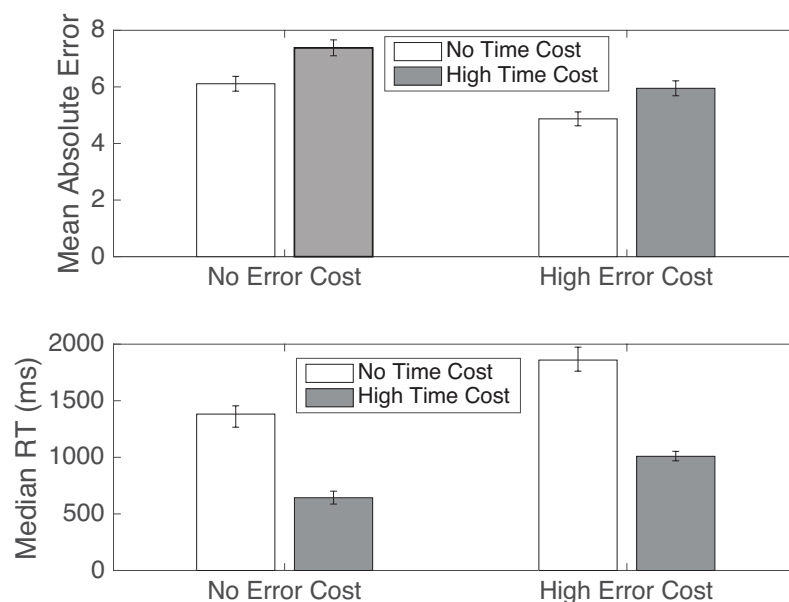
Figure 2.18: Mean absolute errors and reaction times as a function of time cost and error cost indicate an adaptive speed-accuracy tradeoff.

Given that our participants appeared to be sensitive to incentives for speed and accuracy, we asked whether time cost decreased and error cost increased our participants' anchoring biases. To answer this question we performed a repeated-measures ANOVA of our participants' relative adjustments as a function of time cost and error cost. To be precise, we first estimated each participant's relative adjustment separately for each of the four conditions using our linear regression model of anchoring and adjustment (Equation 2.6). We then performed an ANOVA on the estimated relative

adjustments with the factors time cost and error cost (fixed-effects) as well as participant number (random effect) and the interaction effect of time cost and error cost; see Table 2.2. We found that time cost significantly reduced relative adjustment from 50.7% to 31.0% ($F(1,69) = 21.86, p < 0.0001$) whereas error cost significantly increased it from 31.6% to 50.1% ($F(1,69) = 19.49, p < 0.0001$) and the interaction was non-significant. The mean relative adjustments of each condition are shown in Table 2.3. Consequently, as predicted by our theory (Figure 2.14), the anchoring bias increased more rapidly with the true delay when time cost was high or error cost was low (Figure 2.19). This is consistent with the hypothesis that people rationally adapt the number of adjustments to the relative cost of time.[††]

Table 2.2: ANOVA of relative adjustment as a function time cost and error cost.

Source	d.f.	Sum Sq.	Mean Sq.	F	p
error cost	1	0.82461	0.82461	19.49	3.6e-05
time cost	1	0.92484	0.92484	21.86	1.4e-5
error cost x time cost	1	0.04483	0.04483	1.06	0.3069
subject	23	1.77458	0.07716	1.82	0.0293
Error	69	2.91871	0.0423		
Total	95	6.48757			

Table 2.3: Relative size of our participants' adjustments of their initial guesses towards the correct answer by incentive condition with 95% confidence intervals.

	No Error Cost	High Error Cost
No Time Cost	$43.6 \pm 11.2\%$	$57.8 \pm 4.8\%$
High Time Cost	$19.6 \pm 9.0\%$	$42.5 \pm 9.8\%$

The effects of time cost and error cost on our participants' adjustments were also evident from how often their adjustments were insufficient. For this analysis, we only considered trials in which the arrival time suggested that the bus had been missed, that is when the probability of having missed the bus was larger than 0.5. For those trials, adjustments were considered sufficient when the prediction was larger than the expected departure of the second bus minus 2 standard deviations of the delay distributions. We found that the proportion of sufficient adjustments changed substantially with the cost of error and the cost of time (see Figure 2.20). Error cost significantly increased

---

[††]Estimating relative adjustment under the assumption that people anchor on their previous estimate led to the same conclusions.
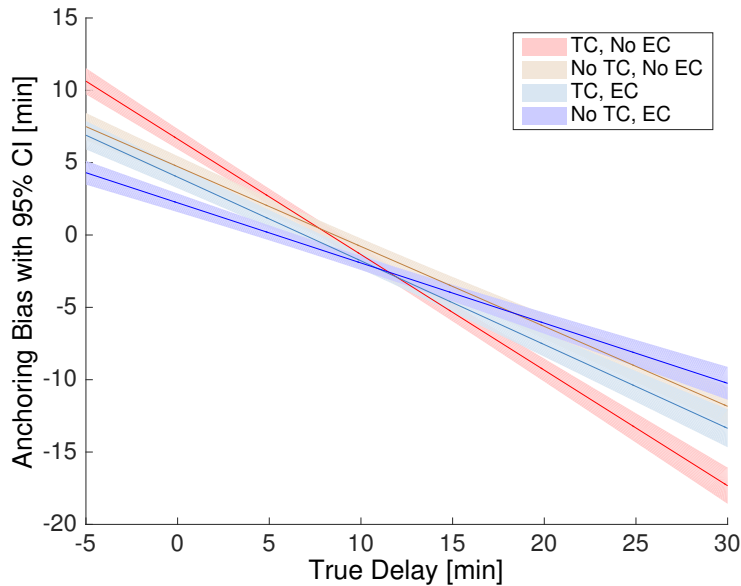
Figure 2.19: Anchoring bias in Experiment 1 by time cost and error cost confirms our theoretical prediction; compare Figure 2.14. The shaded areas are 95% confidence bands. The slope of a line equals one minus the relative adjustment.

the proportion of complete adjustments by $21\% \pm 4\%$ from 56% to 77% ($p < 10^{-6}$), whereas time cost significantly decreased it by $28.6\% \pm 4\%$ from 80.3% to 51.7% ($p = 4 \cdot 10^{-12}$).

2.5.3 COMPUTATIONAL MODELS OF ANCHORING-AND-ADJUSTMENT

To test competing theories of the anchoring bias, we formalized four theories using eight probabilistic models of numerical estimation. Appendix A describes these models in detail; in this section we will give only a brief conceptual overview. The theories range from unbounded Bayesian rationality (theory 1) to random guessing (theory 4) with theories 2 and 3 formalizing intermediate levels of rationality: the sampling hypothesis (theory 3; Vul et al., 2014) and four models of the anchoring-and-adjustment heuristic that range from resource-rational anchoring-and-adjustment to less rational anchoring heuristics like the ones proposed by Epley and Gilovich (2006) and Simmons et al. (2010). By formally comparing these models using Bayesian model selection, we will be able to titrate exactly how rational our participants' estimation strategy was.

According to the first theory, people draw Bayes-optimal inferences and the observed biases
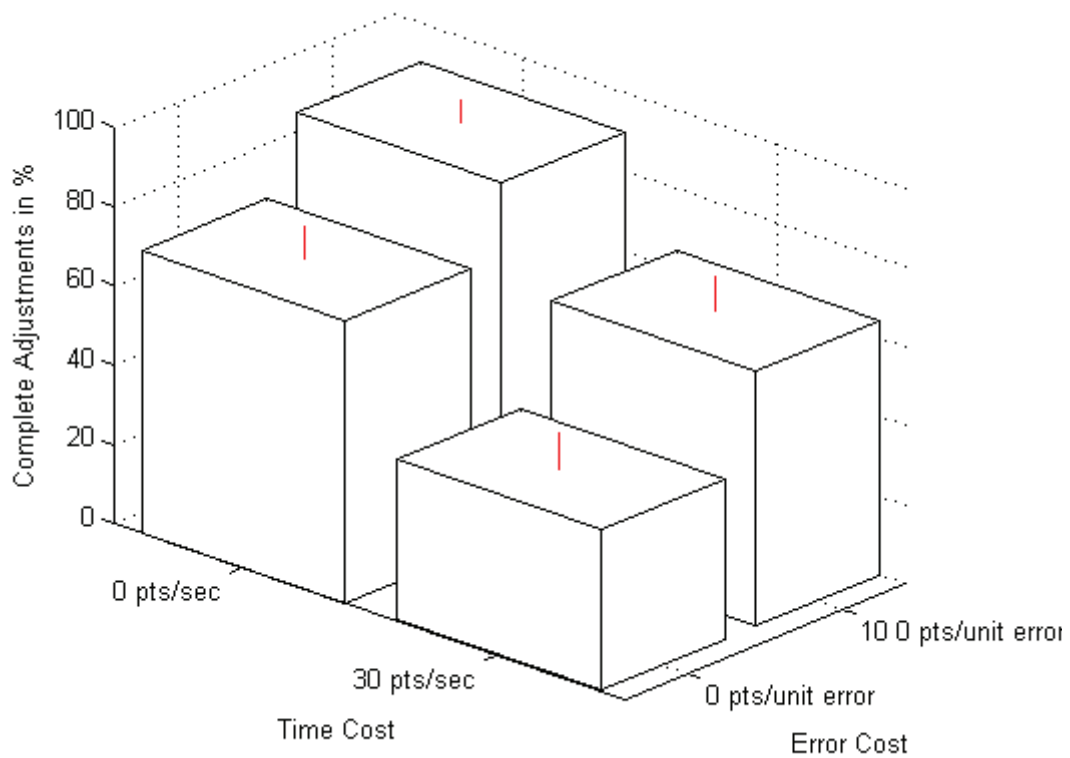
Figure 2.20: This plot shows the relative frequency of complete adjustments as a function of time cost and error cost. The length of the error bars is 1.96 standard errors.

merely reflect a regression towards their prior expectation. We formalized this explanation in terms of Bayesian decision theory (m_{BDT}; Equations A.2-A.5). To connect the deterministic predictions of Bayesian decision theory to people's variable responses, measurement and response errors are included in the model. According to the second theory, people approximate optimal inference by drawing a single sample from the posterior distribution (posterior probability matching, Vul et al., 2014, m_{PPM}, Equations A.6-A.8). However, generating even a single perfect sample can require an intractable number of computations. Therefore, according to the third theory, the mind approximates sampling from the posterior by anchoring-and-adjustment (Lieder et al., 2012). We modeled adjustment using the probabilistic mechanisms illustrated in Figure 2.1. We modified the stopping criterion to model several variants of anchoring-and-adjustment. Existing theories of anchoring-and-adjustments commonly assume that people adjust their estimate until it is sufficiently plausible (Ep-

ley & Gilovich, 2006; Simmons et al., 2010). Our first anchoring-and-adjustment model formalizes this assumption by terminating adjustment as soon as the estimate's posterior probability exceeds a certain plausibility threshold ($m_{A\&As}$, Equations A.9-A.17). The plausibility threshold and the average size of the adjustment are free parameters. According to the second anchoring-and-adjustment model, people make a fixed number of adjustments to their initial guess and report the result as their estimate ($m_{A\&A}$, Equations A.18-A.25). Here the number of adjustments replaces the plausibility-threshold as the model's second parameter. According to the third anchoring-and-adjustment model, people adapt the number of adjustments and the adjustment step size to optimize their speed-accuracy tradeoff ($m_{aA\&A}$, Equations A.26-A.37; Lieder, Griffiths, & Goodman, 2013). The optimal speed-accuracy tradeoff depends on the unknown time $\tau_{adjustment}$ it takes to perform an adjustment, so this time constant is a free-parameter. The fourth anchoring-and-adjustment model extends the third one by assuming that there is an intrinsic error cost in addition to the extrinsic error cost imposed by the experimenter, and this intrinsic cost is an additional model parameter (m_{aAAi}, Equations A.38-A.39). All anchoring models assumed that the anchor in Experiment 1 was the estimate reported in the previous section, that is 7.5 minutes. Finally, we also included a fourth theory. According to this "null hypothesis", our participants chose randomly among all possible responses (m_{random}, Equation A.40).

Except for the null model, the response distributions predicted by our models are a mixture of two components: the distribution of responses expected if people perform the task and the distribution of responses expected when they do not. The relative contributions of these two components are determined by an additional model parameter: the percentage of trials p_{cost} in which participants fail to perform the task. Not performing the task is modeled as random choice according to the null model. Performing the task is modeled according to the assumed estimation strategies described above. For a precise definition and comprehensive explanation of the each model, please consult Appendix A.

2.5.4 MODEL SELECTION

To formally test the four theories—anchoring-and-adjustment, posterior probability matching, Bayesian decision theory, and random choice—and which of the seven models that instantiate them against each other, we performed random-effects Bayesian model selection at the group level (Stephan, Penny, Daunizeau, Moran, & Friston, 2009) and family-level Bayesian model selection (Penny et al., 2010) as implemented in SPM8. For each model we separately approximated the log-

probability of each participant's predictions using the Laplace approximation (Tierney & Kadane, 1986) when applicable, that is when the likelihood function is differentiable with respect to the parameters, and numerical integration of the joint density otherwise. Numerical integration was necessary for discrete-valued parameters such as the number of adjustments. Numerical integration was also necessary for continuous parameters that affect the resource-rational number of adjustments. This is because the likelihood function changes abruptly by a non-differential step when the resource-rational number of adjustments jumps from one number to another. Numerical integration with respect to continuous parameters was performed using the functions *integral* and *integral2* available in Matlab 2013b.

According to Bayesian model selection, adaptive anchoring-and-adjustment with intrinsic error cost (m_{aAAi}) explained our participants' predictions better than any of the alternative models: we can be 99.99% confident that the adaptive anchoring-and-adjustment with intrinsic error is the best model for a larger percentage of people (64.4%) than any of the alternative models; see Figure 2.21, top panel. In addition to this random-effects analysis we also performed a Bayesian fixed effects analysis by computing the group Bayes factor for each pair of models. Reassuringly, this analysis led to the same conclusion: according to the posterior odds ratios, the adaptive anchoring-and-adjustment with intrinsic error cost was at least $\exp(220)$ times as likely as any of the other models we considered. Next, we applied family level inference to determine which theory best explains our data; see Figure 2.21, bottom left panel. According to this method, we can be 99.99% confident that anchoring-and-adjustment is the most probable explanation for a significantly larger proportion of participants (78.2%) than either posterior probability matching (11.0%), Bayesian decision theory (7.2%), or random choice (3.6%). Finally, we compared adaptive to non-adaptive models; see Figure 2.21, bottom right panel. According to the result, we can be 99.86% confident that for the majority of people (79.2%) our adaptive models' predictions are more accurate than the predictions of their non-adaptive counterparts.

Validation of the adaptive control of the number of adjustments

To validate that people perform more adjustments when errors are costly and fewer adjustments when time is costly, as assumed by the adaptive resource-rational model, we computed the maximum-a-posteriori estimates of the parameters of the second anchoring-and-adjustment model (m_{AA}) separately for each of the four incentive conditions. Figure 2.22 shows the estimated number of adjustments as a function of the incentives for speed and accuracy. For five of the six pairs of conditions,

we can be more than 96.9% confident that the number adjustments differ in the indicated direction, and for the sixth pair we can be more than 92% confident that this is the case. Therefore, this analysis supports the conclusion that our participants adapted the number of adjustments to the cost of time and error. To determine whether this pattern is consistent with choosing the number of adjustments adaptively we fit the parameters determining the rational number of adjustments to these estimates. We found that rational resource allocation predicts a qualitatively similar pattern of adjustments for reasonable parameter values (convergence rate: 0.71, time per adjustment: 27ms, assumed initial bias: 6.25min).

2.5.5 DISCUSSION

We observed a bias in people's predictions under uncertainty that increases with time cost and decreases with error cost. This phenomenon is consistent with the interpretation that people use anchoring-and-adjustment to make predictions under uncertainty. Our results suggested that anchoring-and-adjustment is used adaptively: When errors were costly, people invested more time and were more accurate. Their adjustments were larger and their anchoring bias was smaller. By contrast, when time was costly then our participants were faster and less accurate. Their adjustments appeared to be smaller and their anchoring bias was larger. This is consistent with the interpretation that people rationally choose the number of adjustments to optimize their speed-accuracy tradeoff. In fact, the experiment confirmed the predictions of optimal resource-allocation, and the data were best explained by a resource-rational anchoring-and-adjustment model. The anchoring bias may therefore be a consequence of resource-rational computation rather than a sign of human irrationality.

While our results demonstrate that people adaptively tradeoff being biased for being fast, our analysis had to postulate and estimate people's self-generated anchors. Therefore, we cannot be sure whether people really self-generated and adjusted anchors, or whether their responses merely look as if they did so. If people's predictions in Experiment 1 were generated by anchoring-and-adjustment, then we should be able to shift the biases shown in Figure 2.17 by providing different anchors; we tested this prediction in Experiment 2.

2.6   EXPERIMENT 2: PROVIDED ANCHORS

To test whether the biases observed in Experiment 1 resulted from anchoring and to evaluate whether the effects of time cost and error cost also hold for provided anchors, we ran a second experiment in which anchors were provided by asking participants to compare the to-be-predicted delay to a low versus a high number before every prediction. Concretely, this experiment tested two predictions: Our first prediction was that people's anchor will be higher when the number is high than when it is low. Our second prediction was that the bias towards the provided anchor decreases with error cost but increases with time cost.

2.6.1   METHOD

The materials, procedures, models, and data analysis tools used in Experiment 2 were identical to those used in Experiment 1 unless stated otherwise.

PARTICIPANTS

We recruited 60 participants (31 male, 29 female) on Amazon Mechanical Turk. They were between 18 and 60 years old, and their level of education ranged from high school diploma to PhD. Participants were paid $1.25 for participation and could earn a bonus of up to $2.20 for the points they earned in the experiment.

MATERIALS

Experiment 2 was presented as a website programmed in HTML and JavaScript. Experiment 2 was mostly identical to Experiment 1. The relevant changes are summarized below. The complete experiment can be inspected online at http://cocosci.berkeley.edu/mturk/falk/PredictionExperiment2/experiment.html.

PROCEDURE

Experiment 2 proceeded like Experiment 1 except for three changes: First, each prediction was preceded by the question "Do you think he will depart before or after X am?", where X is the an-

chor. This question was presented between the sentence reporting the time the person reached the bus stop and the number line. Participants were required to answer this question by selecting "before" or "after". This is the standard procedure for providing anchors (Jacowitz & Kahneman, 1995; Russo & Schoemaker, 1989; Tversky & Kahneman, 1974). In the two conditions with time cost, participants were given 3 seconds to answer this question before the timer started. Participants were not allowed to make a prediction until they had answered. We incentivized them to take this question serious by awarding +10 points for correct answers and -100 points for incorrect ones. For each participant the anchor was high in half of the trials of each condition and low in the other half. The low anchor was 3 minutes past the scheduled departure of the first bus, and the high anchor was 3 minutes past the scheduled departure of the second bus. The list of anchors was shuffled separately for each block and participant. Second, the 1st, 3rd, 5th,$\cdots$, 2nd-last trial were no longer needed, because they merely served to set the anchor on the even numbered trials of Experiment 1 to a small value. We therefore replaced those trials by 10 trials whose query times tighten the grid of those in the even-numbered trials. Thus for each participant, each block includes ten prediction trials with low anchors and ten prediction trials with high anchors. Third, we increased the base payment and the bonus payment, because Experiment 2 takes longer than Experiment 1. The conversion of points into bonuses remained linear but was scaled up accordingly. The instructions were updated to reflect the changes.

We excluded one participant due to incomplete data, and 16 participants because their answers to our test questions indicated they had misunderstood the time line used to present information and record predictions, or the cost of time or error in at least one condition.[‡‡]

2.6.2   Results

Our participants answered the anchoring questions correctly in 74.8% of the trials. As in Experiment 1, people's predictions were systematically biased: Our participants significantly overestimated delays smaller than 8 min (all $p < 10^{-11}$) and significantly underestimated delays larger than 13 min (all $p < 10^{-4}$); see Figure 2.23. Furthermore, the biases were shifted upwards when the anchor was high compared to when the anchor was low ($z = 7.26, p < 10^{-12}$; see Figure 2.23). This effect was also evident in our participants' average predictions: when the anchor was high, then participants predicted significantly later departures than when the anchor was low: 12.06 ± 0.29 min

---

[‡‡]This exclusion rate would be high in a laboratory experiment, but it is not unusual for long online experiments run on Amazon Mechanical Turk.

versus 10.03 ± 0.15 min ($t(3438) = 6.16, p < 10^{-15}$). To estimate our participants' anchors and quantify their adjustments, we applied the linear regression model described above (Equation 2.6). Overall, our participants' apparent anchor was significantly higher in the high anchor condition (12.69 min) than in the low anchor condition (9.74 min, $p < 10^{-15}$). Our participants' adjustments away from the anchor tended to be small: on average, our participants adjusted their estimate only 29.86% of the distance from the anchor to the correct value when the anchor was low (95% CI: [26.38%; 30.85%]) and 27.25% of this distance when the anchor was high (95% CI: [24.00%; 30.50%]). Thus the relative adjustments were significantly smaller than in Experiment 1 (95% CI: [38.52%, 44.04%]) and they did not differ between the high and low anchor condition ($z = 1.16; p = 0.12$). Thus the linear relationship between the bias and the true delay and difference between the biases for the high versus the low anchor (Figure 2.23) may result from insufficient adjustment away from different anchors. This also explains why the average predictions were higher in the high anchor condition than in the low anchor condition.

Next, we investigated whether people adapted their prediction strategy to the experiment's incentives for speed and accuracy. To get a first impression, we performed a 2-factorial, repeated-measures ANOVA of the prediction errors' absolute values, and the ANOVA models included only the main effects of time cost and error cost and their interaction (fixed effects) and the main effect of participant number (random effect). This analysis confirmed that error cost made our participants' estimates significantly more accurate ($F(1, 3391) = 12.33, p < 0.0001$), but the effect of time cost was not statistically significant ($F(1, 3391) = 1.81, p = 0.185$) and neither was its interaction with the effect of error cost ($F(1, 3391) = 0.0027, p = 0.9583$). [§§] Next we assessed whether the amount by which participants adjusted their initial estimate increased with error cost and decreased with time cost. To answer this question we performed a repeated-measures ANOVA of relative adjustment as a function of time cost and error cost. To be precise, we first estimated each participant's relative adjustment separately for each of the four conditions and the two anchors using our linear regression model of anchoring and adjustment (Equation 2.6). We then performed an ANOVA on the estimated relative adjustments with the factors time cost, error cost, and high vs. low anchor (fixed-effects) as well as participant number (random effect) and the interaction effect of time cost and error cost; see Table 2.4. We found that time cost significantly reduced relative adjustment from 37.2% to 28.2% ($F(1, 297) = 15.5, p = 0.0001$) whereas error cost significantly increased it from 31.2% to 34.2% ($F(1, 297) = 10.39, p = 0.0014$), and the interaction was non-significant. These

---

[§§]Unfortunately, we cannot report an analysis of the reaction times, because they were not measured in the conditions without time cost due to programming error.

findings are consistent with the prediction of our resource-rational theory that the number of adjustments decreases with time cost but increases with error cost regardless of the anchor. The mean relative adjustments of each condition are shown in Table 2.5. Figure 2.24 shows the effects of incentives for speed and accuracy on the anchoring bias in the provided anchors experiment; note that the slope of each line is 1 minus the relative size of the adjustments in the corresponding condition. As predicted by our theory (cf. Figure 2.14) and observed for self-generated anchors (cf. Figure 2.19), the slope of the anchoring bias was largest when time cost was high and errors were not penalized. Table 2.5 summarizes the relative adjustments sizes in the four incentive conditions.

Table 2.4: ANOVA of relative adjustment as a function of time cost and error cost.

Source	d.f.	Sum Sq.	Mean Sq.	F	p
error cost	1	1.0318	1.03178	15.5	0.0001
time cost	1	0.6912	0.69115	10.39	0.0014
subject	42	11.3544	0.27034	4.06	10^{-12}
anchor (high vs. low)	1	0.0774	0.07739	1.16	0.2817
error cost $\times$ time cost	1	0.1066	0.10659	1.6	0.2066
Error	297	19.7643	0.06655		
Total	343	33.0256			

Table 2.5: Relative size of our participants' adjustments of their initial guess towards the correct answer by incentive condition in the experiment with provided anchors with 95% confidence intervals.

	No Error Cost	High Error Cost
No Time Cost	$30.0 \pm 7.4\%$	$44.4 \pm 8.4\%$
High Time Cost	$24.5 \pm 6.5\%$	$32.0 \pm 9.1\%$

2.6.3 TESTING MODELS OF THE ANCHORING BIAS

Consistent with the biases and the effects of time cost and error cost, we found that the two adaptive anchoring-and-adjustment models explained our participants' predictions significantly better than any of the alternative models; see Figure 2.25, top panel. Concretely, the first adaptive anchoring-and-adjustment model (m_{aAA}) was the best explanation for 36.9% of our participants, and the adaptive anchoring-and-adjustment model with an additional intrinsic error cost parameter (m_{aAAi}) was the best explanation for another 24.8% of our participants. Thus for the majority

of our participants, responses were best explained by adaptive anchoring-and-adjustment. Further-more, we can be 85.9% confident that the first adaptive anchoring-and-adjustment is the best model for a larger percentage of people than any of the alternative models. In addition to this random effects analysis, we also ran a Bayesian fixed-effects analysis by computing the group Bayes factors. This analysis confirmed that the two adaptive anchoring-and-adjustment models explain the data substantially better than any of the alternatives, but among these two models it strongly favored the more complex model with intrinsic error cost: According to the posterior-odds ratios this model is at least 10^{30} times as likely as any other model we considered. In conclusion, we found that most participants performed adaptive anchoring-and-adjustment (m_{aAA} and m_{aAAi}) and while the con-tribution of the intrinsic error cost is negligible in many participants it is crucial in others. Next, we asked which theory best explains our participants' responses; see Figure 2.25, bottom left panel. Ac-cording to family-level Bayesian model selection, we can be 99.99% confident that anchoring-and-adjustment is the most probable explanation for a significantly larger proportion of people (76.9%) than either posterior probability matching (10.6%), Bayesian decision theory (10.4%), or random choice (2.1%). Furthermore, we can be 98.5% confident that for the majority of people (67.6%) our adaptive models' predictions are more accurate than the predictions of their non-adaptive coun-terparts; see Figure 2.25, bottom right panel.

2.6.4 Discussion

Our participants' predictions were significantly biased towards the provided anchors. When the anchor was high, their predictions and biases were shifted upwards compared to when it was low. This bias increased linearly with the distance from the anchor to the correct value. Furthermore, this experiment also confirmed our second prediction: the bias towards the provided anchor de-creased with error cost but increased with time cost (compare Figures 2.14 and 2.24). Thus the bias towards the provided anchors and the effects of time cost and error cost were qualitatively the same as with self-generated anchors (Figure 2.19). Contrary to the claims by Epley and Gilovich (2006), our results suggest that anchoring-and-adjustment is sufficient to explain the anchoring bias towards provided as well as self-generated anchors.

While time cost had an effect on the imputed number of adjustments, the effect of time cost on absolute error was not statistically significant. This might have been because the timer started three seconds after the anchoring question and the number line were presented. Our rationale was to en-sure that our participants encode the anchor before predicting the departure time and we found that

it takes about three seconds to read, think about, and answer the anchoring question. However, this change might have reduced the time pressure experienced by our participants and thereby diminished the effect of time cost on accuracy relative to Experiment 1.

Interestingly, our model-based analysis suggested that our participants' effective anchors were less extreme than the values we provided. One possible reason is that people after sometimes discard the provided anchor and generate their own anchor instead after having stated that the provided anchor is too high or too low. Having stated the direction in which the correct value deviates from the anchor could potentially also increase people's propensity to make adjustments consistent with this judgment. Since our participants' direction judgments were mostly correct, this effect would increase adjustment, but adjustments were smaller than in Experiment 1. However, it is also conceivable that our analysis picked up this omnipresent additional adjustment as a shift in the anchor.

Despite the qualitative commonalities between the results of our two experiments with self-generated versus provided anchors, there were quantitative differences: In three of the four conditions, our participants' adjustments were significantly smaller for provided anchors than for self-generated anchors. There are at least two possible complementary explanations: First, self-generated anchors are probably much more variable than the initial guesses elicited by provided anchors, and the anchoring biases towards high versus low self-generated anchors might cancel each other out. Second, people probably treat provided anchors not only as initial guesses but also as conversational hints that the correct value is close to the provided anchor (Y. C. Zhang & Schwarz, 2013). Based on this hint people may either strategically decrease the number of adjustments or assign a higher plausibility to estimates close to the provided anchor. The latter could be modeled as a Bayesian inference from the hint on the to-be-predicted value, but this rational inference alone would be insufficient to account for our data because the effect of the anchor type disappeared when time cost was high and error cost was zero (cf. Table 2.3 with Table 2.5).

Thus, resource-rational anchoring-and-adjustment is a promising process model of numerical estimation. It can explain the plethora of anchoring effects summarized in Table 2.1 from empirically supported first principles: probabilistic inference by an iterative (sampling) algorithm and optimal resource-allocation. The resulting models enable new insights into old and new empirical phenomena.

2.7 General Discussion

Anchoring and adjustment is one of the classic heuristics reported by Tversky and Kahneman (1974) and it seems hard to reconcile with rational behavior. In this article, we have argued that this heuristic can be understood as a signature of resource-rational information processing rather than a sign of human irrationality. We have supported this conclusion by a resource-rational analysis of numerical estimation, simulations of anchoring phenomena with a resource-rational process model, two novel experiments that confirmed the predictions of our rational account of anchoring, and quantitative model comparisons against alternative explanations of anchoring. We showed that anchoring-and-adjustment can be interpreted as a Markov chain Monte Carlo algorithm–a rational approximation to rational inference. We found that across many problems the optimal speed-accuracy tradeoff of this algorithm entails performing so few adjustments that the resulting estimate is biased towards the anchor. Our simulations demonstrated that resource-rational anchoring-and-adjustment, which adaptively chooses the number of adjustments to maximize performance net the cost of computation, provides a unifying explanation for ten different anchoring phenomena (see Table 2.1). Finally, our experiments confirmed that people rationally adapt the number of adjustments to the relative cost of time.

Although we explored the implications of limited time and finite cognitive resources assuming an abstract computational architecture based on sampling, the results of our mathematical analysis are more general and the algorithms we derived illustrate general properties of resource-rational information processing. Other iterative inference mechanisms such as (stochastic) gradient descent, variational Bayes, predictive coding (Friston, 2009; Friston & Kiebel, 2009), and probabilistic computation in cortical microcircuits (Habenschuss et al., 2013) also have the property of diminishing returns for additional computation. Therefore the qualitative predictions shown in Figures 2.3–2.6 are valid now only for the abstract computational architecture that we chose to analyze but characterize bounded rationality for a more general class of cognitive architectures. Therefore, our results support the adaptive allocation of finite computational resources and the resource-rationality of bias regardless of the specific cognitive mechanism that people use to draw inferences.

We discuss the implications of our results for general theoretical questions. We start with the conclusion that people use anchoring-and-adjustment more widely than previously assumed, that is they adjust not only from self-generated anchors but also from provided anchors. Next, we discuss how our model is related to previous theories of anchoring and how they can be integrated into our resource-rational framework. We then turn to two questions about rationality: First, we discuss

existing evidence for the hypothesis that anchors are chosen resource-rationally and how it can be tested in future experiments. Second, we argue that resource-rationality, the general theory we have applied to explain the anchoring bias, provides a more adequate normative framework for cognitive strategies than classical notions of rationality. We close with directions for future research.

2.7.1 People adjust from provided and self-generated anchors

In contrast to most heuristics, anchoring-and-adjustment is a very flexible strategy. It can be quick and biased by performing only a few adjustments, or accurate and slow by performing many adjustments. Intuitively, people should perform more adjustments and be less biased when they are motivated to be accurate. Therefore, the reduction of the bias with financial incentives has been used to operationalize anchoring-and-adjustment: Epley and Gilovich (2005) found no evidence that the bias towards a provided anchor decreases with financial incentives and concluded that therefore people use anchoring-and-adjustment only with self-generated but not with provided anchors. By contrast, in our experiments financial incentives increased the number of adjustments regardless of whether the anchor was self-generated (Experiment 1; Figure 2.19) or provided (Experiment 2; Figure 2.24). How is this finding compatible with previous studies in which financial incentives failed to reduce the anchoring bias in Epley and Gilovich (2005); Tversky and Kahneman (1974)? According to our simulations and empirical data, the reason is that people know much less about the quantities for which Epley and Gilovich (2005) decided to provide anchors than for those for which people were found to generate their own anchors. In our experiments with self-generated versus provided anchors we eliminated the confounding effect of uncertainty by having people estimate the same quantities with and without being provided an anchor. Consistent with Simmons et al. (2010), we found that the anchoring bias decreased with financial incentives regardless of whether we provided an anchor or not. Thus our results suggest that resource-rational anchoring-and-adjustment is a unifying mechanisms for the anchoring biases observed for self-generated as well as provided anchors. Our simulations show that this conclusion is compatible with the results reviewed by Epley and Gilovich (2006), because the effect of financial incentives declines with the uncertainty about the quantity to be estimated. This explanation is similar to the argument by Simmons et al. (2010), but our formal model does not need to assume that people reason about the direction of their adjustments. Last but not least, our findings suggest that incentives are more effective at debiasing than previously thought as long as people are sufficiently knowledgable.

2.7.2 RELATION TO PREVIOUS THEORIES OF ANCHORING AND ADJUSTMENT

Previous models of anchoring-and-adjustment (Epley & Gilovich, 2006; Simmons et al., 2010) assumed that adjustment terminates when the plausibility of the current estimate exceeds a threshold. Here we formalized this idea by the anchoring-and-adjustment model with a simple stopping rule (m_{AAs}, Equations A.9-A.12). Importantly, this model was not supported by our experimental data; see Figures 2.21 and 2.25. Instead, our data supported adaptive anchoring-and-adjustment according to which the number of adjustments is chosen in advance such as to optimize the strategy's expected speed-accuracy tradeoff. From an information processing perspective the limitation of models postulating that adjustment stops when plausibility exceeds a threshold is that there is no single threshold that works well across all estimation problems. Depending on the level of uncertainty successful estimation requires different thresholds. A threshold that is appropriate for low uncertainty will result in never-ending adjustment in a problem with high uncertainty. Conversely, a threshold that is appropriate for a problem with high uncertainty would be too liberal when the uncertainty is low. In addition, Simmons et al. (2010) postulate that people reason about the direction of their adjustment whereas resource-rational anchoring-and-adjustment does not. It would be interesting to see whether an extension of our model that incorporates directional information would perform better in numerical estimation and better predict human behavior. We will return to this idea when we discuss directions for future research.

According to the selective-accessibility theory of anchoring (Strack & Mussweiler, 1997), comparing an unknown quantity to the provided anchor increases the accessibility of anchor-consistent knowledge and the heightened availability of anchor-consistent information biases people's estimates. There is no quantitative mathematical model of selective accessibility that could be tested against our resource-rational anchoring-and-adjustment model using the data we have collected. The evidence that some anchoring biases result from selective accessibility (Strack & Mussweiler, 1997) does not undermine our analysis, because the existence of selective accessibility would not rule out the existence of anchoring-and-adjustment and vice versa. In fact, from the perspective of resource-rational probabilistic inference a mechanism similar to selective accessibility is likely to co-exist with anchoring-and-adjustment. Concretely, we have formalized the problem of numerical estimation of some quantity X as minimizing the expected error cost of the estimate $\hat{x}$ with respect to the posterior distribution $P(X|K)$ where K is the entirety of the person's relevant knowledge. This problem can be decomposed into two sub-problems: conditioning on relevant knowledge to evaluate (relative) plausibility and searching for an estimate with high plausibility. It appears unlikely that the mind can solve the first problem by simultaneously retrieving and instantly incorpo-

rating each and every piece of knowledge relevant to estimating X. Instead, the mind might have to sequentially recall and incorporate pieces $K^{(1)}, K^{(2)}, K^{(3)}, \cdots$ of its knowledge to refine $P(X)$ to $P(X|K^{(1)})$ to $P(X|K^{(1)}, K^{(2)})$ to $P(X|K^{(1)}, K^{(2)}, K^{(3)})$, and so forth. This process could be modeled as bounded using a sequential Monte Carlo algorithm (Doucet, De Freitas, & Gordon, 2001) and bounded conditioning (Horvitz, Suermondt, & Cooper, 1989).

Furthermore, it would be wasteful not to consider the knowledge that has been retrieved to answer the comparison question in the estimation task and impossible to retrieve all of the remaining knowledge. Selective accessibility may therefore result from the first process. Yet, regardless of how the first problem is solved, the mind still needs to search for an estimate $\hat{x}$ with high posterior probability, and this search process might be implemented by something like anchoring-and-adjustment. Furthermore, the knowledge retrieved in the first step might also guide the generation of an anchor. Importantly, both processes are required to generate an estimate. Therefore, we agree with (Simmons et al., 2010) that selective accessibility and anchoring-and-adjustment might coexist and both of them might contribute to the anchoring bias.

In summary, our resource-rational analysis of estimation sheds new light on classic notions of anchoring-and-adjustment (Epley & Gilovich, 2006; Tversky & Kahneman, 1974), explaining why they work and why people use them. Furthermore, our framework is sufficiently general to incorporate and evaluate the extensions proposed by Simmons et al. (2010) and Strack and Mussweiler (1997) and many others. Exploring these extensions is an interesting direction for future work.

2.7.3   ARE ANCHORS CHOSEN RATIONALLY?

Anchoring-and-adjustment has two components: generating an anchor and adjusting from it. Our experiments and simulations supported the conclusion that adjustment is resource-rational. Thus, a natural next question is whether anchors are also generated resource-rationally.

Self-generated anchors are usually close to the correct value, but provided anchors can be far off. For instance, it appears irrational that people can be anchored on their social security number when they estimate how much they would be willing to pay for a commodity (Ariely et al., 2003). Yet, the strategy failing people in this specific instance may nevertheless be resource-rational overall for at least four reasons: First, it is sensible to assume that the experimenter is reasonable and cooperative. Therefore her utterances should follow the Gricean maxims. Specifically, according to Grice's maxim of relation the stated anchor should be relevant (Y. C. Zhang & Schwarz, 2013). Furthermore, as a

rational information-seeking agent the experimenter should ask the question whose answer will be most informative. The most informative anchor to compare the true value to would be at the center of the experimenter's belief distribution. This too suggests that it is reasonable to treat the provided anchor as a starting point. Second, subsequent thoughts and questions are usually related. So it is reasonable to use the answer to a preceding question as the starting point for next thought. This holds for sequences of arithmetic operations such as $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ for which people anchor on their intermediate results when they are forced to respond early (Tversky & Kahneman, 1974) and in many other cases too. Third, when the provided anchor is the only number available in working memory, then using it may be faster and require less effort than generating a new one. Last but not least, one's beliefs may be wrong and the anchor may be more accurate. This was the case in Russo and Shoemaker's experiment: People overestimated the year in which Attila the Hun was defeated in Europe so much that the anchor was usually closer to the correct value (A.D. 451) than the mean of their unbiased estimates (A.D. 953.5). For these reasons, the observation that people anchor on irrelevant values provided in psychological experiments does not imply that anchors are selected irrationally. Anchor selection could be well adapted to the real-world. Consequently, anchoring biases in everyday reasoning would be much more benign than those observed in the laboratory. This is probably true, because most anchoring experiments violate people's expectation that the experimenter will provide relevant information, provide negligible incentives for accuracy, and ask people to estimate quantities about which they know very little.

There also is empirical evidence suggesting that people do not always use the provided value as their anchor. For instance, our model-based analysis of Experiment 2 suggested that people's effective anchors were less extreme than the provided values. This suggests that our participants did not always use the provided number as their anchor. Furthermore, in the experiment by Strack and Mussweiler (1997) the provided anchor influenced the participants' estimates only when it was semantically related to the quantity to be estimated. Pohl (1998) found that the anchoring bias was absent when the anchor was perceived as implausible, and Hardt and Pohl (2003) found that the bias was smaller on trials where the anchor's judged plausibility was below the median plausibility judgment. Thus, at least under some circumstances, people appear to discard the provided value when it appears irrelevant or misleading.

However, realizing that the provided anchor is implausible and generating a better anchor require knowledge, effort, and time. Therefore, when people are asked to estimate a quantity they know almost nothing about, it may be resource-rational for them to anchor on whatever the experimenter suggested. This seems applicable to most anchoring experiments, because participants are usually

so uncertain that they do not even know in which direction to adjust from the provided anchor (Simmons et al., 2010). If you cannot even tell whether the correct value is larger or smaller than the anchor, how could you generate a better one? The effect of the anchor is largest in people with little knowledge and high uncertainty about the quantity to be estimated (Jacowitz & Kahneman, 1995; Wilson et al., 1996). These people would benefit from a better anchor, but they cannot easily generate one, because they lack the relevant knowledge. Conversely, our simulation of the effect of knowledge suggested that people knowledgeable enough to generate good anchors, will perform well even if they start from a highly implausible anchor. Although this argument is speculative and has yet to be made precise it suggests that, at least in some situations, self-generating an anchor might not be worth the effort regardless of one's knowledge.

In conclusion, existing data are not necessarily inconsistent with the idea that anchors are chosen resource-rationally. Thus, whether anchors are chosen rationally is still an open question. Experimental and theoretical approaches to this question are an interesting avenue for future research that we will discuss below.

The experiments reported in this chapter provide further support for resource-rationality as a descriptive theory of human cognition. Previous experiments supported the prediction of resource-rationality that mental algorithms tolerate bias in exchange for speed when accuracy is not crucial (Lieder, Goodman, & Griffiths, 2013; Lieder et al., 2012). Here we went one step further and tested whether the human mind rationally allocates its computational resources according to the utility of being accurate and the cost of time. Our empirical data confirmed this prediction. This is in line with the finding of near-optimal speed-accuracy tradeoffs in perceptual decision-making (Bogacz et al., 2010). The key difference is that we studied the control of reasoning whereas Bogacz et al. (2010) studied the collection of sensory information. Resource-rationality is a general framework applicable to all cognitive abilities. Even though resource-rationality is a very recent approach, it has already shed some light on a wide range of cognitive abilities and provides a unifying framework for the study of intelligence in psychology, neuroscience, and artificial intelligence (Gershman et al., 2015). For example, we have recently applied the resource-rational framework to decision-making (Lieder, Hsu, & Griffiths, 2014), planning (Lieder, Goodman, & Huys, 2013), and strategy selection (Lieder & Griffiths, 2015; Lieder, Plunkett, et al., 2014). In conclusion, resource-rationality appears to be a promising framework for normative and descriptive theories of human cognition.

2.7.4 Directions for future research

The question to which extent anchors are chosen resource-rationally is one interesting avenue for future research. The hypothesis that anchors are chosen rationally predicts that if everything else is equal people will choose a relevant anchor over an irrelevant one. This could be probed by providing people with two anchors rather than just one. Alternatively, one could manipulate the ease of self-generating a good anchor and test whether this ease decreases the bias towards an implausible provided anchor. To analyze such experiments, the models developed could be used to infer which anchor people were using from the pattern of their responses.

Future studies could also leverage people's reaction times to further test whether the number of iterations is predetermined before adjustment begins against the alternative hypothesis that people decide whether or not to make another adjustment based on the plausibility of the current estimate as assumed by earlier theories (Epley & Gilovich, 2006; Simmons et al., 2010). Our model also predicts a multiplicative interaction between opportunity cost and error cost such that the anchoring bias is proportional to the ratio of time cost over error cost. Qualitatively, this means that the effect of error cost should increase with opportunity cost and the effect of opportunity cost should increase with time cost. However, when both are increased or decreased by the same factor, then the anchoring bias should remain constant.

An additional direction for future work is to extend the adaptive anchoring-and-adjustment model. This could be done in several ways. First, the model could be extended by mechanisms for choosing and generating anchors. Second, the model could be extended by specifying *how* the mind approximates optimal resource allocation. A third extension of our models might incorporate directional information into the proposal distribution as in the Hamiltonian Monte Carlo algorithm (Neal, 2011) to better capture the effects of direction uncertainty discovered by Simmons et al. (2010). A fourth extension might capture the sequential incorporation of relevant knowledge by iterative conditioning and explore its connection to the selective accessibility theory of the anchoring bias (Strack & Mussweiler, 1997). A fifth frontier is to make resource-rational anchoring-and-adjustment more adaptive: How can the proposal distribution and a mechanism for choosing the number of adjustments be learned from experience? Can better performance be achieved by adapting the proposal distribution from one adjustment to the next? Finally, our resource-rational anchoring-and-adjustment only uses a single sample, but it can be generalized to using multiple samples. Each of these extensions might improve the performance of the estimation strategy and it is an interesting question of whether or not those improvements would bring its predictions closer to human behav-

ior. Future studies might also evaluate additional alternatives to our model, such as an anchoring model with adaptive plausibility threshold or algorithms that directly approximate the most probable estimate rather than a sample from the posterior distribution.

Most previous models of heuristics are formulated for the domain in which the corresponding bias was discovered. For instance, previous models of anchoring-and-adjustment were specific to numerical estimation (Epley & Gilovich, 2006; Simmons et al., 2010). Yet, everyday reasoning is not restricted to numerical estimation and anchoring also occurs in very different domains such as social cognition (Epley et al., 2004). This highlights the challenge that models of cognition should be able to explain not only what people do in the laboratory but also their performance in the real-world. Heuristics should therefore be able to operate on the complex, high-dimensional semantic representations people use in everyday reasoning. Resource-rational anchoring-and-adjustment lives up to this challenge, because Markov-chain Monte Carlo methods are as applicable to semantic networks (Abbott, Austerweil, & Griffiths, 2012) as they are to single numbers. In fact, resource-rational anchoring-and-adjustment is a very general mechanism that can operate over arbitrarily complex representations and might be deployed not only for numerical estimation but also in many other cognitive faculties such as memory retrieval, language understanding, social cognition, and creativity. For instance, resource-rational anchoring-and-adjustment may be able to explain the hindsight bias in memory recall (Hardt & Pohl, 2003; Pohl, 1998), primacy effects in sequential learning (Abbott & Griffiths, 2011), and the dynamics of memory retrieval (Abbott et al., 2012; Bourgin, Abbott, Griffiths, Smith, & Vul, 2014).

2.7.5   CONCLUSION

Resource-rational anchoring-and-adjustment provides a unifying, parsimonious, and principled explanation for a plethora of anchoring effects including some that were previously assumed to be incompatible with anchoring-and-adjustment. Interestingly, we discovered this cognitive strategy purely by applying resource-rational analysis to estimation under uncertainty. It is remarkable that the resulting model is so similar to the anchoring-and-adjustment heuristic. Our experiments confirmed that people rationally adapt the number of adjustments to the environment's incentives for speed and accuracy. Resource-rational anchoring-and-adjustment thereby reconciles the anchoring-bias with people's adaptive intelligence and Bayesian models of reasoning under uncertainty. Concretely, the anchoring bias may reflect the optimal speed-accuracy tradeoff when errors are benign, which is true of most, if not all, laboratory tasks. Yet, when accuracy is important and speed is not

crucial, then people perform more adjustments and the anchoring bias decreases. In conclusion, the anchoring bias may be a window on resource-rational computation rather than a sign of human irrationality. Being biased can be resource-rational, and heuristics can be discovered by resource-rational analysis.
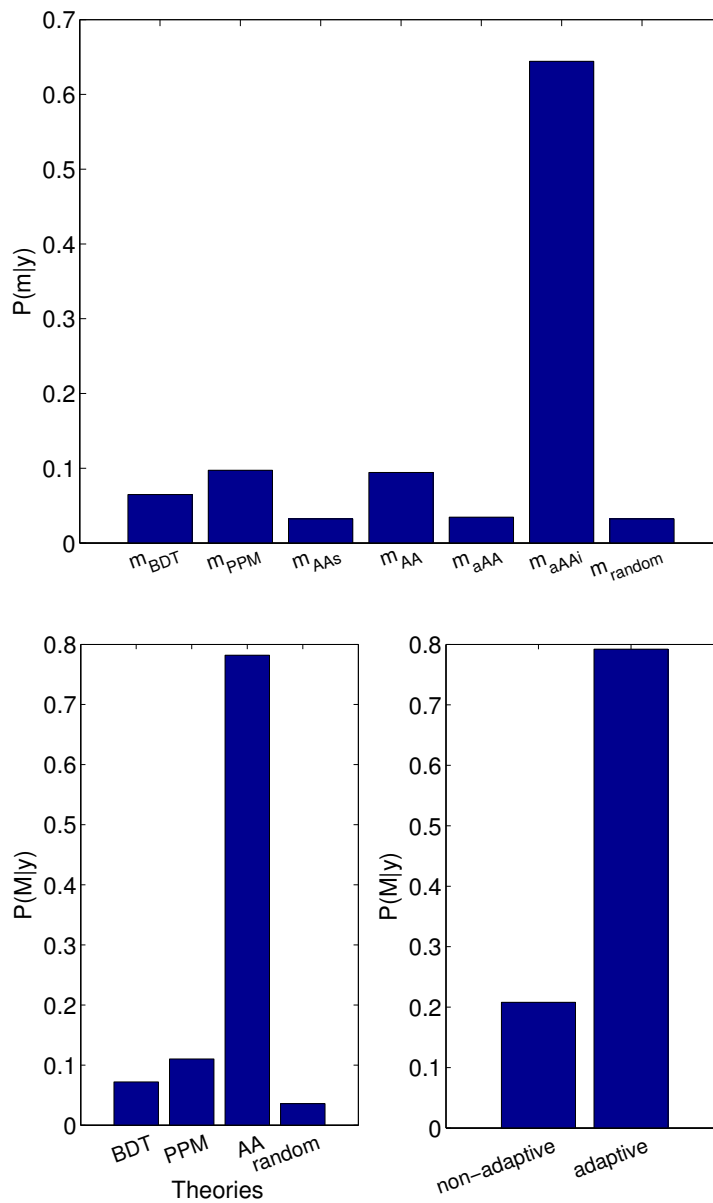
Figure 2.21: Results of Bayesian model selection given the data from Experiment 1. The top panel shows the posterior probabilities of individual models. The bottom left panel shows the posterior probabilities of the four theories (BDT: Bayesian decision theory, PPM: posterior probability matching, AA: anchoring-and-adjustment, random: predictions are chosen randomly). The bottom right panel shows the posterior probabilities of adaptive versus non-adaptive models.
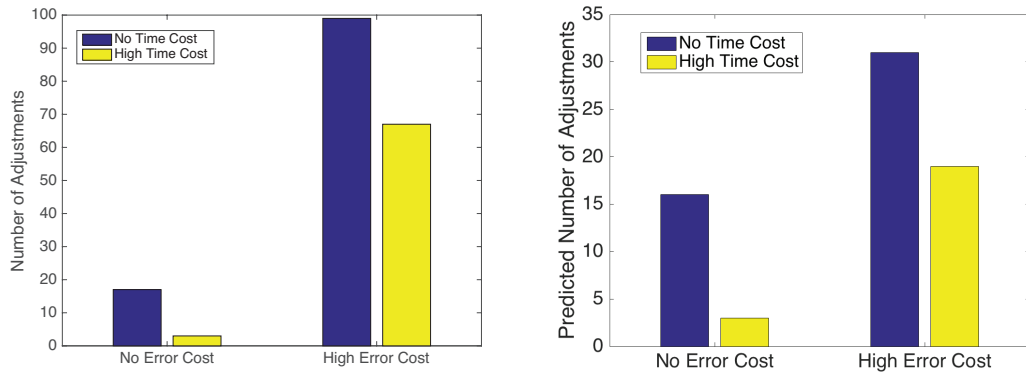
Figure 2.22: Estimated (left panel) and predicted (right panel) number of adjustments.
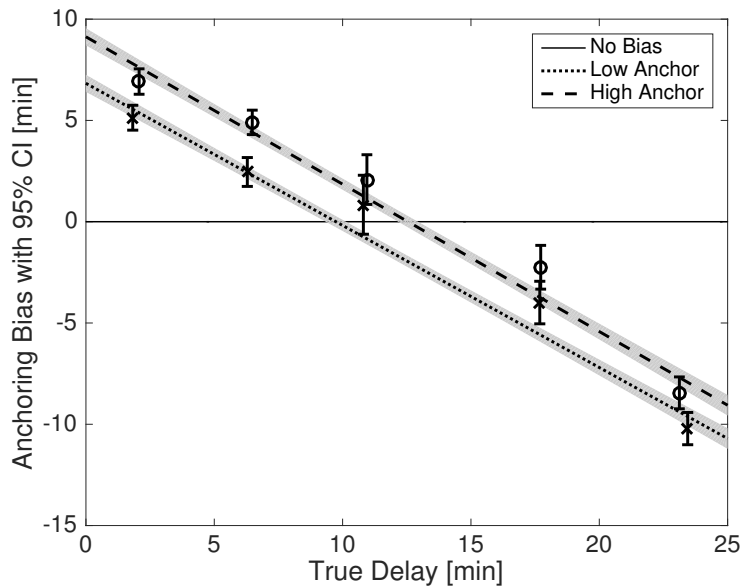
Figure 2.23: Biases when the provided anchor was high versus low. Solid lines show the results of linear regression. Shaded areas are 95% confidence bands, the diamonds with error bars are the average biases within a five minute window and their 95% confidence intervals; that is ±1.96 SEM.
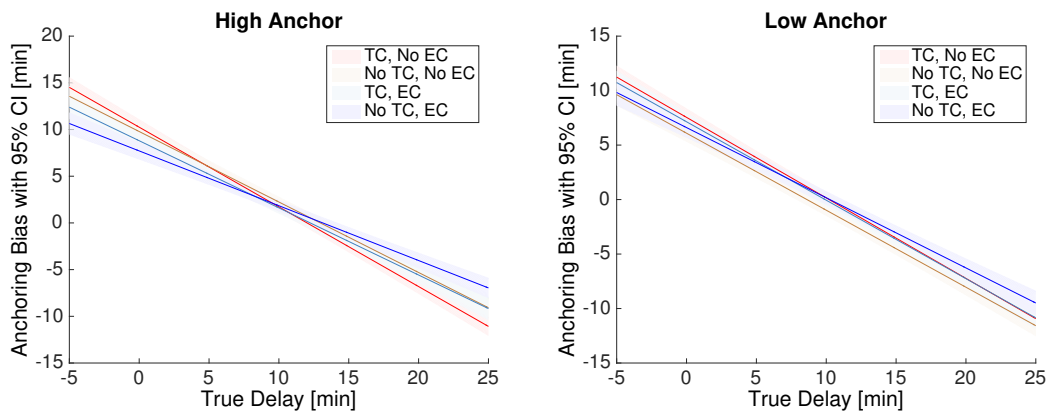
Figure 2.24: Effect of incentives for speed and accuracy when a high anchor was provided confirm our theory's prediction; cf. Figure 2.14.
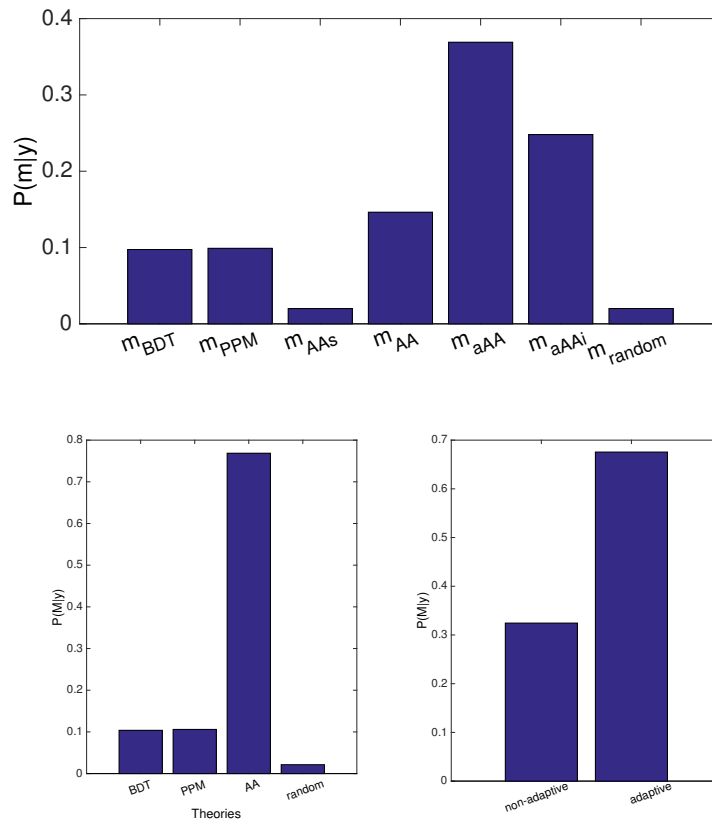
Figure 2.25: Model selection results for Experiment 2. The top panel shows the posterior model probabilities. The bottom panel shows the results of Bayesian inference on the level of model families.

3

A resource-rational perspective on availability biases[*]

In addition to the anchoring bias analyzed in the previous chapter, Tversky and Kahneman's ground-breaking paper "Judgment under uncertainty: heuristics and biases" (Tversky & Kahneman, 1974) reported another way in which human judgment violates the laws of probability theory: The *availability bias* is the phenomenon that people overestimate the probability of events that come to mind easily (Tversky & Kahneman, 1973). It leads people to overestimate the frequency of extreme events (Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978) which in turn contributes to overreactions to the risk of terrorism (Sunstein & Zeckhauser, 2011) and other threats (Lichtenstein et al., 1978; Rothman, Klein, & Weinstein, 1996). Such availability biases result from the fact that not all memories are created equal: while most unremarkable events are quickly forgotten, the strength of a memory increases with the magnitude of its positive or negative emotional valence (Cruciani, Berardi, Cabib, & Conversi, 2011). This may be why memories of extreme events, such as a traumatic car accident (Brown & Kulik, 1977; Christianson & Loftus, 1987) or a big win in the casino, come to mind much more easily (Madan, Ludvig, & Spetch, 2014) and affect people's decisions more strongly (Ludvig, Madan, & Spetch, 2014) than moderate events, such as the 2476th time you drove

---

[*]This chapter is based on Lieder, Griffiths, and Hsu (2017).

home safely and the 1739$^{\text{th}}$ time a gambler lost \$1 (Thaler & Johnson, 1990).

The availability bias is commonly assumed to be irrational, but here we propose that it might reflect the rational use of finite time and limited cognitive resources (Griffiths et al., 2015). This chapter explores the implications of these bounded resources within the resource-rational framework introduced in Chapter 1. According to our mathematical analysis, the availability bias could serve to help decision-makers focus their limited resources on the most important eventualities. In other words, we argue that the overweighting of extreme events ensures that the most important possible outcomes (i.e., those with extreme utilities) are always taken into account even when only a tiny fraction of all possible outcomes can be considered. Concretely, we show that maximizing decision quality under time constraints requires biases compatible with those observed in human memory, judgment, and decision-making. Without those biases the decision-maker's expected utility estimates would be so much more variable that her decisions would be significantly worse. This follows directly from a statistical principle known as the *bias-variance tradeoff* (Hastie, Tibshirani, & Friedman, 2009).

Starting from this principle, we derive a rational process model of memory encoding, judgment, and decision making that we call *utility-weighted learning* (UWL). Concretely, we assume that the mind achieves a near-optimal bias-variance tradeoff by approximating the *optimal importance sampling* algorithm (Geweke, 1989; Hammersley & Handscomb, 1964) from computational statistics. This algorithm estimates the expected value of a function (e.g., a utility function) by a weighted average of its values for a small number of possible outcomes. To ensure that important potential outcomes are taken into account, optimal importance sampling optimally prioritizes outcomes according to their probability and the extremity of their function value. The resulting estimate is biased towards extreme outcomes but its reduced variance makes it more accurate. To develop our model, we apply optimal importance sampling to estimating expected utilities. We find that this enables better decisions under constrained resources. The intuitive reason for this benefit is that overweighting extreme events ensures that the most important possible outcomes (e.g., a catastrophe that has to be avoided or an epic opportunity that should be seized) are always taken into account even when only a tiny fraction of all possible outcomes can be considered.

According to our model, each experience o creates a memory trace whose strength w is proportional to the extremity of the event's utility $u(o)$ (i.e., $w = |u(o) - \bar{u}|$ where $\bar{u}$ is a reference point established by past experience). This means that when a person experiences an extremely bad event (e.g., a traumatic accident) or an extremely good event (e.g., winning the jackpot) the resulting memory trace will be much stronger than when the utility of the event was close to zero (e.g., lying in bed

and looking at the ceiling). Here, we refer to events such as winning the jackpot and traumatic car accidents as 'extreme' not because they are rare or because their utility is far from zero but because they engender a large positive or large negative *difference* in utility between one choice (e.g., to play the slots) versus another (e.g., to leave the casino).

In subsequent decisions (e.g., whether to continue gambling or call it a day), the model probabilistically recalls past outcomes of the considered action (e.g., the amounts won and lost in previous rounds of gambling) according to the strengths of their memory traces. As a result, the frequency with which each outcome is recalled is biased by its utility even though the recall mechanism is oblivious to the content of each memory.

Concretely, the probability that the first recalled outcome is an instance of losing \$1 would be proportional to the sum of its memory traces' strengths. Although this event might have occurred very frequently, each of its memory traces would be very weak. For instance, while there might be 1345 memory traces their strengths would be small (e.g., $|u(-\$1) - \bar{u}|$ with $\bar{u}$ close to $u(-\$1)$). Thus, the experience of losing \$1 in the gamble would be only moderately available in the gambler's memory (total memory strength $1345 \cdot |u(-\$1) - \bar{u}|$). Therefore, the one time when the gambler won \$1000 might have a similarly high probability of coming to mind because its memory trace is significantly stronger (e.g., one memory trace of strength $|u(\$1000) - \bar{u}|$). According to our model, this probabilistic retrieval mechanism will *sample* a few possible outcomes from memory. These simulated outcomes (e.g., $o_1 = \$1000, o_2 = \$ - 1, \cdots, o_5 = \$1000$) are then used to estimate the expected utility of the considered action by a weighted sum of their utilities where the theoretically derived weights partly correct for the utility-weighting of the memory traces (i.e., $\hat{U} = \sum_i w_i \cdot u(o_i)$ with $w_i = \frac{1}{|u(o_i) - \bar{u}|}$). Finally, the considered action is chosen if and only if the resulting estimate of the expected utility gain is positive.

Our model explains why extreme events come to mind more easily, why people overestimate their frequency, and why they are overweighted in decision-making. It captures published findings on biases in memory recall, frequency estimation, and decisions from experience (Erev et al., 2010; Ludvig et al., 2014; Madan et al., 2014) as well as three classic violations of expected utility theory in decisions from description. Our model is competitive with the best existing models of decisions from experience and correctly predicted the previously unobserved correlation between events' perceived extremity and the overestimation of their frequencies. The empirical evidence that we present strongly supports the model's assumption that the stronger memory encoding of events with extreme utilities causes biases in memory recall that in turn lead to biases in frequency estimation and decision-making. Concretely, people remember extreme events more frequently than equally frequent events

of moderate utility, overestimate their frequency, and overweight them in decision-making (Ludvig et al., 2014). Furthermore, the magnitude of overweighting increases significantly with the magnitude of the memory bias (Madan et al., 2014), and we found that the extent to which people overestimate an event's frequency correlates significantly with its extremity. The theoretical significance of our analysis is twofold: it provides a unifying mechanistic and teleological explanation for a wide range of seemingly disparate cognitive biases and it suggests that at least some heuristics and biases might reflect the rational use of finite time and limited cognitive resources (Griffiths et al., 2015).

The remainder of this chapter proceeds as follows: We start by deriving a novel decision mechanism as the rational use of finite time under reasonable, abstract assumptions about the mind's computational architecture. We show that the derived mechanism captures people's availability biases in frequency judgment and memory recall. Next, we demonstrate that the same mechanism can also account for three classic violations of expected utility theory and evaluate it against alternative models of decisions from description. We proceed to show that our model can also capture the heightened availability, overestimation, and overweighting of extreme events in decisions from experience. Finally, we show that utility-weighted sampling can emerge from a biologically-plausible learning mechanism that captures the temporal evolution of people's risk preferences in decisions from experience and evaluate it against alternative models of decisions from experience. We conclude with implications for the debate on human rationality and directions for future research.

3.1 Resource-rational decision-making by utility-weighted sampling

According to expected utility theory (von Neumann & Morgenstern, 1944), decision-makers should evaluate each potential action a by integrating the probabilities $P(o|A = a)$ of its possible outcomes o with their utilities $u(o)$ into the action's *expected utility* $\mathbb{E}_{p(O|A=a)}[u(O)]$. Unlike simple laboratory tasks where each choice can yield only a small number of possible payoffs, many real-life decisions have infinitely many possible outcomes.[†] As a consequence, the expected utility of action a becomes an integral:

$$\mathbb{E}_{p(O|A=a)}[u(O)] = \int p(o|a) \cdot u(o) \, do. \tag{3.1}$$

[†]People often cope with this complexity by partitioning possible outcomes into chunks like "stock goes up" vs. "stock goes down". We do not consider this approximation to be an inherent component of the problem itself, but rather as useful component of many heuristic strategies.

In the general case, this integral is intractable to compute. Below we investigate how the brain might approximate the solution to this intractable problem.

3.1.1   Sampling as a decision strategy

To explore the implications of resource constraints on decision-making under uncertainty, we model the cognitive resources available for decision-making within a formal computational framework that has been successfully used to develop rational process models of human cognition and can capture the variability of human performance, namely *sampling* (Griffiths, Vul, & Sanborn, 2012). Sampling methods can provide an efficient approximation to integrals such as the expected utility in Equation 3.1 (Hammersley & Handscomb, 1964), and mental simulations of a decision's potential consequences can be thought of as samples. The idea that the mind handles uncertainty by sampling is consistent with neural variability in perception (Fiser et al., 2010) and the variability of people's judgments (Denison et al., 2013; Griffiths & Tenenbaum, 2006; Vul et al., 2014). For instance, people's predictions of an uncertain quantity X given partial information y are roughly distributed according to its posterior distribution $p(X|y)$ as if they were sampled from it (Griffiths & Tenenbaum, 2006; Vul et al., 2014). Such variability has also been observed in decision-making: in repeated binary choices from experience animals chose each option stochastically with a frequency roughly proportional to the probability that it will be rewarded (Herrnstein & Loveland, 1975). This pattern of choice variability, called *probability matching*, is consistent with the hypothesis that animals perform a single simulation and chose the simulated action whenever its simulated outcome is positive. People also exhibit probability matching when the stakes are low, but as the stakes increase their choices transition from probability matching to maximization (Vulkan, 2000). This transition might arise from people gradually increasing the number of samples they generate to maximize the amount of reward they receive per unit time (Vul et al., 2014). Decision mechanisms based on sampling from memory can explain a wide range of phenomena (N. Stewart et al., 2006). Concordant with recent drift-diffusion models (Shadlen & Shohamy, 2016) and query theory (Johnson, Häubl, & Keinan, 2007; Weber et al., 2007), this approach assumes that preferences are constructed (Payne, Bettman, & Johnson, 1992) through a sequential, memory-based cognitive process.

Assuming that people make decisions by sampling, we can express time and resource-constraints as a limit on the number of samples, where each sample is a simulated outcome: According to our theory, the decision-maker's primary cognitive resource is a probabilistic simulator of the environment. The decision-maker can use this resource to anticipate some of the many potential futures

that could result from taking one action versus another, but each simulation takes a non-negligible amount of time. Since time is valuable and the simulator can perform only one simulation at a time, the cost of using this cognitive resource is thus proportional to the number of simulations (i.e. samples).

If a decision has to be based on only a small number of simulated outcomes, what is the optimal way to generate them? Intuitively, the rational way to decide whether to take action a is to simulate its consequences o according to one's best knowledge of the probability p that they will occur and average the resulting gain in utility $\Delta u(o)$ to obtain an estimate of $\Delta \hat{U}_s^{\text{RS}}(a)$ of the expected gain or loss in utility for taking action a over not taking it, that is

$$\Delta \hat{U}_s^{\text{RS}}(a) = \frac{1}{s} \sum_{i=1}^{s} \Delta u(o_i), \quad o_1, \cdots, o_s \sim p(O). \tag{3.2}$$

This decision strategy, which we call *representative sampling* (RS), generates an unbiased utility estimate. Yet – surprisingly – representative sampling is insufficient for making good decisions with very few samples. Consider, for instance, the choice between accepting versus declining a game of Russian roulette with the standard issue six-round NGant M1895 revolver. Playing the game will most likely, i.e. with probability $p_1 = \frac{5}{6}$, reward you with a thrill and save you some ridicule ($\Delta u(o_1) = 1$) but kill you otherwise ($p_2 = \frac{1}{6}$, $\Delta u(o_2) = -10^9$). Ensuring that representative sampling declines a game of Russian roulette at least 99.99% of the time, would require 51 samples – potentially a very time-consuming computation.

Like Russian roulette, many real-life decisions are complicated by an inverse relationship between the magnitude of the outcome and its probability (Pleskac & Hertwig, 2014). Many of these problems are much more challenging than declining a game of Russian roulette, because their probability of disaster is orders of magnitude smaller than $\frac{1}{6}$ and it may or may not be large enough to warrant caution. Examples include risky driving, medical decisions, diplomacy, the stock market, and air travel. For some of these choices (e.g., riding a motor cycle without wearing a helmet) there may be a one in a million chance of disaster while all other outcomes have negligible utilities:

$$\Delta u(o_d) = -10^9, p(o_d) = 10^{-6}, \quad \forall i \neq d : |\Delta u(o_i)| \leq 1. \tag{3.3}$$

If people decided based on n representative samples, they would completely ignore the potential disaster with probability $1 - (1 - 10^{-6})^n$. Thus to have at least a 50% chance of taking the potential disaster into account they would have to generate almost 700000 samples. This is clearly infeasible;

thus one would almost always take this risk even though the expected utility gain is about -1000. In conclusion, representative sampling is insufficient for resource-bounded decision-making when some of the outcomes are highly improbable but so extreme that they are nevertheless important. Therefore, the robustness of human decision-making suggests that our brains use a more sophisticated sampling algorithm—such as importance sampling.

Importance sampling is a popular sampling algorithm in computer science and statistics (Geweke, 1989; Hammersley & Handscomb, 1964) with connections to both neural networks (Shi & Griffiths, 2009) and psychological process models (Shi et al., 2010). It estimates a function's expected value with respect to a probability distribution p by sampling from an importance distribution q and correcting for the difference between p and q by down-weighting samples that are less likely under p than under q and up-weighting samples that are more likely under p than under q. Concretely, *self-normalized* importance sampling (Robert & Casella, 2009) draws s samples $x_1, \cdots, x_s$ from a distribution q, weights the function's value $f(x_j)$ at each point x_j by the *weight* $w_j = \frac{p(x_j)}{q(x_j)}$ and then normalizes its estimate by the sum of the weights:

$$X_1, \cdots, X_s \sim q, \quad w_j = \frac{p(x_j)}{q(x_j)} \tag{3.4}$$

$$\mathbb{E}_p[f(X)] \approx \hat{E}_{q,s}^{\text{IS}} = \frac{1}{\sum_{j=1}^s w_j} \cdot \sum_{j=1}^s w_j \cdot f(x_j). \tag{3.5}$$

With finitely many samples, this estimate is generally biased. Following Zabaras (2010), we approximate its bias and variance by

$$\text{Bias}[\hat{E}_{q,s}^{\text{IS}}] \approx \frac{1}{s} \cdot \int \frac{p(x)^2}{q(x)} \cdot (\mathbb{E}_p[f(x)] - f(x)) \, dx \tag{3.6}$$

$$\text{Var}[\hat{E}_{q,s}^{\text{IS}}] \approx \frac{1}{s} \cdot \int \frac{p(x)^2}{q(x)} \cdot (f(x) - \mathbb{E}_p[X])^2 \, dx. \tag{3.7}$$

We hypothesize that the brain uses a strategy similar to importance sampling to approximate the expected utility gain $\mathbb{E}_{p(O|A=a)}[\Delta u(O)]$ of taking action a and approximate the optimal decision $a^\star = \arg\max_a \mathbb{E}_{p(O|A=a)}[\Delta u(O)]$ by

$$\hat{a}^\star = \arg\max_a \overline{\Delta U}_{q,s}^{\text{IS}}(a), \quad \overline{\Delta U}_{q,s}^{\text{IS}}(a) \approx \mathbb{E}_{p(O|a)}[\Delta u(o)] \tag{3.8}$$

$$\overline{\Delta U}_{q,s}^{\text{IS}}(a) = \frac{1}{\sum_{j=1}^s w_j} \sum_{j=1}^s w_j \cdot \Delta u(o_j), \quad o_1, \cdots, o_s \sim q. \tag{3.9}$$

Note that importance sampling is a family of algorithms: each importance distribution q yields a different estimator, and two estimators may recommend opposite decisions. This leads us to investigate which distribution q yields the best decisions.

3.1.2 WHICH DISTRIBUTION SHOULD WE SAMPLE FROM?

Representative sampling is a special case of importance sampling in which the simulation distribution q is equal to the outcome probabilities p. Representative sampling fails when it neglects crucial eventualities. Neglecting some eventualities is necessary, but particular eventualities are more important than others. Intuitively, the importance of potential outcome o_i is determined by $|p(o_i) \cdot u(o_i)|$ because neglecting o_i amounts to dropping the addend $p(o_i) \cdot u(o_i)$ from the expected-utility integral (Equation 3.1). Thus, intuitively, the problem of representative sampling can be overcome by considering outcomes whose importance ($|p(o_i) \cdot u(o_i)|$) is high and ignoring those whose importance is low.

Formally, the agent's goal is to maximize the expected utility gain of a decision made from only s samples. The utility foregone by choosing a sub-optimal action can be upper-bounded by the error in a rational agent's utility estimate. Therefore the agent should minimize the expected squared error of its estimate of the expected utility gain $\mathbb{E}\left[\Delta U\right]$, which is the sum of its squared bias and variance , that is $\mathbb{E}\left[(\overline{\Delta U}_{q,s}^{\text{IS}} - \mathbb{E}\left[\Delta U\right])^2\right] = \text{Bias}\left[\overline{\Delta U}_{q,s}^{\text{IS}}\right]^2 + \text{Var}\left[\hat{\Delta U}_{q,s}^{\text{IS}}\right]$ (Hastie et al., 2009). As the number of samples s increases, the estimate's squared bias decays much faster ($O(s^{-2})$) than its variance ($O(s^{-1})$); see Equations 3.6-3.7. Therefore, as the number of samples s increases, minimizing the estimator's variance becomes a good approximation to minimizing its expected squared error.

According to variational calculus the importance distribution

$$q^{\text{var}}(o) \propto p(o) \cdot |\Delta u(o) - \mathbb{E}_p[\Delta U]| \tag{3.10}$$

minimizes the variance (Equation 3.7) of the utility estimate in Equation 3.9 (Geweke, 1998; Zabaras, 2010; see Appendix B). This means that the optimal way to simulate outcomes in the service of estimating an action's expected utility gain is to over-represent outcomes whose utility is much smaller or much larger than the action's expected utility gain. Each outcome's probability is weighted by how disappointing ($\mathbb{E}_p[\Delta U] - \Delta u(o)$) or elating ($\Delta u(o) - \mathbb{E}_p[\Delta U]$) it would be to a decision-maker anticipating to receive the gamble's expected utility gain ($\mathbb{E}_p[\Delta U]$). But unlike in *disappointment theory* (Bell, 1985; Loomes & Sugden, 1984, 1986), the disappointment or elation is not

added to the decision-maker's utility function but increases the event's subjective probability by prompting the decision-maker to simulate that event more frequently. Unlike in previous theories, this distortion was *not* introduced to describe human behavior but derived from first principles of resource-rational information processing: Importance sampling over-simulates extreme outcomes to minimize the mean-squared error of its estimate of the action's expected utility gain. It tolerates the resulting bias because it is more important to shrink the estimate's variance.

Unfortunately, importance sampling with q^{var} is intractable, because it presupposes the expected utility gain $\mathbb{E}_p[\Delta U]$ that importance sampling is supposed to approximate. However, the average utility $\overline{\Delta u}$ of the outcomes of previous decisions made in a similar context could be used as a proxy for the expected utility gain $\mathbb{E}_p[\Delta U]$. That quantity has been shown to be automatically estimated by model-free reinforcement learning in the midbrain (Schultz, Dayan, & Montague, 1997). Therefore, people should be able to sample from the approximate importance distribution

$$\tilde{q}(o) \propto p(o) \cdot \left| \Delta u(o) - \overline{\Delta u} \right| . \tag{3.11}$$

This distribution weights each outcome's probability by the extremity of its utility. Thus, on average, extreme events will be simulated more often than equiprobable outcomes of moderate utility. We therefore refer to simulating potential outcomes by sampling from this distribution as *utility-weighted sampling*.

3.1.3 Utility-weighted sampling

Having derived the optimal way to simulate a small number of outcomes (Equation 3.11), we now turn to the question how those simulated outcomes should be used to make decisions under uncertainty. The general idea is to estimate each action's expected utility gain from a small number of simulated outcomes, and then choose the action for which this estimate is highest.

If the simulated outcomes were drawn representatively from the outcome distribution p, then we could obtain an unbiased expected utility gain estimate by simply averaging their utilities (Equation 3.2). However, since the simulated outcomes were drawn from the importance distribution $\tilde{q}$ rather than p, we have to correct for the difference between these two distributions by computing a weighted average instead (Equation 3.5). Concretely, we have to weight each simulated outcome o_j by the ratio $w_j = \frac{p(o_j)}{\tilde{q}(o_j)}$ of its probability under the outcome distribution p over its probability under the importance distribution $\tilde{q}$ from which it was sampled. Thus, the extreme outcomes

that are overrepresented among the samples from $\tilde{q}$ will be down-weighted whereas the moderate outcomes that are underrepresented among the samples from $\tilde{q}$ will be up-weighted. Because $\tilde{q}(o) \propto p(o) \cdot \left| \Delta u(o) - \overline{\Delta u} \right|$, the weight w_j of outcome o_j is $\frac{1}{|\Delta u(o) - \overline{\Delta u}|/z}$ for some constant z. Since the weighted average in Equation 3.5 is divided by the sum of all weights, the normalization constant z cancels out. Hence, given samples $o_1, \cdots, o_s$ from the utility-weighted sampling distribution $\tilde{q}$, the expected utility gain of an action or prospect can be estimated by

$$\overline{\Delta U}_{\tilde{q},s}^{\text{IS}} = \frac{1}{\sum_{j=1}^{s} 1/\left| \Delta u(o_j) - \overline{\Delta u} \right|} \cdot \sum_{j=1}^{s} \frac{\Delta u(o_j)}{\left| \Delta u(o_j) - \overline{\Delta u} \right|}. \tag{3.12}$$

If no information is available a priori, then there is no reason to assume that the expected utility gain of a prospect whose outcomes may be positive or negative should be positive, or that it should be negative. Therefore, in these situations, the most principled guess an agent can make for the expected utility gain $\mathbb{E}_p[\Delta U]$ in Equation 3.10 – before computing it – is $\overline{\Delta u} = 0$. Thus, when the expected utility gain is not too far from zero, then the importance distribution q^{var} for estimating the expected utility gain of a single prospect can be efficiently approximated by

$$\tilde{q}(o) \propto p(o) \cdot |\Delta u(o)| . \tag{3.13}$$

This approximation simplifies the UWS estimator of a prospect's expected utility gain (Equation 3.12) into

$$\Delta \hat{U}_{\tilde{q},s}^{\text{IS}} = \frac{1}{\sum_{j=1}^{s} 1/|\Delta u(o_j)|} \cdot \sum_{j=1}^{s} \text{sign} \left(\Delta u(o_j) \right) , \quad o_j \sim \tilde{q}, \tag{3.14}$$

where $\text{sign}(x)$ is -1 for $x < 0$, 0 for $x = 0$, and $+1$ for $x > 0$.

This utility-weighted sampling mechanism succeeds where representative sampling failed. For Russian roulette, the probability that a sample drawn from the utility-weighted sampling distribution (Equation 3.13) considers the possibility of death (o_2) is

$$q(o_2) = \frac{p(o_2) \cdot |\Delta u(o_2)|}{p(o_2) \cdot |\Delta u(o_2)| + p(o_2) \cdot |\Delta u(o_2)|} = \frac{1/6 \cdot |-10^9|}{5/6 \cdot |1| + 1/6 \cdot |-10^9|} > 0.9999. \tag{3.15}$$

Consequently, utility-weighted sampling requires only 1 rather than 51 samples to recommend the correct decision at least 99.99% of the time, because the first sample is almost always the most important potential outcome (i.e., death). In this case, the utility estimate defined in Equation

3.14 would be $1/|10^9| \cdot -1 = -10^9$ and its expected value for a single sample is also very close to -10^9. While this mechanism is biased to overestimate the risk of playing Russian roulette ($\mathbb{E}[U] = -10^9/6 + 5/6 > -10^9$), that bias is beneficial because it makes it easier to arrive at the correct decision. Likewise, a single utility-weighted sample suffices to consider the potential disaster (Equation 3.3) at least 99.85% of the time, whereas even 700,000 representative samples would miss the disaster almost half of the time. Thus, utility-weighted sampling would allow people to make good decisions even under extreme time pressure. This suggests that to achieve the optimal bias-variance tradeoff (Hastie et al., 2009) the sampling distribution has to be biased towards extreme outcomes. This bias reduces the variance of the utility estimate enough to enable better decisions than representative sampling whose expected utility gain estimate is unbiased but has high variance.

To apply the utility-weighted sampling model to decisions people face in life and experiments, we have to specify the utility $u(o)$ of the outcomes o. To do so, we interpret an outcome's utility as the subjective value that the decision-maker's brain assigns to it in the choice context. Concretely, we follow the proposal of Summerfield and Tsetsos (2015) that the brain represents value in an efficient neural code. This proposal is based on psychophysical and neural data (Louie, Grattan, & Glimcher, 2011; Louie, Khaw, & Glimcher, 2013; Mullett & Tunney, 2013) and fits into our resource-rational framework: The brain's representational bandwidth is finite, because the possible range of neural firing rates is limited. Efficient coding makes rational use of the brain's finite representational bandwidth by adapting the neural code to the range of values that have to be represented in a given context. This implies rescaling the values of potential outcomes such that all of them lie within the representational bandwidth. If the representational bandwidth is 1 and the largest and the smallest possible values in the current context c are $o_c^{\max}$ and $o_c^{\min}$ respectively, then the utility of an outcome o should be represented by

$$u(o) = \frac{o}{o_c^{\max} - o_c^{\min}} + \varepsilon, \tag{3.16}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$ is neural noise that reflects uncertainty about the outcome's value. Since it is the neural representation of value rather than value itself that drives choice, we interpret $u(o)$ as the subjective utility of outcome o in context c. We will consistently use this formal definition of utility (Equation 3.16) in this and all following sections.

Our basic UWS model of how people estimate a prospect's expected utility thus has only two parameters: the number of samples s and the unreliability σ_ε of the decision-maker's representation of utility.

Having derived a resource-rational mechanism for estimating expected utilities, we now translate it into a decision strategy. Many real-world decisions and most laboratory tasks involve choosing between two actions a_1 and a_2 with uncertain outcomes $O^{(1)} \in \{o_1^{(1)}, o_2^{(1)}, \cdots, o_{n_1}^{(1)}\}$ and $O^{(2)} \in \{o_1^{(2)}, o_2^{(2)}, \cdots, o_{n_2}^{(2)}\}$ that depend on the unknown state of the world. Consider, for example, the choice between two lottery tickets: the first ticket offers a 1% chance to win $\$1000$ at the expense of a 99% risk to lose $\$1$ ($O^{(1)} \in \{-1, 1000\}$) and the second ticket offers a 10% chance to win $\$1000$ at the expense of a 90% risk to lose $\$100$ ($O^{(2)} \in \{-100, 1000\}$). According to expected utility theory, one should choose the first lottery (taking action a_1) if $\mathbb{E}[u(O^{(1)})] > \mathbb{E}[u(O^{(2)})]$ and the second lottery (action a_2) if $\mathbb{E}\left[u(O^{(1)})\right] < \mathbb{E}\left[u(O^{(2)})\right]$. This is equivalent to taking the first action if the expected utility difference $\mathbb{E}\left[u(O^{(1)} - u(O^{(2)}))\right]$ is positive and the second action if it is negative. The latter approach can be approximated very efficiently by focusing computation on those outcomes for which the utilities of the two actions are very different and ignoring events for which they are (almost) the same. For instance, it would be of no use to simulate the event that both lotteries yield $\$1000$ because it would not change the decision-maker's estimate of the differential utility and thus have no impact on her decision. To make rational use of their finite resources, people should thus use utility-weighted sampling to estimate the expected value of the two actions' differential utility $\Delta U = u(O^{(1)}) - u(O^{(2)})$ as efficiently as possible. This is accomplished by sampling pairs of outcomes from the bivariate importance distribution

$$q_\Delta^\star(O^{(1)}, O^{(2)}) \propto p(O^{(1)}, O^{(2)}) \cdot \left| u(O^{(1)}) - u(O^{(2)}) - \mathbb{E}\left[\Delta U\right]\right|, \qquad (3.17)$$

integrating their differential utilities according to

$$\Delta \hat{U}_{q,s}^{IS} = \frac{1}{\sum_{j=1}^{s} w_j} \sum_{j=1}^{s} w_j \cdot \left(u(o_j^{(1)}) - u(o_j^{(2)})\right), \quad o_1, \cdots, o_s \sim q_\Delta\left(O^{(1)}, O^{(2)}\right), \quad (3.18)$$

and then choosing the first action if the estimated differential utility is positive, that is

$$\hat{a}^\star = \begin{cases} 1 & \text{if } \Delta \hat{U}_{q,s}^{IS} > 0 \\ 2 & \text{if } \Delta \hat{U}_{q,s}^{IS} < 0 \\ 1 \text{ with } 50\% \text{ probability and } 2 \ 50\% \text{ probability} & \text{if } \Delta \hat{U}_{q,s}^{IS} = 0 \end{cases} \qquad (3.19)$$

Note that each simulation considers a pair of outcomes: one for the first alternative and one for the second alternative. This is especially plausible when the outcomes of both actions are determined by a common cause. For instance, the utilities of wearing a shirt versus a jacket on a hike are both primarily determined by the weather. Hence, reasoning about the weather naturally entails reasoning about the outcomes of both alternatives simultaneously and evaluating their differential utilities in each case (e.g. rain, sun, wind, etc.) instead of first estimating the utility of wearing a shirt and then starting all over again to estimate the utility of wearing a jacket.

Given that there is no a priori reason to expect the first option to be better or worse than the second option, $\mathbb{E}\left[\Delta U\right]$ is 0 and the equation simplifies to

$$q_\Delta(O^{(1)}, O^{(2)}) \propto p(O^{(1)}, O^{(2)}) \cdot \left| u(O^{(1)}) - u(O^{(2)}) \right|. \tag{3.20}$$

This distribution captures the fact that the decision-maker should never simulate the possibility that both lotteries yield the same amount of money– no matter how large it is. It does not overweight extreme utilities per se, but rather pairs of outcomes whose utilities are very different. Its rationale is to focus on the outcomes that are most informative about which action is best. For instance, in the example above, our UWS model of binary choice overweights the unlikely event in which the first ticket wins $1000 and the second ticket loses $100. Plugging the optimal importance distribution (Equation 3.20) into the UWS estimate for the expected differential utility yields an intuitive heuristic for choosing between two options. Formally, the optimal importance sampling estimator for the expected value of the differential utility ($\mathbb{E}\left[\Delta U\right]$) is

$$\Delta \hat{U}^{\text{IS}}_{\tilde{q},s} = \frac{1}{\sum_{j=1}^{s} 1 / \left| u\left(o_j^{(1)}\right) - u\left(o_j^{(2)}\right) \right|} \cdot \sum_{j=1}^{s} \text{sign}\left(u\left(o_j^{(1)}\right) - u\left(o_j^{(2)}\right) \right), \quad o_j \sim q_\Delta, \tag{3.21}$$

where $\text{sign}(x)$ is $+1$ for positive x and -1 for negative x. If the heightened availability of extreme events roughly corresponded to the utility-weighted sampling distribution (Equation 3.20), then the decision rule in Equation 3.21 could be realized by the following simple and psychologically plausible heuristic for choosing between two actions:

1. Imagine a few possible events
   (e.g., 1. Ticket 1 wins and ticket 2 loses. 2. Ticket 2 wins and ticket 1 loses. 3. Ticket 1 winning and ticket 2 losing comes to mind again. 4. Both tickets lose.).

2. For each imagined scenario, evaluate which action would fare better
   (1. ticket 1, 2. ticket 2, 3. ticket 1, 4. ticket 1).

3. Count how often the first action fared better than the second one (3 out of 4 times).

4. If the first action fared better more often than the second action, then choose the first action, else choose the second action (Get ticket 1!).

As a quantitative example, consider how UWS would choose between a ticket with a 10% chance of winning $99 and a 90% chance of losing $1 versus winning $1 for sure. If, the frequency with which events come to mind reflects utility-weighted sampling, then people could simply tally whether winning came to mind more often than losing. According to UWS, winning should came to mind about 86% of the time whereas losing should come to mind only about 14% of the time (the derivation of these simulation frequencies is provided in Appendix B). Hence, if the decision-maker imagined the outcome of choosing the gamble twice, there would be a 71.4% chance that winning came to mind twice, a 26.2% chance that winning and losing each came to mind once, and an only 2.4% chance of imagining losing twice. In the first case, the heuristic would always choose the gamble, in the second case it would choose it half of the time, and in the last case it would always decline the gamble. Hence, simply tallying which option (gambling vs. playing it safe) the imagined outcomes favored more frequently (and breaking ties at random) would be sufficient to make the correct decision 84% of the time despite having imagined the outcome only twice. Appendix B provides a complete description of this worked example and applies UWS to the general case of choosing between a gamble and its expected value.

The overweighting of outcomes that strongly favor one action over another in UWS is similar to the effect of anticipated regret in regret theory (Loomes & Sugden, 1982), but in UWS extremity changes the frequency with which an event is simulated and does *not* affect its utility. Magnifying the subjective probabilities of extreme events makes UWS more similar to salience theory (Bordalo, Gennaioli, & Shleifer, 2012) according to which pairs of payoffs that are very different receive more attention than pairs of payoffs that are similar. Yet, while salience theory provides a descriptive account of binary choice frequencies in decisions from description, UWS additionally provides a resource-rational mechanistic account of decisions from experience, memory recall, and frequency judgments.

3.1.5 SUMMARY AND PROSPECTIONS

In summary, our analysis suggested that the rational use of finite cognitive resources implies that extreme events should be overrepresented in decision-making under uncertainty. Utility-weighted

sampling is a rational process model that formalizes this prediction. This biased mechanism leads to better decisions than its unbiased alternative (i.e. representative sampling). Utility-weighted sampling thereby enables robust decisions under time constraints that prohibit the careful consideration of many possible outcomes.

We have derived two versions of utility-weighted sampling: The first version estimates the expected utility gain of a single action. The second version chooses between two actions. Although both mechanisms overweight extreme events their notions of extremity are different. The UWS mechanism for estimating the expected utility gain of a single action overweights individual outcomes with extreme utilities. By contrast, the UWS mechanism for choosing between two actions overweights pairs of outcomes whose utilities are very different. In the remainder of this chapter, the first mechanism is applied to simulate frequency judgment, pricing, and decisions from experience, and the second mechanism is applied to simulate binary decisions from description. Despite this difference, we can interpret the first mechanism as a special case of the second one, because its importance distribution (Equation 3.11) compares the utility of the prospect's outcomes against the average utility of alternative actions. Hence, UWS always overweights events that entail a large difference between the utility of the considered action and some alternative. The frequency with which a state has been experienced or its stated probability also influence how often it will be sampled. Thus, impossible and highly improbable states are generally unlikely to be sampled. However, states with high differential utility are sampled more frequently than is warranted by how often they have been experienced or their stated probability. This increases the probability that improbable states with extreme differential utility will be considered. We support the proposed mechanism by showing that it can capture people's memory biases for extreme events, the overestimation of the frequency of extreme events, biases in decisions from description, and biases in decisions from experience.

3.2 Biases in frequency judgment confirm predictions of UWS

If people remembered the past as if they were sampling from the UWS distribution (Equation 3.11), they would recall their best experience and their worst experience much more frequently than an unremarkable one (Madan et al., 2014, cf.). If people relied on such a biased memory system to estimate frequencies and assess probabilities, then their estimate $\hat{f}_k$ of the frequency $f_k = p(o_k)$ of the event o_k would be

$$\hat{f}_k = \frac{\sum_{i=1}^{s} w_i \cdot \delta(o_i = o_k)}{\sum_{i=1}^{s} w_i}, \quad w_i = \frac{1}{|u(o_i)|}, \quad o_i \sim \tilde{q}, \tag{3.22}$$

where $\tilde{q}(o) \propto p(o) \cdot |u(o)|$ is the utility-weighted sampling distribution. Since $\tilde{q}$ over-represents each event o_k proportionally to its extremity $|u(o_k) - \bar{u}|$, that is $\frac{\tilde{q}(o_k)}{p(o_k)} \propto |u(o_k) - \bar{u}|$, we predict that people's relative over-estimation $\frac{\hat{f}_k}{f_k}$ is a monotonically increasing function of the event's extremity $|u(o_k) - \bar{u}|$. Formally, the bias (Equation 3.6) of utility-weighted probability estimation (Equation 3.22) implies that the relative amount by which people overestimate an event's frequency (i.e., $\frac{\hat{f}_k - f_k}{f_k}$) should increase with the event's extremity ($|u(o_k)| - \bar{u}$), according to

$$\frac{\hat{f}_k - f_k}{f_k} = \frac{1}{s}\left(c - \frac{1}{|u(o_k) - \bar{u}|}\right), \tag{3.23}$$

where c is an upper bound on people's relative overestimation. This predicts that people should overestimate the frequency of an event more the more extreme it is regardless of its frequency. In this section, we test this prediction against people's judgments: we first report an experiment suggesting that frequency overestimation increases with perceived extremity, and then we show that UWS can capture the finding that overestimation occurs regardless of the event's frequency (Madan et al., 2014).

3.2.1 Frequency overestimation increases with perceived extremity

Lichtenstein et al. (1978) and Pachur, Hertwig, and Steinmann (2012) found that people's estimates of the frequencies of lethal events are strongly correlated with how many instances of each event they can recall. Furthermore, Lichtenstein et al. (1978) also found that overestimation was positively correlated with the number of lives lost in a single instance of each event, the likelihood that an occurrence of the event would be lethal, and the amount of media coverage it would typically attract. We hypothesize that extremity-weighted memory encoding contributed to these effects. If this were true, then overestimation should increase with perceived extremity. Here, we test this prediction of UWS in a new experiment that measures perceived extremity and correlates it with the biases in people's frequency estimates.

METHODS   We recruited 100 participants on Amazon Mechanical Turk. Participants received a baseline payment of \$1.25 for about 30 minutes of work. Participants were asked to estimate how many American adults had experienced each of 39 events in 2015 as accurately as possible and accurate frequency estimation was incentivized by a performance dependent bonus of up to \$2. In addition, participants judged each event's valence (good or bad) and extremity (0: neutral – 100: ex-

treme). The 39 events comprised 30 stressful life events from Hobson et al. (1998), four lethal events (suicide, homicide, lethal accidents, and dying from disease/old age), three rather mundane events (going to the movies, headache, and food-poisoning), and two attention-checks. As a reference, participants were told the total number of American adults and how many of them retire each year.

To assess overestimation we compared our participants' estimates to the true frequencies of the events according to official statistics.[‡] The complete experiment can be inspected online.[§] Out of 100 participants 22 failed one or more attention checks (number of Americans elected president, number of Americans who slept between 2h and 10h at least once) and were therefore excluded.

RESULTS AND DISCUSSION   A significant rank correlation[¶] between the average extremity judgments of the 37 events and average relative overestimation $\frac{\hat{f}_k - f_k}{f_k}$ confirmed our model's prediction (Spearman's $\rho = 0.46, p = .0045$, see Figure 3.1), and we observed the same effect at the level of individual judgments (Spearman's $\rho = 0.14, p < 10^{-12}$). The frequencies of the five most extreme events, that is murder (93.3%), suicide (92.6% extreme), dying in an accident (90% extreme), the death of one's partner (86% extreme), and suffering a major injury or serious illness (85% extreme) were *over*estimated by a factor of 159 ($p = 0.0001$), 9 ($p = 0.0026$), 35 ($p = 0.0035$), 1.01 ($p = 0.03$), and -0.22 ($p = 0.25$) respectively. By contrast, the prevalences of the five least extreme events, that is headache (20% extreme), change in work responsibilities (21% extreme), getting a traffic ticket (26% extreme), moving flat (26% extreme), and career change (32% extreme) were *under*estimated by 4% ($p = 0.42$), 1% ($p = 0.95$), 10% ($p = 0.52$), 52% ($p < 0.0001$), and 24% ($p = 0.0211$) respectively. Like Rothman et al. (1996), we found that people overestimate the frequency of suicide (overestimated by 927%) more heavily than the frequency of divorce (overestimated by 27%). According to our theory, this is because suicide is perceived as more extreme than divorce (92.6% extreme vs. 59% extreme).

Furthermore, we found that the effect of extremity on overestimation also holds across the three categories the events were drawn from (see Figure 3.2): people significantly underestimated the frequency of mundane events ($t(233)$=-3.66,p=0.0003) while overestimating the frequency of stressful life events ($t(2338) = 2.02, p = 0.0433$) and lethal events ($t(311) = 5.46, p < 10^{-7}$). Two-

---

[‡]This data was obtained from Hobson & Delunas (2001),www.cdc.gov/nchs/fastats/deaths.htm, www.mpaa.org/resources/3037b7a4-58a2-4109-8012-58fca3abdf1b.pdf,www.cdc.gov/foodborneburden/, and Rasmussen, Jensen, Schroll, & Olesen (1991).

[§]http://cocosci.berkeley.edu/mturk/falk/freq_estimation_revised.html

[¶]We analyzed this relationship using Spearman's rank correlation, since we cannot assume that people's extremity judgments follow a ratio scale.
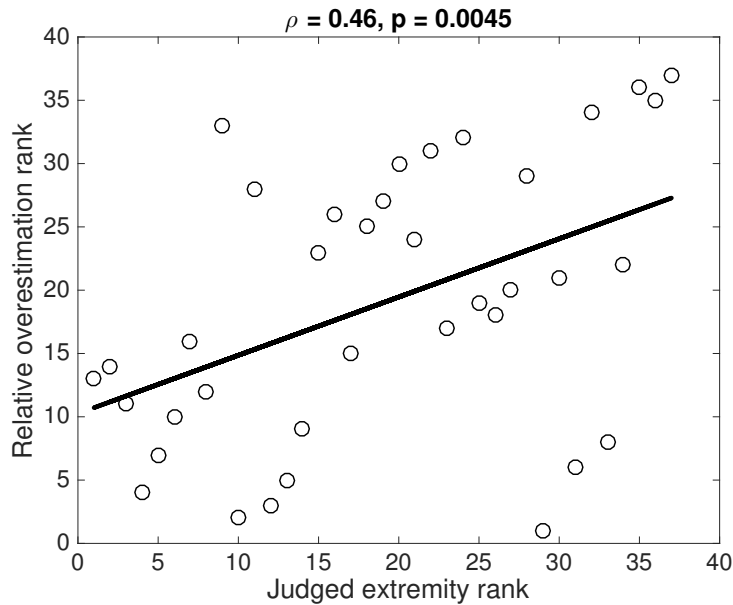
Figure 3.1: Relative overestimation $\left((\hat{f}_k - f_k)/f_k\right)$ increases with perceived extremity $(|u(o_k)|)$. Each circle represents one event's average ratings.

sample t-tests confirmed that relative overestimation was larger for stressful life events than for mundane events $(t(2571) = 3.16, p = 0.0016)$ and even larger for lethal events $t(544) = 12.70, p < 10^{-15})$. Figure 3.2 illustrates that overestimation and perceived extremity increased together.

While people's judgments were biased for the events studied here, there are many quantities, such as the length of poems, for which people's predictions are unbiased (Griffiths & Tenenbaum, 2006). This is consistent with UWS because unlike monetary gains and losses they impart no (dis)utility on their observer. For instance, hearing that a poem is 8 lines long carries virtually the same utility as hearing that another poem is 25 lines long. Hence, for such quantities, UWL would predict effectively unbiased memory encoding, recall, and prediction. Our theory's ability to differentiate situations where human judgment is biased from situations where it is unbiased speaks to its validity.

In conclusion, the experiment confirmed our theory's prediction that an event's extremity increase the relative overestimation of its frequency. However, additional experiments are required to disentangle the effects of extremity and low probability, because these two variables were anticorrelated $(\rho(36) = -0.67, p < 0.0001)$. To address this problem, we examined our model's
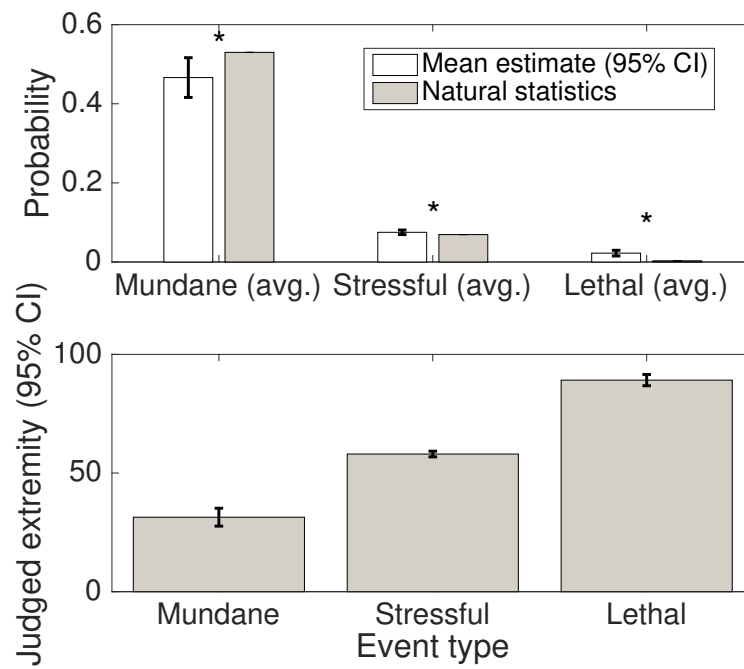
Figure 3.2: Judged frequency and extremity by event type.

predictions using two published studies that kept frequency constant across events (Madan et al., 2014).

3.2.2 UWS captures overestimation of extreme events regardless of frequency

The results reported above supported the hypothesis that people overestimate the frequency of extreme events, but most extreme events in that experiment were also rare. Therefore our findings could also be explained by postulating that people overestimate extreme events only because they are rare (Hertwig, Pachur, & Kurzenhäuser, 2005). This possibility is supported by empirical evidence for regression to the mean effects in frequency estimation (Attneave, 1953; Hertwig et al., 2005; Lichtenstein et al., 1978; H. Zhang & Maloney, 2012). Yet, extremity per se also contributes to overestimation: Madan et al. (2014) found that people overestimate the frequency of an extreme event relative to a non-extreme event even when both were equally frequent. The hypothesis that people overestimate the frequency of extreme events because those events are rare cannot account for this finding, but utility-weighted sampling can. To demonstrate this, we simulated the experiments by Madan et al. (2014) using utility-weighted sampling.

In the first experiment by Madan et al. (2014) participants repeatedly chose between two doors. Each door probabilistically generated one of two outcomes, and different doors were available on different trials. There were a total of four doors generating a sure gain of $+20$ points, a sure loss of -20 points, a risky gain offering a 50/50 chance of $+40$ or 0, and a risky loss offering a 50/50 chance of 0 or -40 points. In most trials participants either chose between the risky and the sure gain (gain trials) or between the risky and the sure loss (loss trials). After each choice, participants were shown the number of points earned, and they received no additional information about the options. After 6 blocks of 48 such choices participants were asked to estimate the probability with which each door generated each of the possible outcomes and to report the first outcome that came to their mind for each of the four doors. In their second experiment Madan et al. (2014) shifted all outcomes from Experiment 1 by $+40$ points.

We estimated the two parameters of the UWS model (i.e., the number of samples s and the noisiness σ_ε of the utility function) from the choice frequencies reported by Madan et al. (2014) using the maximum-likelihood principle. While participants had to learn the outcome probabilities from experience, the model developed so far assumes known probabilities. We thus restricted our analysis to the last block of each experiment. For each experiment, our model defines a likelihood function over the number of risky choices in gain trials and the number of risky choices in loss trials. We max-

Table 3.1: UWS simulation of people's memory recall (A) and frequency estimates (B) after the Experiments by Madan et al. (2014).

A

Comes to mind first:	Extreme Gain vs. Neutral	Extreme Loss vs. Neutral
Experiment 1	64.5% vs. 35.5%	71% vs. 29%
Experiment 2	70.0% vs. 30%	72.6% vs. 27.4%

B

Estimated Frequency of ...	Extreme Gain vs. Neutral	Extreme Loss vs. Neutral
Experiment 1	83.0% vs. 17.0%	87.5% vs. 12.5%
Experiment 2	87.5% vs. 12.5%	90.0% vs. 10.0%

Table 3.2: UWS captures people's risk preferences in the Experiments by Madan et al. (2014).

Risky Choices in	Gain Trials	Loss Trials
Experiment 1:	UWS: 54% People: 45%	UWS: 36%, People: 35%
Experiment 2:	UWS: 60% , People: 55%	UWS: 31%, People: 14%

imized the product of these likelihood functions with respect to our model's parameters using grid search over possible numbers of samples and global optimization with respect to σ_ε. The resulting parameter estimates were $s = 4$ samples and $\sigma_\varepsilon = 0.05$.

With these parameters, utility-weighted sampling correctly predicted that extreme outcomes come to mind first more often than the equally frequent moderate outcomes; see Table 3.1A. Next, we simulated people's frequency estimates according to Equation 3.22. UWS correctly predicted that people overestimate the frequency of extreme outcomes relative to the equally frequent moderate outcome; see Table 3.1B. In addition, UWS captured that participants were more risk-seeking for gains than for losses (see Table 3.2), and a later section investigates this phenomenon in more detail.

3.2.3 Summary and Discussion

The findings presented in this section provide strong support for our hypothesis that utility-weighting is the reason why people over-represent extreme events: First, Experiment 1 showed that there is a significant correlation between an event's utility and the degree to which people overestimate its frequency. Second, the data from Madan et al. (2014) rule out the major alternative explanation that

people overestimate the frequency of extreme events only because they are rare and also demonstrate that the overestimation is mediated by a memory bias for events with extreme utility. Furthermore, we found that the adaptive bias predicted by our theory exists not only in decision-making but also in frequency estimation and memory.

A parsimonious explanation for these three phenomena could be that the over-representation of extreme events results from a known bias in learning: emotional salience enhances memory formation (Cruciani et al., 2011). While overestimation has been previously explained by high "availability" of salient memories (Tversky & Kahneman, 1973), our theory specifies what exactly the availability of events should correspond to – namely their importance distribution $\tilde{q}$ (Equation 3.13) – and why this is useful. Our empirical findings were consistent with utility-weighted sampling but inconsistent with the hypothesis that the bias in frequency estimation is merely a reflection of the regression to the mean effect (Hertwig et al., 2005). While alternative accounts of why people overestimate the frequency of extreme events, such as selective media coverage (Lichtenstein & Slovic, 1971), can explain the overestimation of certain lethal events, they cannot account for the data of Madan et al. (2014). Thus at least part of the overestimation of extreme events appears to be due to utility-weighted sampling. Hence, an event's extremity may sway people's decisions by increasing their propensity to remember it, and this is clearly distinct from extremity's potential effects on the subjective utility of anticipated outcomes (Bell, 1985; Loomes & Sugden, 1982, 1984, 1986).

Our model's predictions are qualitatively consistent with the data of Madan et al. (2014)(Madan et al., 2014) but often more extreme. This difference might result from the idealistic assumption that there is no forgetting. This issue will be revisited with a more realistic learning model later in the chapter.

3.3   BIASES IN DECISIONS FROM DESCRIPTION

According to decision theory, an event's probability determines its weight in decision-making under uncertainty. Therefore, the biased probability estimates induced by utility-weighted sampling suggest that people should overweight extreme events in decisions under uncertainty. We will test this prediction in the domain of decisions from experience. Since this will require a model of learning, we model decisions from description as an intermediate step towards building a model of decisions from experience.

In the decisions from description paradigm participants choose between gambles that are de-

scribed by their payoffs and outcome probabilities (Allais, 1953; Kahneman & Tversky, 1979). Typically participants make binary choices between pairs of gambles or between a monetary gamble and a sure payoff. While people could, in principle, make these decisions by computing and comparing the gamble's expected values, ample empirical evidence demonstrates that they do not. Instead, people might reuse their strategies for everyday decisions. Everyday decisions are usually based on memories of past outcomes in similar situations. Hence, if people reused their natural decision strategies, then their decisions from description should be affected by the availability biases that have been observed in memory recall and frequency judgments. Our section on utility-weighted learning in decision from experience provides a precise, mechanistic account of how these biases arise from biased memory encoding. Here, we assume that similar mechanisms are at play in decisions from description. For instance, it is conceivable that the high salience of large differential payoffs in decisions from description (Bordalo et al., 2012) attracts a disproportionate amount of people's attention, making them more memorable, and increasing the frequency with which they will be considered. We think that such mechanisms could roughly approximate the utility-weighting prescribed by our model, at least for simple gambles whose outcomes are displayed appropriately.

In this section, we therefore apply UWS to decisions from description, validate the resulting model on the data from the Technion choice prediction competition (Erev et al., 2010), and demonstrate that it can capture three classic violations of expected utility theory.

3.3.1   Validation on decisions from description

We validated the utility-weighted sampling model of binary choices (Equations 3.18–3.21) with the stochastic normalized utility function defined in Equation 3.16 against people's decisions from description in the Technion choice prediction tournament (Erev et al., 2010). There are many factors that influence people's responses that are outside the scope of our model. These include accidental button presses, mind-wandering, misperception, and the occasional use of additional decision strategies that might be well adapted to the specific problems to which they are applied (Lieder & Griffiths, 2015, 2017). We therefore extended UWS to allow for an unknown proportion of choices (p_{random}) that are determined other factors. We model the net effect of those choices as choosing either option with a probability of 0.50.

We fitted the number of samples s, the noisiness σ_ε of the utility function, and the percentage of trials in which people choose at random to the training data of the Technion choice prediction competition. The maximum likelihood estimates of these model parameters were $s = 10$ samples,

$\sigma_\varepsilon = 0.1703$, and $p_{\mathrm{random}} = 0.07$. We then used these parameter estimates to predict people's choices in the decision problems of the test set of the Technion choice prediction competition. Figure 3.3 shows our model's predictions and compares them to people's choice frequencies. On average across the 60 problems, people chose the risky option about $46.75 \pm 3.98\%$ of the time and the UWS model chose the risky option about $48.92 \pm 2.56\%$ of the time. This difference was not statistically significant $(t(59) = -1.03, p = 0.31)$ suggesting that the predictions of UWS were unbiased. While there was no bias—on average—the predictions of UWS were regressed towards 50/50 compared to people's choice frequencies: On problems where people were risk-seeking UWS chose the risky option less often than people $(66.11\%$ vs. $79.20\%, t(24) = -6.48, p < .0001)$. But on problems where people were risk-averse, UWS chose the risky option more often than people $(35.54\%$ vs. $21.97\%, p < .0001)$.

Our model predicted people's choice frequencies more accurately than cumulative prospect theory (CPT; Tversky, & Kahneman, 1992) or the priority heuristic (Brandstätter, Gigerenzer, & Hertwig, 2006): Its mean squared error $(\mathrm{MSD}_{\mathrm{UWS}} = 0.0266)$ was significantly lower than for cumulative prospect theory $(\mathrm{MSD}_{\mathrm{CPT}} = 0.0837, t(59) = -5.4, p < .001)$ or the priority heuristic $(\mathrm{MSD}_{\mathrm{priority}} = 0.1437, t(59) = -4.9, p < .001)$. Furthermore, the predicted risk preference agreed with people's risk preferences in 87% of the trials (CPT: 93%, priority heuristic: 81%) and the predicted choice frequencies were highly correlated with people's choice frequencies $(r_{\mathrm{UWS}}(59) = 0.88, p < 10^{-15}$ versus $r_{\mathrm{CPT}} = 0.86$ and $r_{\mathrm{priority}} = 0.65)$. Our model's predictive accuracy was similar to those of the best existing models, namely stochastic cumulative prospect theory with normalization $(r = 0.92, \mathrm{MSD} = 0.0116)$ and Haruvy's seven parameter logistic regression model that won the competition $(r = 0.92, \mathrm{MSD} = 0.0126)$, although the differences were still statistically significant $(t(59) = 3.5, p < .001$ and $t(59) = 3.97, p < .001)$. In addition to performing about as well as the best existing models our model is distinctly principled: UWS is the only accurate mathematical process model that is derived from first principles. All alternative models that perform similarly well were tailored to capture known empirical phenomena or fail to specify the mechanisms of decision-making.

Having estimated our model's parameters and validated it, we now proceed to demonstrate that it can explain three paradoxes in risky choice, namely the Allais paradox (Allais, 1953), the fourfold pattern of risk preferences (Tversky & Kahneman, 1992), and preference reversals (Lichtenstein & Slovic, 1971).

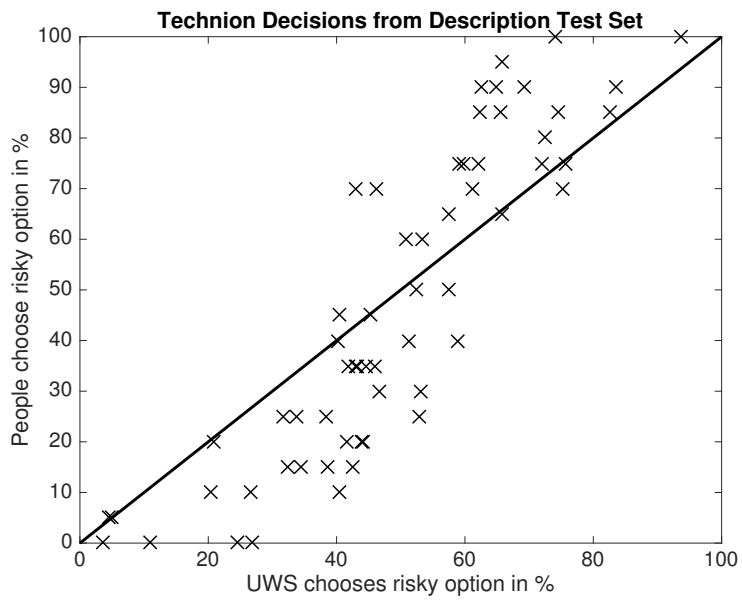**Figure 3.3:** Predictions of UWS on the test set of the Technion choice prediction tournament for decisions from description according to the parameters estimated from the training set. Each data point reports the frequency with which UWS (horizontal axis) versus people (vertical axis) chose the risky option in one of the 60 decision problems of the Technion competition, and the solid line is the identity line.

Table 3.3: The Allais gambles: Participants choose between lottery L_1 and lottery L_2 for $z = 2400$ versus $z = 0$.

	(o_1, p_1)	(o_2, p_2)	(o_3, p_3)
$L_1(z):$	$(z, 0.66)$	$(2500, 0.33)$	$(0, 0.01)$
$L_2(z):$	$(z, 0.66)$	$(2400, 0.34)$	

3.3.2   THE ALLAIS PARADOX

In the two lotteries $L_1(z)$ and $L_2(z)$ defined in Table 3.3 the chance of winning z dollars is exactly the same. Yet, when $z = 2400$ most people prefer lottery L_2 over lottery L_1, but when $z = 0$ the same people prefer L_1 over L_2. This inconsistency is known as the Allais paradox (Allais, 1953).

We simulated people's choices between both pairs of lotteries according to utility-weighted sampling with the parameters estimated from the Technion training set. To do so, we computed the probability p and utility difference ΔU for each possible pair of outcomes of the first lottery L_1 and the second lottery L_2. Since the outcomes of the two lotteries are statistically independent, the probability that the first lottery yields outcome O_1 while the second lottery yields O_2 is $P(O_1) \cdot P(O_2)$. To apply UWS to predict people's choices between the two lotteries, we determined all possible values of the differential utility ΔU and their respective probabilities. For instance, when $z = 0$, then the possible differential utilities are $0, -u(2400), u(2500) - u(2400)$, and $u(2500)$ (see Tables 3.3 and 3.4). In this case, ΔU is $-u(2400)$ if the first or the third outcome is drawn for the first lottery and the second outcome is drawn for the second lottery. The probability of the first scenario is $p_1 \cdot p_2 = 0.66 \cdot 0.34$ and the probability of the second scenario is $p_3 \cdot p_2 = 0.01 \cdot 0.34$; hence the probability of $\Delta U = -u(2400)$ is $0.67 \cdot 0.34$. Next, we computed the simulation frequency $\tilde{q}(\Delta U)$ which is proportional to $p(\Delta U) \cdot |\Delta u|$. For instance, in this example, $\mathbb{E}\left[\tilde{q}(\Delta U = -u(2400))\right] \propto 0.67 \cdot 0.34 \cdot u(2400)$ and normalizing this probability distribution yields $\mathbb{E}\left[\tilde{q}(\Delta U = -u(2400))\right] = 0.5$ suggesting that this extreme eventuality would occupy half of the decision-maker's mental simulations even though its probability is less than 23%. This corresponds to overweighting this event by a factor of 2.19. Table 3.4 presents these numbers for all differential utilities possible with $z = 2400$ or $z = 0$.

Our simulations with UWS predicted people's seemingly inconsistent preferences in the Allais paradox. For the first pair of lotteries ($z = 2400$), UWS preferred the second lottery to the first one, choosing L_2 55.66% of the time and L_1 only 44.34% of the time. But for the second pair of lotteries ($z = 0$), UWS choose the first lottery more often than the second one (50.38% vs.

Table 3.4: Utility-weighted sampling explains the Allais paradox.

ΔU	p	$\mathbb{E}\left[\tilde{q}\right]$	$\mathbb{E}\left[\tilde{q}\right]/p$
0	0.66	0	0
$u(2500) - u(2400)$	0.33	0.58	1.8
$-u(2400)$	0.01	0.42	42

$z = 2400$:

ΔU	p	$\mathbb{E}\left[\tilde{q}\right]$	$\mathbb{E}\left[\tilde{q}\right]/p$
0	$0.66 \cdot 0.67$	0	0
$-u(2400)$	$0.67 \cdot 0.34$	0.5	2.19
$u(2500) - u(2400)$	$0.33 \cdot 0.34$	0.01	0.08
$u(2500)$	$0.33 \cdot 0.66$	0.49	2.26

$z = 0$:

Note: The agent's simulation yields $\Delta U = \Delta u$ with probability $\tilde{q}(\Delta u) \propto p(\Delta u) \cdot |\Delta u|$ where p is Δu's objective probability.

49.62%). Table 3.4 shows how our theory explains why people's preferences reverse when z changes from 2400 to 0: According to the importance distribution $\tilde{q}$ (Equation 3.13), people overweight the event for which the utility difference between the two gambles' outcomes (O_1 and O_2) is largest ($\Delta U = u(O_1) - u(O_2)$). Thus when $z = 2400$, the most over-weighted event is the possibility that gamble L_1 yields $o_1 = 0$ and gamble L_2 yields $o_2 = 2400$ ($\Delta U = -u(2400)$); consequently the bias is negative and the first gamble appears inferior to the second ($\mathbb{E}[\Delta \hat{U}_{\tilde{q},2}^{\text{IS}}] = -0.0294$ which corresponds to $\$ -75.54$). But when $z = 0$, then L_1 yielding $o_1 = 2500$ and L_2 yielding $o_2 = 0$ ($\Delta U = +u(2500)$) becomes the most over-weighted event making the first gamble appear superior ($\mathbb{E}[\Delta \hat{U}_{\tilde{q},2}^{\text{IS}}] = +0.0013$ which corresponds to $\$3.25$). Our model's predictions are qualitatively consistent with the empirical findings by Kahneman and Tversky (1979) but less extreme; this is primarily because fitting the model to the data from the Technion choice prediction Tournament led to large number of samples ($s = 10$) and the predicted availability biases decrease with the number of samples; for a smaller number of samples, the model predictions would have been closer to the empirical data.

3.3.3 THE FOURFOLD PATTERN OF RISK PREFERENCES

Framing outcomes as losses rather than gains can reverse people's risk preferences (Tversky & Kahneman, 1992): In the domain of gains people prefer a lottery (o dollars with probability p) to its expected value (*risk seeking*) when $p < .5$, but when $p > .5$ they prefer the expected value (*risk*

aversion). In contrast, in the domain of losses, people are risk averse for $p < .5$ but risk seeking for $p > .5$. This phenomenon is known as the *fourfold pattern of risk preferences*. Formally, decision-makers are risk seeking when they prefer a gamble $(p, o; 0)$ which yields \$$o$ with probability p and nothing otherwise to its expected value $p \cdot o$ dollars, and risk averse if they prefer receiving the expected value for sure to playing the gamble. We therefore determined the risk preferences predicted by utility-weighted sampling by simulating choices between such gambles and their expected values. Concretely, we used the gambles $(p, o; 0)$ for $0 < p < 1$ and $-1000 < o < 1000$ and applied UWS with the parameters estimated from the Technion choice prediction tournament. Appendix B illustrates how utility-weighted sampling makes these decisions and how this leads to inconsistent risk preferences.

We found that utility-weighted sampling predicts the fourfold pattern of risk preferences (Tversky & Kahneman, 1992); see Figure 3.4. To understand how utility-weighted sampling explains this phenomenon, remember that it estimates the expected value of the differential utility ΔU by sampling from the importance distribution $\tilde{q}(\Delta u) \propto |\Delta u| \cdot p(\Delta u)$. The differential utility of choosing a gamble that yields o with probability p over its expected value $p \cdot o$ is

$$\Delta U = \begin{cases} u(o) - u(p \cdot o) & \text{with probability } p \\ -u(p \cdot o) & \text{with probability } 1 - p \end{cases}. \tag{3.24}$$

Utility-weighted sampling thus overweights the gain/loss o of the lottery if p is small, because then $|u(o) - u(p \cdot o)| > |u(p \cdot o)|$. Conversely, it underweights the gain/loss o if p is large, because then $|u(o) - u(p \cdot o)| < |u(p \cdot o)|$. Concretely, when choosing between a two-outcome gamble and its expected value, UWS simulates the outcome of the gamble as if winning and losing were equally probable even when the probability of winning is much larger or much smaller than 0.5 (see Appendix B). On top of this over-simulation of the more extreme outcome, the noise term of the utility function (Equation 3.16) stochastically flips the sign of the differential utilities of some of the simulated outcomes. When the probability of winning is close to 0 or 1, then this happens almost exclusively for the outcome whose differential utility is closer to zero. Combined with the over-simulation of the more extreme outcome this asymmetry renders the decision-maker's bias positive (risk-seeking) for improbable gains and probable losses but negative (risk-aversion) for probable gains and improbable losses (see Figure 3.4). Appendix B elaborates this explanation with detailed worked examples.

In everyday life the fourfold pattern of risk preferences manifests itself in the apparent para-

128

dox that people who are so risk-averse that they buy insurance can also be so risk-seeking that they play the lottery. Our simulations resolved this apparent contradiction: First, we simulated the decision whether or not to play the Powerball lottery.[‖] The jackpot is at least $40 million, but the odds of winning it are less than 1:175 million. In brief, people pay $2 to play a gamble whose expected value is only $1. We simulated how much people would be willing to pay for a ticket of the Powerball lottery according to UWS. We found that UWS overestimates the value of a lottery ticket by more than a factor of 2 more than 36% of the time. Thus, a person who evaluates lottery tickets often should consider them underpriced about one third of the time. Applied to choice, UWS predicts that people buy lottery tickets almost every second time they consider it (P^{UWS}(buy lottery ticket) $= 0.497$), because they over-represent the possibility of winning big. Next, we applied UWS to predict how much the same people would be willing to pay for insurance. Our simulation assumed that the total insured loss follows the heavy-tailed power-law distribution of debits (N. Stewart et al., 2006) over the range from $1 to $1 000 000. To simplify the application of UWS to this continuous distribution, we set the reward expectancy $\bar{u}$ to zero and assumed that the simulation distribution is not affected by noise. We determined the certainty equivalents of the utility-weighted sampling estimates of the utility of an insurance against a loss drawn from this distribution. To do so, we applied the inverse of the utility function to the UWS estimates of the expected disutility of the hazard. We found that UWS overestimates the expected hazard about 80% of time, and it overestimates it by a factor of at least 2 in 64% of all cases. Therefore, most people should be motivated to buy insurance even when they just bought a lottery ticket. The prediction of utility-weighted sampling for whether people actually decide to buy an overpriced insurance policy are more moderate, because the high price of insurance makes the possibility of paying nothing and losing nothing more salient. Nevertheless, UWS predicts that people would be willing to buy insurance for 130% of its expected value about 37.3% of the time. Thus 90% of customers would buy 130% overpriced insurance after considering at most 5 offers.

Utility-weighted sampling thereby resolves the paradox that people who are so risk-seeking that they buy lottery tickets can also be so risk-averse as to buy insurance by suggesting that people overweight extreme events regardless of whether they are gains (as in the case of lotteries) or losses (as in the case of insurance).

[‖]The payoffs and probabilities of this lottery were modeled according to http://www.calottery.com/play/draw-games/powerball.
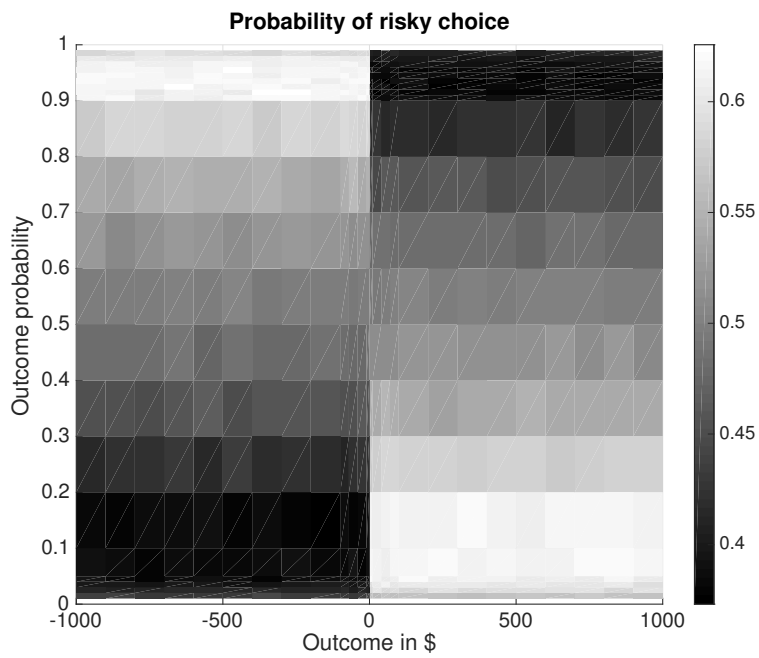
Figure 3.4: Utility-weighted sampling predicts the fourfold pattern of risk preferences. The color scale indicates the probability people make the risky choice as a function of the probability and dollar value of the outcome.

3.3.4 Preference reversals

When people first price a risky gamble and a safe gamble with similar expected value and then choose between them, their preferences are inconsistent almost 50% of the time: most people price the risky gamble higher than the safe one, but many of them nevertheless choose the safer one (Lichtenstein & Slovic, 1971). This inconsistency does not result from mere randomness, as preference reversals in the opposite direction are rare.

To evaluate whether our theory can capture this inconsistency, we simulated the pricing of a safe gamble offering an 80% chance of winning \$1 and a risky gamble offering a 40% chance of winning \$2, and the subsequent choice between them according to UWS with the parameters estimated from the Technion choice prediction tournament for decisions from description. Since the largest and the smallest possible outcome are $o_{\max} = 2$ and $o_{\min} = 0$ respectively, the utility function from Equation 3.16 becomes $u(o) = \frac{o}{2} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma = 0.17)$.

We assumed that people price a gamble by estimating its expected utility gain $\overline{\Delta U}_{q,s}^{IS}$ according to Equation 3.14 and then convert the resulting utility estimate into its monetary equivalent. Plugging the payoffs and outcome probabilities of the safe gamble in to Equation 3.14 reveals that, for the safe gamble, winning ($o = 1$) and losing ($o = 0$) would be simulated with the frequencies

$$q_{\text{safe}}(o = 1) = \frac{0.8 \cdot |u(\$1)|}{0.8 \cdot |u(\$1)| + 0.2 \cdot |u(\$0)|}, \text{ and} \tag{3.25}$$

$$q_{\text{safe}}(o = 0) = \frac{0.2 \cdot |u(\$0)|}{0.8 \cdot |u(\$1)| + 0.2 \cdot |u(\$0)|}, \tag{3.26}$$

respectively. For the risky gamble the possibility of winning is over-represented more:

$$q_{\text{risky}}(o = 2) = \frac{0.4 \cdot |u(\$2)|}{0.4 \cdot |u(\$2)| + 0.6 \cdot |u(\$0)|}, \text{ and} \tag{3.27}$$

$$q_{\text{risky}}(o = 0) = \frac{0.6 \cdot |u(\$0)|}{0.4 \cdot |u(\$2)| + 0.6 \cdot |u(\$0)|}. \tag{3.28}$$

Each simulated decision-maker sampled 10 possible outcomes. We then applied Equation 3.14 to translate the 10 samples from q_{safe} into the UWS estimate of the expected utility gain of playing the safe gamble ($\overline{\Delta u}_{q_{\text{safe}},10}^{IS}$) and the 10 samples from q_{risky} into the UWS estimate of the expected utility gain of playing the risky gamble ($\overline{\Delta u}_{q_{\text{risky}},10}^{IS}$). Finally, we converted each estimated utility gain into the equivalent monetary amount m by inverting the utility function u without adding any noise,

that is

$$m_{\text{risky}} = u^{(-1)}(\overline{\Delta u}^{\text{IS}}_{q_{\text{risky}},10}) = (o_{\max} - o_{\min}) \cdot \overline{\Delta u}^{\text{IS}}_{q_{\text{risky}},10} \tag{3.29}$$

$$m_{\text{safe}} = u^{(-1)}(\overline{\Delta u}^{\text{IS}}_{q_{\text{safe}},10}) = (o_{\max} - o_{\min}) \cdot \overline{\Delta u}^{\text{IS}}_{q_{\text{safe}},10}. \tag{3.30}$$

Each value of m_{risky} corresponds to one participant's judgment of the fair price for the risky gamble and likewise for the values of m_{safe}.

To simulate choice, we applied the UWS model for binary decisions from description (Equations 3.20-3.21) with the parameters estimated from the Technion choice prediction tournament. To choose between the risky versus the safe gamble, this model estimates the expected differential utility $\mathbb{E}\left[u(O_{\text{risky}}) - u(O_{\text{safe}})\right]$ directly instead of estimating the gambles' expected utilities $\mathbb{E}\left[u(O_{\text{risky}})\right]$ and $\mathbb{E}\left[u(O_{\text{safe}})\right]$ separately. Consequently, it overweights pairs of outcomes whose utilities are very different instead of individual outcomes whose utilities are far from 0. Concretely, it simulates pairs of outcomes (i.e., one outcome for the risky gamble and one outcome for the safe gamble) according to the distribution q_Δ defined in Equation 3.20, which weights their joint probability by the absolute value of their difference in utility. The differential utilities $\Delta u_1, \cdots, \Delta u_{10}$ of the simulated outcome pairs are then translated into an estimate of the difference between the expected utility of the risky gamble versus the safe gamble according to Equation 3.21. If the resulting decision variable $\Delta \hat{U}^{\text{IS}}_{q_\Delta,10}$ is positive, the simulated decision-maker chooses the risky gamble, if it is negative they choose the safe gamble, and if it is 0 then they choose randomly.

Since the utilities $u(o)$ that drive the overweighting of extreme outcomes are stochastic (Equation 3.16), we conducted $100\,000$ simulations to average over a large number of utility-weighted sampling distributions q. Each simulation generated one price for the safe gamble, one price for the risky gamble, and one simulated choice between the two. At the beginning of each simulation, the utilities $u(0)$, $u(1)$, and $u(2)$ were drawn from $\mathcal{N}(\mu = \frac{o}{2}, \sigma = 0.17)$ for each possible outcome $o \in \{0, 1, 2\}$ and plugged into Equations 3.25–3.28 to yield the distributions the decision-maker would sample from in that simulation. Within each simulation, the sampled outcomes were evaluated by independent applications of the noisy utility function (Equation 3.16). Hence, even when the same outcome was sampled multiple times in a simulation, its subjective utility could be different every time.

UWS predicted that 42% of participants should reverse their risk preference from pricing to choice. In 66% of these reversals the model prices the risky gamble higher but choose the safe one.

As a result, utility-weighted sampling typically prices the risky gamble higher than the safe gamble (67% of the time), but it choses the safe gamble almost every second time (49% of the time). The rational decision mechanism of utility-weighted sampling weights events differently depending on whether it is tasked to perform pricing versus choice. Given that its shift in attention is a rational adaption to the task, the inconsistency between people's apparent risk preferences in pricing versus choice is consistent with resource-rationality.

While the laboratory experiments that demonstrated the effects simulated above can be criticized as artificial because their stakes were low or hypothetical, the overweighting of outcomes with extreme differential utility has also been observed in high-stakes, financial decisions whose outcomes do count (Post, van den Assem, Baltussen, & Thaler, 2008), and UWS can capture those effects as well (see Section "Deal or No Deal: Overweighting of extreme events in real-life high-stakes economic decisions" of Appendix B).

3.3.5 SUMMARY

In this section, we have shown that utility-weighted sampling accurately predicts people's decisions from description across a wide range of problems including those that elicit inconsistent risk preferences. Our utility-weighted sampling model of decisions from description rests on three assumptions: Its central assumption is that expected utilities are estimated by importance sampling. In addition, we assumed that binary choices from description are made by directly estimating the differential utility of choosing the first option over the second option. This assumption was important to predict the fourfold pattern of risk preferences, preference reversals, and the Allais paradox. Finally, we assumed that the mapping from payoffs to utilities is implemented by efficient coding. This assumption is not critical to the simulations reported here, but it will become important in our simulations of decisions from experience in the next section.

3.4 OVERWEIGHTING OF EXTREME EVENTS IN DECISIONS FROM EXPERIENCE

In decisions outside the laboratory we are rarely given a list of all possible outcomes and their respective probabilities. Instead, we have to estimate these probabilities from past experience. When people learn outcome probabilities from experience their risk preferences are systematically different than when the probabilities are described to them (Hertwig & Erev, 2009). For instance, people

overweight rare outcomes in decisions from description but tend to underweight them in decisions from experience (Hertwig, Barron, Weber, & Erev, 2004).

A common paradigm for studying decisions from experience is repeated binary choices with feedback. In this paradigm, the outcomes and their probabilities are initially unknown and must be learned from experience. Madan et al. (2014)(Madan et al., 2014) discovered an interesting memory bias in this paradigm: people remember extreme outcomes more often than moderate ones and overestimate their frequency. Ludvig et al. (2014) showed that people also overweight the same extreme outcomes in their decisions when their probability is $\frac{1}{2}$. Above we showed that utility-weighted sampling can account for the memory biases discovered by Madan et al. (2014), and in this section we investigate whether utility-weighted sampling can also account for the corresponding biases in decisions from experience by simulating the experiments by Ludvig et al. (2014). Our analysis suggests that biased memory encoding serves to help people make future decisions more efficiently by making the most important desiderata come to mind first.

Ludvig et al. (2014) conducted a series of four experiments. In each of the four experiments people made a series of decisions from experience. For instance, Experiment 1 comprised 5 blocks with 48 choices each. There were a total of four options: a sure gain of $+20$ points, a sure loss of -20 points, a risky gain offering a 50/50 chance of $+40$ or 0, and a risky loss offering a 50/50 chance of 0 or -40 points. In most trials participants either chose between the risky and the sure gain (gain trials) or between the risky and the sure loss (loss trials). After each choice subjects were shown the number of points earned, and they received no additional information about the options. Experiments 2-4 used different outcomes but were otherwise similar. In Experiment 2 the absolute values of all outcomes of Experiment 1 were shifted by 5 points. In Experiment 3 the gain and loss trials were supplemented by extreme gain trials and extreme loss trials whose outcomes were double the outcomes in Experiment 1. Experiment 4 had a loss condition in which all outcomes were losses (4L) and a gain condition in which all outcomes were gains (4G). Both conditions comprised risky gambles in which only the high outcome was extreme (HX), gambles in which only the low outcome was extreme (LX), and gambles in which both outcomes were extreme (BX).

To simulate these experiments, we assumed that Ludvig et al.'s participants had learned the outcome probabilities in the first four blocks and modeled their choice frequencies in the final block of each experiment. We can therefore model each individual decision as the choice between two lotteries each of which is defined by the value of the high outcome o^{high}, the probability p^{high} of receiving

it, and the low outcome o^{low}:

$$l_1 = \left(o_1^{\text{high}}, p_1^{\text{high}}, o_1^{\text{low}} \right) \tag{3.31}$$

$$l_2 = \left(o_2^{\text{high}}, p_2^{\text{high}}, o_2^{\text{low}} \right). \tag{3.32}$$

We model utility-weighted sampling as simulating s possible outcomes of each action a by sampling from the importance distribution defined in Equation 3.11:

$$\hat{o}_1^{(a)}, \cdots, \hat{o}_s^{(a)} \sim q(o|a) \propto p(o|a) \cdot |u(o) - \bar{u}|, \tag{3.33}$$

where $\bar{u}$ is the average outcome experienced by the participant. The simulated utilities are then combined into estimates of each action's expected utility gain according to Equation 3.12, and the option with the highest expected utility gain estimate is chosen. Our model defines the likelihood of individual choices in terms of two parameters: the number of samples s, and the noise variance σ_ε^2 of the brain's representation of utilities. We estimated these parameters from the choice frequencies in the final blocks of each condition of Experiments 1-4 by the maximum-likelihood method.

The results of fitting our model to the data of Ludvig et al. (Figure 3.5) revealed that utility-weighted sampling can capture the effects in all of the experiments with a single set of parameters (i.e. $s = 2$ samples, and a noise standard deviation of $\sigma_\epsilon = 0.65$) and the fit is robust to small changes in these parameters. Most importantly, utility-weighted sampling predicts that people are more risk seeking when the extreme outcome is high than when the extreme outcome is low. This explains why participants were more risk seeking for gains than for losses (Experiments 1-2). Experiment 3 combined trials in which the outcomes were twice the outcomes in Experiment 1 (3X) with the original trials from Experiment 1 (3NX). Our model correctly predicted the two main effects: more risk seeking on extreme gain trials than on extreme loss trials (3X) and a substantially smaller difference in risk seeking between their non-extreme counterparts (3NX).

UWS also captured the finding that the effect for the non-extreme outcomes is substantially smaller than in Experiment 1 even though the options were identical. According to our model, the context of the extreme outcomes in Experiment 3 suppresses the difference between the non-extreme gain and loss trials, because each outcome is divided by the range of all outcomes that need to be represented; see Equation 3.16. Since the range of outcomes is twice as large in Experiment 3 than in Experiment 1, the difference between the rewards of the non-extreme outcomes in Experiment 3 is only half as large as in Experiment 1. Consequently the noise in the reward signals can
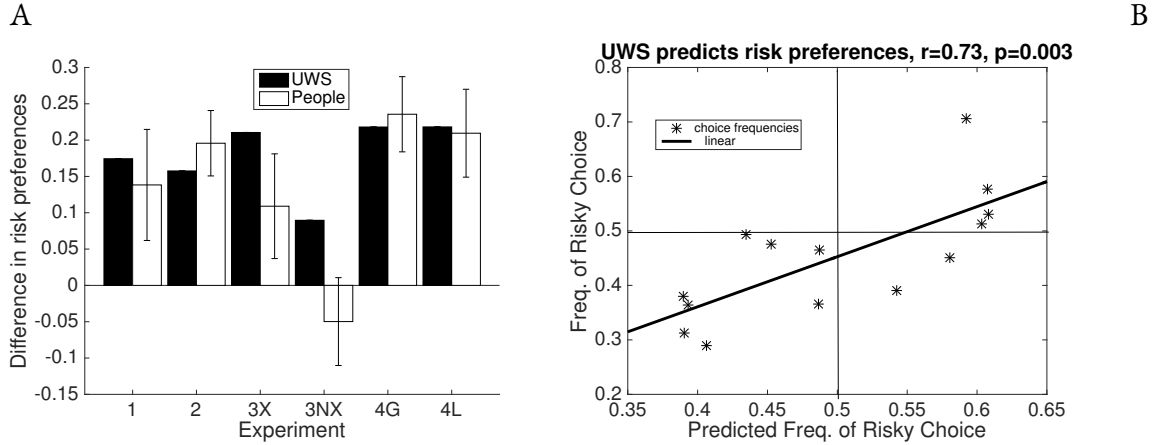
Figure 3.5: Differential risk preferences in the experiments by Ludvig et al. (2014). (A) Observed and predicted patterns of risk preferences. (B) Scatterplot combining all experimental conditions.

overturn the signal in Experiment 3NX more often than in Experiment 1. For Experiment 4 utility-weighted sampling correctly predicted more risk seeking when the high outcome was extreme and the low outcome was moderate (HX; $p_{\text{risky choice}} = 0.61$) than vice versa (LX; $p_{\text{risky choice}} = 0.39$), and an intermediate amount of risk seeking when both outcomes were extreme (BX, $p_{\text{risky choice}} = 0.49$). Utility-weighted sampling predicted this pattern of risk preferences regardless of whether all outcomes were gains (Experiment 4G) or all outcomes were losses (Experiment 4L). Utility-weighted sampling predicts all of these effects from the assumption that the brain's simulation mechanism is biased towards outcomes with extreme utility. Future models might be able to achieve a better fit, but to our knowledge utility-weighted sampling is the only theory to date that captures at least the qualitative effects observed by Ludvig et al. (2014).

In the experiments by Ludvig et al. (2014) all outcome probabilities were equal to 0.5. In prospect theory (Kahneman & Tversky, 1979) probability weighting only depends on the magnitude of the probability. Hence, it cannot overweight the 50% chance of one event and underweight the 50% chance of the other event at the same time. UWS, by contrast, can explain the effects, because it predicts that extreme events will always be overweighted regardless of their probability. This highlights a critical difference between UWS and prospect theory: In prospect theory over- versus underweighting depends on the value of the probability but is independent of the utility. By contrast, in UWS the over- or under-weighting is determined by the outcome's utility but is independent of its probability. Cumulative prospect theory (Tversky & Kahneman, 1992) captures the effect of extremity on overweighting in principle, but it doesn't capture this effect when there are only two possible

outcomes.

To apply our theory to the empirical data, we had to choose a utility function. We chose the stochastic normalized utility function defined in Equation 3.16 because of its neuroscientific underpinnings and its ability to explain context-sensitive preferences in value-based decision-making (Summerfield & Tsetsos, 2015). Concretely, UWS combined with a context-insensitive utility function, such as a simple linear function of the outcome, or the concave utility function of prospect theory, would be unable to explain why people's preference for the risky gamble +40/0 over the safe option +20 is lower in Experiment 3 than in Experiment 1 even though the choices are exactly the same. In addition to the normalization by the dynamic range, the noise term is also necessary, because otherwise any scaling of the utility function is canceled out by the normalization of the sampling distribution. Therefore, there appears to be no simpler or more conventional utility function that can explain the qualitative features of the data of Ludvig et al. (2014) than the normalized stochastic utility function defined in Equation 3.16. Given this utility function, UWS predicts that the overweighting of the gain (+40) in the choice between a 50/50 chance to gain 40 or 0 vs. 20 for sure in Experiment 1 would disappear if there were only gain trials so that the average outcome would be 20 which is exactly in the middle between 0 and 40.

3.5   Utility-weighted learning from experience

So far, we have shown that utility-weighted sampling can capture biases in frequency judgment, decision-making, and memory recall. Our explanation postulates that the brain samples from an importance distribution that weights each outcome's probability by the absolute value of the extremity of the outcome's utility, but it remains unclear whether and how the brain could implement this mechanism. We have speculated that there may be a common root to these biases: the enhancement of learning by emotional salience. Consistent with this mechanism, memory consolidation is enhanced when the reward associated with an experience is larger (Adcock, Thangavel, Whitfield-Gabrieli, Knutson, & Gabrieli, 2006). Adcock et al. (2006) found that this modulation of memory consolidation is mediated by the release of dopamine from the ventral tegmental area. The enhancement of learning by emotional salience implies that extreme events, such as the terrorism, natural disasters, and traumatic accidents, are engraved more deeply into our memory than mundane events. A single extreme experience, such as a traumatic event, in a neutral context can instill an enduring association that is much stronger than the association formed with a mundane event that occurred more frequently in the same context. Based on this idea we propose a biologically-plausible learning mech-

anism that tunes neural networks to sample from the importance distribution of utility-weighted sampling.

3.5.1 UWS CAN EMERGE FROM REWARD-MODULATED ASSOCIATIVE PLASTICITY

Utility-weighted sampling can be implemented using a stochastic winner-take-all network (Nessler, Pfeiffer, Buesing, & Maass, 2013, c.f.) whose units represent potential outcomes and receive inputs from units representing the alternatives of the choice (a). The weight $w_{a,o}$ of connection between the input units representing alternative a and the output units representing outcome o encode the strength of the association between alternative a and outcome o. The weights w thereby determine the relative frequency with which the network simulates each outcome for each alternative. In this section, we propose a learning rule for the weights w that tunes the network to simulate outcomes according to utility-weighted sampling (Equation 3.33).

We assume that the initial association strengths w are zero, and that choosing an alternative a and receiving a rewarding outcome o reinforces their association $w_{a,o}$. The association strengthens more the more surprising the outcome is (Courville, Daw, & Touretzky, 2006). Our model captures this effect by updates that are proportional to the absolute value of the reward prediction error $\mathrm{PE}(o)$:

$$w_{a,o}(t+1) = \begin{cases} (1-\gamma) \cdot (w_{a,o}(t) + \alpha \cdot |\mathrm{PE}(o)|) & \text{if } A(t) = a \text{ and } O(t) = o \\ (1-\gamma) \cdot w_{a,o}(t) & \text{if } A(t) = a \text{ and } O(t) \neq o \end{cases}, \qquad (3.34)$$

where $A(t)$ and $O(t)$ are the chosen alternative and the outcome in trial t, α is the learning rate, and γ is the forgetting rate. The reward prediction error is the difference between the experienced reward $r(o)$ and reward expectancy $\bar{r}(t)$:

$$\mathrm{PE}(o) = r(o_r) - \bar{r}(t), \qquad (3.35)$$

where $r\left(o(t)\right)$ is the subjective utility of outcome o defined in Equation 3.16, and $\bar{r}(t)$ is the reward expectancy $\bar{u}(t)$ associated with any trial in the experiment. It can therefore be thought of as a recency-weighted average over all rewards regardless of the choices that generated them. We assume that this expectancy is learned independently from the alternative-outcome associations by temporal difference learning, that is

$$\bar{u}(t+1) = \bar{u}(t) + \eta \cdot \mathrm{PE}, \qquad (3.36)$$

where η is a learning rate and the reward prediction error PE is conveyed by phasic dopamine signals from the ventral tegmental area to the ventral striatum and the frontal lobe (Niv, 2009). This concludes the learning part of our model.

To model decision-making we assume that the rate at which units representing alternative a activate units representing outcome o is proportional to the strength of their connection, that is

$$P(\hat{O} = o|A = a) = \frac{w_{a,o}}{\sum_{o=1}^{n} w_{a,o}} \propto w_{a,o} \tag{3.37}$$

The learning rule (Equation 3.34) increases the weight $w_{a,o}$ with probability $p(o|a)$ by an increment proportional to $|\text{PE}(o)|$. Therefore, the probability that outcome o will be simulated when considering action a (i..e, $P(\hat{O} = o|A = a)$) converges to $p(o|a) \cdot |\text{PE}(o)| = p(o|a) \cdot |u(o) - \bar{u}| \propto q^{\text{UWS}}$, where $u(o) = r(o)$. In this way, the network gradually learns to perform utility-weighted sampling (Equation 3.11). The simulated outcomes could be read out by a decision network that chooses the alternative with the highest value of the utility estimate defined in Equation 3.12. Thus, after sufficient learning the simulation network and the decision network jointly perform utility-weighted sampling. The above equations are meant as an abstract specification of network properties rather than the definition of a concrete neural network, but they suggest a way in which the brain might learn to perform utility-weighted sampling.

Having proposed a learning mechanism that can give rise to utility-weighted sampling, we will now evaluate its predictions against the temporal dynamics of people's risk preferences in repeated decisions from experience.

3.5.2   TEMPORAL DYNAMICS OF RISK PREFERENCES

Above, we simulated people's risk preferences in the final blocks of the experiments by Ludvig et al. (2014) assuming that the participants had already learned the utility-weighted sampling distribution. Here, we test whether the utility-weighted learning (UWL) model can predict this learning outcome and capture the temporal evolution of people's risk preferences from the first block through the last block. The utility-weighted learning model predicts participants' choice probabilities as a function of seven parameters: the number of samples s, the uncertainty σ_ε about utilities, the learning rate α, the forgetting rate γ, the initial reward expectancy $\bar{r}(0)$, the rate η at which the reward expectancy $\bar{r}$ is being updated, and the probability of random choice p_{random}. To estimate these parameters, we fitted the block-by-block choice frequencies reported by Ludvig et al. (2014) by maximum-likelihood

estimation.

The parameter estimates were $s = 1$ samples, learning rate $\alpha = 1$, forgetting rate $\gamma = 0.375$, noise standard deviation $\sigma_\varepsilon = 0.1$, initial reward expectancy 3, TD learning rate $\eta = 0.05$, probability of random choice $p_{random} = 0.64$. We found that utility-weighted learning captures several qualitative properties of how people's risk preferences changes with experience: Our simulations of Experiments 1-2 captured that people gradually become more risk-averse on loss trials but more risk-seeking on gain trials (Figure 3.6A). Our simulations of Experiment 3 captured that this effect is reduced when gains and losses are non-extreme in the context in which they occur (Figure 3.6B), and the simulation of Experiment 4 captured that more experience makes people more risk-seeking when the high outcome is extreme, but more risk-averse when the low outcome is extreme, even if all outcomes are gains or all outcomes are losses (Figure 3.6C). According to utility-weighted learning the determinant of risk-seeking is that the high outcome is farther away from the learned reward expectancy than the low outcome. The reward expectancy tracks to average across all recent outcomes. Thus, UWL predicts risk seeking when the high outcome is farther away from the average outcome than the low outcome.

3.5.3 PREDICTING MEMORY BIASES

Earlier in this chapter, the experiments by Madan et al. (2014) was simulated according to utility-weighted sampling. We found that UWS correctly predicted the qualitative differences between moderate and extreme events in frequency estimation and memory recall, but its predictions were more extreme than the biases observed in people. In this section we revisit these effects with the utility-weighted learning model. In addition, the utility-weighted learning model also allows us to simulate the relationship between memory biases and risk preferences, as well as the effect of recent outcomes on risky choice.

Concretely, we fitted the UWL model to the block-by-block choice frequencies in Experiments 1 and 2 by Madan et al. (2014) using the maximum-likelihood method. We then used the resulting parameter estimates to predict participants' frequency estimates and memory biases. To do so, we modeled people's frequency estimates according to utility-weighted sampling as defined in Equation 3.22. Likewise, participants' answers to the memory recall question were modeled by the outcome that was sampled most frequently; if two or more outcomes occurred equally frequently one of them was chosen at random.

A

B

C

Ludvig et al. (2014), Experiment 4
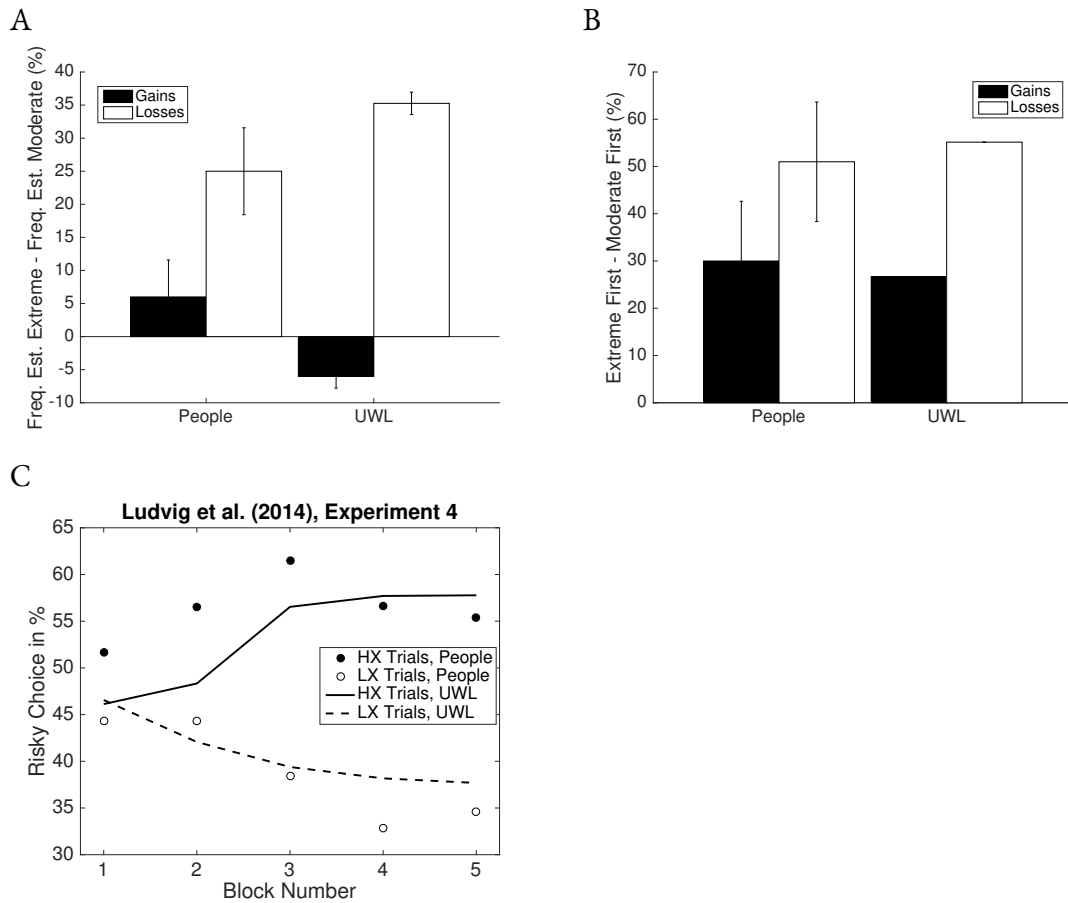
Figure 3.6: The utility-weighted learning (UWL) model captures the temporal evolution of people's risk preferences in the experiments by Ludvig et al. (2014). (A) UWL captures that participants in Experiments 1-2 became increasingly more risk seeking on gain trials but more risk averse on loss trials. (B) UWL captures that in Experiment 3 risk seeking increased primarily on trials with extreme outcomes. (C) UWL captures that in Experiment 4 people became risk seeking when the larger outcome was extreme and risk averse when the smaller outcome was extreme regardless of whether the outcomes were gains or losses.

The maximum likelihood parameter estimates indicated increased accuracy motivation: more simulations ($s = 2$), faster learning ($\alpha = 9$), and slower forgetting ($\gamma = 0$). The estimated standard deviation of the noise was $\sigma_\varepsilon = 0.1$, the estimated initial reward expectancy $\bar{r}(0)$ was 7, the estimated rate at which the reward expectancy is updated was 0.5, and the estimated probability of random choice was 0. With these parameters our model captured people's memory biases (see Figure 3.7) and their relationship with risk seeking: Even though the risky choice generated the moderate outcome (0 points) and the extreme outcome (± 40 points) equally often, for most people the extreme outcome came to mind first (Figure 3.7B), and their frequency estimates were significantly higher for the extreme loss than for the moderate outcome (Figure 3.7A). This was not the case for the high gain (+40), because according to the parameter estimates participants entered the experiment with the expectation that outcomes would average 560 points. As a comparison with Table 3.1 shows, the predictions of UWL are closer to the empirical data than the predictions of the basic UWS model.

In addition, our model correctly predicted that people who recalled the extreme gain first were more risk seeking on gain trials than people who remembered the moderate outcome first ($56.32 \pm 0.24\%$ vs. $50.83 \pm 0.26\%$ risky choices) whereas people who remembered the extreme loss first were less risk seeking on loss trials than people who remembered the moderate outcome first ($31.83\% \pm 0.34\%$ vs. $33.67 \pm 0.34\%$ risky choices). The simulated frequency estimates were significantly correlated with the model's preference for the risky option: The higher the model estimated the frequency of the extreme loss to be the fewer risky choices it made on loss trials ($r = -0.4419, p < 10^{-15}$). Conversely, risk seeking on gain trials increased with the estimated frequency of the extreme gain ($r = 0.23, p < 10^{-15}$). Utility-weighted learning also captured that people were more risk seeking when the most recent risky choice in the same context yielded the good outcome than when it yielded the bad outcome: For gain trials UWS predicted 8.6% higher risk seeking after receiving the high gain (+40) than after winning nothing on the previous risky gain trial. Conversely, UWS predicted 6.0% less risk seeking following the large loss (-40) compared to no loss on the previous risky loss trial.

Finally, we simulated Experiment 2 from Madan et al. (2014) according to the same parameters. This experiment was identical to Experiment 1 except that all outcomes were shifted by +40 points so that there were no negative outcomes. Our model correctly predicted that this manipulation changes none of the qualitative effects observed in Experiment 1, and our model now correctly predicted that people overestimate the frequency of the extreme gain relative to the neutral outcome (UWS: 56.4% versus 43.2%).
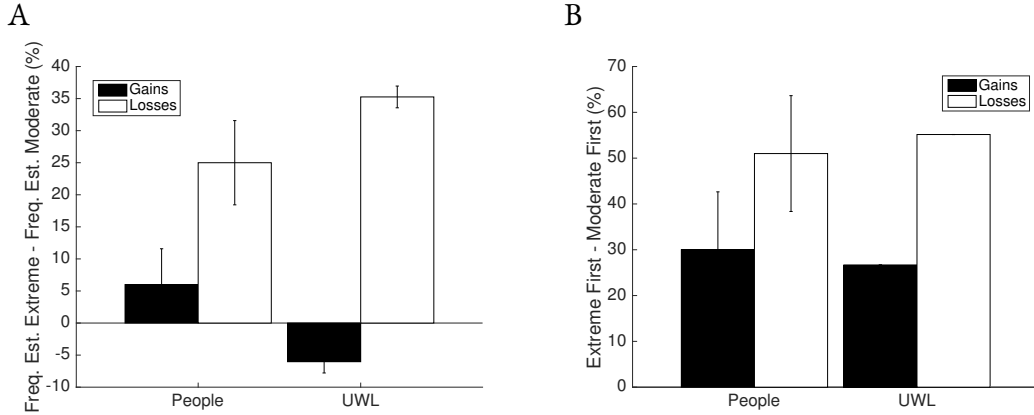
Figure 3.7: Utility-weighted learning (UWL) predicts the biased memory recall and frequency estimates observed by Madan et al. (2014). Error bars denote 95% confidence intervals. (A) Difference between estimated frequencies of extreme versus moderate outcomes. (B) Proportion of people who recalled the extreme outcome first minus proportion who recalled the moderate one first.

3.5.4 VALIDATION ON DECISIONS FROM EXPERIENCE

Having shown that UWL predicts people's biases in memory recall and frequency estimation more accurately than the original UWS model and captures the temporal dynamics of people's risk preferences in repeated decisions from experience and the effect of recent outcomes on risky choice, we now evaluate UWL against alternative models of repeated decisions from experience. To do so, we use data from the Technion choice prediction tournament as we did for our basic utility-weighted sampling model of decisions from description. As before, we fit our utility-weighted learning model to the training set by maximum-likelihood estimation, evaluate its predictive accuracy on the test set, and perform formal model comparisons against the best models from the competition. The only difference is that we now use the data sets and models from the Technion tournament on repeated decisions from experience rather than decisions from description.

The parameter estimates were as follows: learning rate $\alpha = 2$, number of samples $s = 9$, forgetting rate $\gamma = 0$, standard deviation of the noise $\sigma_\varepsilon = 0.1$, probability of random choice $p_{\text{random}} = 0.12$, initial reward expectancy $\bar{r}(0) = 3$, and $\eta = 0.05$. We set our model's parameters to these values and evaluated its predictions against people's choice frequencies on the test set; see Figure 3.8. Our model's predictions agreed with people's risk preferences for 90% of the decision problems. The correlation between the predicted and observed choice frequencies was
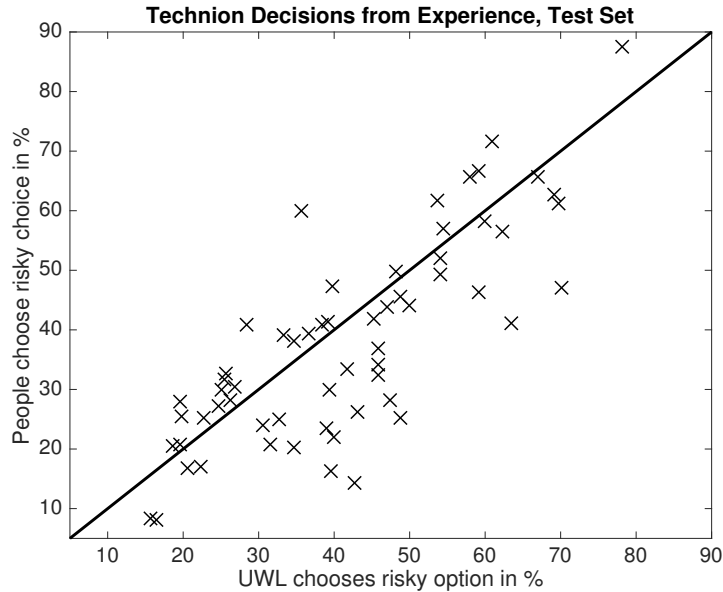
Figure 3.8: Utility-weighted learning (UWL) predictions for test set of Technion prediction tournament for repeated decisions from experience. Each data point reports the risky choice frequency of UWL (horizontal axis) versus people (vertical axis) for one of the 60 decision problems, and the solid line is the identity line.

$r = 0.80$, and the mean-squared error of the predicted choice frequencies was MSD $= 0.0120$. Our model thereby explained the data substantially better than the basic reinforcement learning model that Erev et al. (2010) considered as a baseline (66% agreement, $r = 0.51$, MSD $= 0.0263$; $t(59) = -3.2, p = .002$), and not significantly worse than the best model in the competition: the explorative sampler with recency (MSD $= 0.0066, t(59) = 1.65, p = .1, 86\%$ agreement, $r = 0.89$). While the best model was provided as a baseline, the best submission was the ACT-R model of instance-based learning (MSD: 0.08, $r = 0.89$). After the competition, Lejarraga, Dutt, and Gonzalez (2012) introduced an improved instance-based learning model that performed slightly better than the exploratory sampler with recency (MSD $= 0.006, 86\%$ agreement, $r = 0.89$) and its mean-squared error was significantly lower than that of our model ($t(118) = -2.21, p = 0.01$). The predictive accuracy of the normalized reinforcement learning model was comparable to the performance of our model (MSD: 0.0087, 84% agreement, $r = 0.84$). Additional analyses comparing the risk preferences of UWL to those of people are provided in the Appendix B.

3.5.5 Discussion

We hypothesized that utility-weighted sampling arises from biased memory encoding. In this section, we formalized this proposal by a biologically-plausible learning rule that we call utility-weighted learning (UWL). The empirical data of Ludvig et al. (2014) and Madan et al. (2014) provides four strong pieces of evidence for our hypothesis that the over-representation of extreme events results from utility-weighted memory encoding: First, people overweight outcomes with extreme utilities in decisions from experience relative to equally probably outcomes with moderate utilities. Second, this overweighting emerged gradually through learning and the time course of learning matched the predictions of our utility-weighted-learning model. Third, participants displayed biases in memory recall that matched the biases of their decisions and our model captured both. Fourth, as predicted by our model, there was a significant correlation between the magnitude of each participant's bias in memory recall and the bias in their choice frequencies. This is consistent with our model's assumption that the overweighting of events with extreme utilities and their heightened availability in memory have a common cause: utility-weighted memory encoding. While the correlation between biases in memory and choice does not imply causation, our model's assumption that utility-weighted memory encoding causes memory biases that in turn cause biases in decision-making does offer a plausible explanation for this phenomenon. Under this assumption the covariation of the ease with which extreme events come to mind could plausible arise from individual differences in the sensitivity to reward and punishment (Corr, 2004): The higher a person's reward sensitivity, the more biased their memory encoding will be. The more biased the strengths of a person's memories are in favor of extreme events, the more easily they will be recalled, and this in turn increases their decision weights.

We found that our model explained the temporal dynamics of of people's risk preferences and memory biases in repeated decisions from experience and evaluated the utility-weighted learning model against people's choice frequencies in a wide range of decisions problems. UWL was competitive with the best existing models of decisions from experience. Together with the findings presented in previous sections, the results in this section show that utility-weighted sampling can provide a unifying, mechanistic explanation for a wide range of biases in decisions from description and decisions from experience. This is important for two reasons. First, it is often implied that decisions from description and decisions from experience rely on separate mechanisms, and second our most influential theories of decision-making are *not* mechanistic.

Although the experiments simulated here had only two possible outcomes, the UWL learning

model is equally applicable to decisions with many possible outcomes and one example thereof can be found in the Section "Payoff-variability effects in decisions with very many possible outcomes" of the Appendix B.

The proposed learning mechanism is similar to the Pearce-Hall model of classical conditioning (Pearce & Hall, 1980) in that both update the strength of a stimulus-reward association by an amount proportional to the absolute value of a reward prediction error. However, there are several important differences. Most importantly, our model learns the conditional probabilities of multiple possible outcomes given a single cue whereas the Pearce-Hall model learns to predict the intensity of a single reward or punishment given multiple cues. Consequently, in the Pearce-Hall model, the reward prediction is derived from the learned associations. By contrast, in our model the reward prediction is learned independently of the cue-outcome associations. Furthermore, the Pearce-Hall model uses the reward prediction error from the previous trial whereas our model uses the reward prediction error from the current trial. The two models also differ in the remaining terms of their learning rules.

To fit the temporal dynamics of risk preferences with learning, we had to make a number of assumptions about the underlying learning mechanisms. The details of this proposal are not essential to our theory and may be revised and simplified in future versions of the utility-weighted learning model. Instead, the utility-weighted learning model should be seen as a proof of principle that utility-weighted sampling can emerge from reward-modulated associative learning in the brain.

3.6   General Discussion

While the resource-rational analysis presented in the previous chapter addressed the question "How long should you think given that your time is valuable?", the resource-rational analysis presented in this chapter answered the complementary question "Given that you can consider only a limited number of things, which ones should you think about?". The analysis revealed that in order to take the most important outcomes into account most of the time, it is necessary for us to think about the potential outcomes of our decisions in a way that is biased towards events that are extremely good or extremely bad. According to this argument, our seemingly irrational availability biases are a rational adaptation to the constraints imposed by our limited time and finite processing speed. These cognitive constraints, make utility-weighted sampling a resource-rational mechanisms for decision-making under risk and uncertainty. The analysis presented here incorporated an additional

cognitive constraint, namely the finite representational resources that are available to encode the outcome of each simulation. The rational use of this finite representational bandwidth by efficient coding scales reward values by their dynamic range (Summerfield & Tsetsos, 2015). This makes extremity context-dependent, thereby explaining why potential outcomes that are over-weighted in some contexts are under-weighted in others.

Utility-weighted sampling explains not only how we are able to make sensible decisions under severe time pressure but also why we overestimate the frequency of extreme events and have inconsistent risk preferences. Utility-weighted sampling explains why extreme events come to mind first and why people overestimate their frequencies and overweight them in decisions under uncertainty. Our model captures how people's risk preferences depend on valence (gains versus losses), probability, the elicitation method (pricing versus choice), and on whether probabilities are described or experienced. Utility-weighted sampling can thus explain preference reversals, the Allais paradox, and the fourfold pattern of risk preferences. In addition, our utility-weighted learning model captures the temporal dynamics of people's risk preferences during repeated decisions from experience, the effect of recent outcomes on risky choice, and the relationship between memory biases and risk preferences.

Our model's predictive validity in the Technion choice prediction tournaments for repeated decisions from description and decisions from experience was competitive with, although not quite as good as, the fit of the models that won these competitions. Yet, while most of these models were specific to their competition, our model was derived from first principles, it also applies to more complex decisions with (infinitely) many possible outcomes, and it can simultaneously explain a much wider range of biases in decision-making, judgment, and memory than ever attempted before. In addition, our model does not just describe risk preferences but specifies the underlying (neuro)computational mechanisms. The biases explained by our model include newly discovered phenomena (Ludvig et al., 2014; Madan et al., 2014) that have not been modeled before as well as classic findings that were previously explained separately.

The remainder of this chapter synthesizes and discusses the results presented above. We start by showing that the difference between our theory's predictions for decisions from description versus decisions from experience captures the description-experience gap. We then discuss the similarities and differences between UWS and previous theories of inconsistent risk preferences. Afterwards, we take a step back and discuss how the work presented here instantiates the general resource-rational approach to modeling cognitive mechanisms. Next, we discuss the connections between our theory and theory of ecological rationality. Finally, we acknowledge the limitations of our analysis, discuss

directions for future work, and conclude.

3.6.1 UTILITY-WEIGHTED SAMPLING CAPTURES THE DESCRIPTION-EXPERIENCE GAP

People's risk preferences in decisions from description and decisions from experience are systematically different. This difference is known as the *description-experience gap* (Hertwig & Erev, 2009). Most prominently, people appear to overweight small probabilities in decisions from description but underweight them in decisions from experience. Having applied utility-weighted sampling to decisions from description and decisions from experience, we are now in a position to evaluate whether the difference between the UWS model of binary decisions from description (Equations 3.20-3.21) and the utility-weighted learning model (Equations 3.34-3.37) captures the description experience gap. To do so, we computed the difference between the two models' predictions on the test set of the Technion choice prediction tournament (Figure 3.3 versus Figure 3.8) and compared it against the difference between people's choice frequencies in these two conditions.

Figure 3.9 shows that the difference between the two models correctly predicted the sign of the description-experience gap on 95% of the decision problems in the test set of the Technion choice prediction tournament. The correlation between the predicted and the actual description-experience gaps was $r = 0.8853$ ($p < 10^{-15}$), and the mean squared deviation was 0.0361. Our model of decisions from experience captures the effects of either not experiencing, or gradually forgetting rare outcomes. This explains why rare events tend to receive less weight in decisions from experience than in decisions from description. For instance, in problems 1-5 where the probability of the high outcome is at most 0.1, people and utility-weighted sampling are more risk-seeking when the probabilities are described than when they are experienced. According to our models, there is another difference: In decisions from description people over-simulate eventualities in which the outcomes of two choices are extremely different. In decisions from experience, by contrast, people simulate the possible outcomes of each option independently, so that utility-weighted sampling over-simulates each option's most extreme outcome even when they are identical. Thus, when choosing between losing a moderate amount for sure and the chance of winning a small amount or losing a large amount, UWS is more risk seeking in decisions from description than in decisions from experience, and this correctly predicts the positive description-experience gap in problems 30-34 (see Figure 3.9 and Erev et al., 2010). According to our theory, the description-experience gap is not only due to the fact that rare events in decisions from experience sometimes go unnoticed or are gradually discounted or forgotten but also due to difference between overweighting unusually large and un-
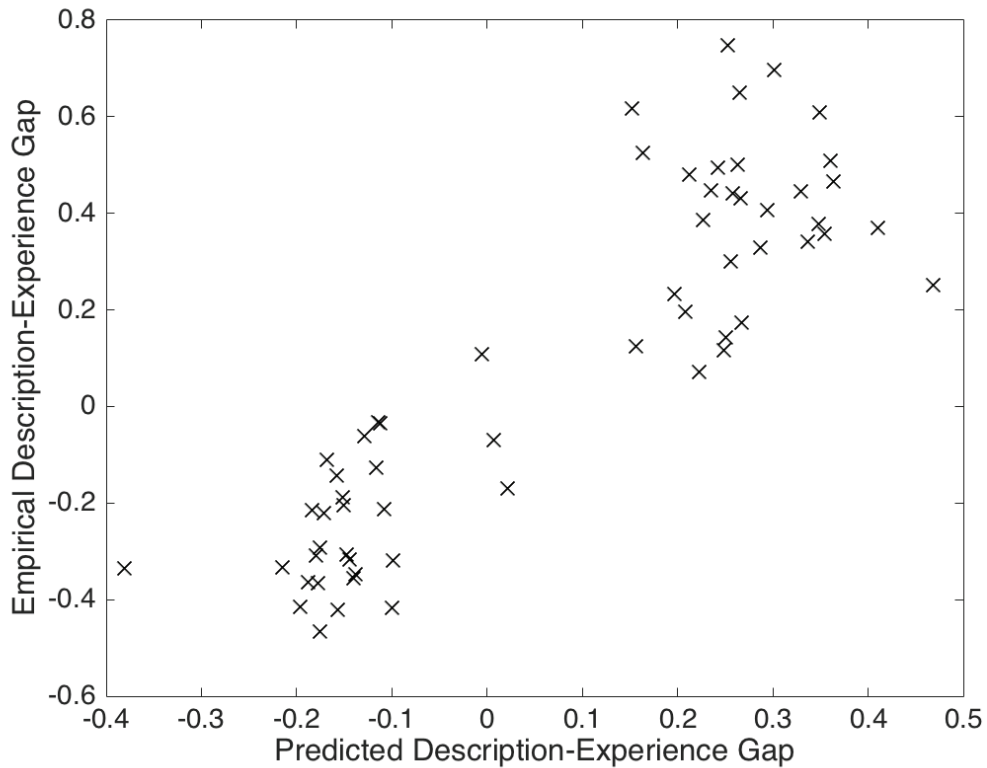
Figure 3.9: Utility-weighted sampling captures the gap between people's risk preferences in decisions from description and decisions from experience. 95% agreement, $r = 0.8853$, mean-squared error 0.0361.

usually small outcomes in decisions from experience versus overweighting of pairs of outcomes with large utility differences in binary decisions from description. Recent empirical evidence for the important contribution of memory biases in favor of extreme events to the description-experience gap (Madan, Ludvig, & Spetch, 2016) strongly supports our model's explanation. Furthermore, Kellen, Pachur, and Hertwig (2016) found that people are more sensitive to the payoffs and less sensitive to their probabilities in decisions from experience than in decisions from description even when the difference between experienced frequencies and described probabilities is controlled for. This too is consistent with the overweighting of extreme payoffs in decisions from experience.

3.6.2 COMPARISON TO PREVIOUS THEORIES OF JUDGMENT AND DECISION-MAKING

Unlike previous theories of decision-making, our model is both *normative and mechanistic*. In contrast to descriptive theories of choice, our approach has been to explore the implications of limited cognitive resources for the mechanisms by which people *should* make decisions under uncertainty. In contrast to most normative theories of choice, we have engaged with people's limited cognitive resources and derived a process model. This makes our theory the first rational process model (Griffiths et al., 2015) of cognitive biases in decision-making. The proposed mechanism for decisions from experience is psychologically plausible in that it relies on the well-known availability bias in memory recall (Tversky & Kahneman, 1973). Furthermore, we have shown that UWS naturally emerges from a biologically-plausible reward-modulated associative plasticity mechanism that is driven by the reward prediction error conveyed by dopamine (Schultz et al., 1997). But unlike most process models, UWS was derived from first principles and instantiates rational information processing.

Our theory provides the first rational perspective on the heightened availability of extreme events and the cognitive biases in judgment and decision-making that result from it. We have shown that it can explain a wide range of phenomena in memory, judgment, learning, decisions from description, and decisions from experience. Subsets of these phenomena, such as the simulated violations of expected utility theory in decisions from description were already accounted for by previous theories, but our model is the first to provide a unifying explanation for all of them, and none of the previous theories could explain why events with extreme utilities should be remembered first and sway people's decisions. As far as we know, UWS is the first theory that can simultaneously explain decisions from description and decisions from experience, and it reconciles the discrepancies between them. In particular, no previous theory was able to reconcile the reflection effect in decisions from description (risk aversion for a 50% chance of a large gain but risk seeking for a 50% chance of a large loss) with the exact opposite of this effect in decisions from experience (Ludvig et al., 2014; Madan et al., 2016). We think that our theory is unique in providing the first rational process model of availability biases in judgment and decision-making and offering a unifying explanation for a very wide range of seemingly disparate phenomena, but it builds on previous work (Bordalo et al., 2012; Griffiths et al., 2015; Hertwig et al., 2005; Lichtenstein et al., 1978; Ludvig et al., 2014; Madan et al., 2014; Pachur et al., 2012; N. Stewart et al., 2006; Tversky & Kahneman, 1973; Vul et al., 2014) and has commonalities with many existing theories of judgment and decision-making. We provide a detailed discussion of how our theory is similar to and different from previous accounts of memory, frequency judgment, decisions from description, and decisions from experience in the Appendix B. Table 3.5 summarizes these comparisons in terms of the range of phenomena explained by UWS and some previous mod-

els and theories, namely the availability-by-recall model (Hertwig et al., 2005; Pachur et al., 2012), the regressed-frequency model (Hertwig et al., 2005), the value-assessment model (Barron & Erev, 2003), instance-based learning theory (T. C. Stewart, West, & Lebiere, 2009), the exploratory sampler with recency (Erev et al., 2010), the contingent average and trend (CAT) model (Plonsky, Teodorescu, & Erev, 2015), the decision-by-sampling model (N. Stewart et al., 2006), salience theory (Bordalo et al., 2012), the priority heuristic (Brandstätter et al., 2006), regret theory (Loomes & Sugden, 1982), prospect theory (Kahneman & Tversky, 1979), stochastic cumulative prospect theory (SCPT; Erev, et al., 2010), dynamic prospect theory (Post et al., 2008), disappointment theory (Bell, 1985; Loomes & Sugden, 1984, 1986), and the 3-moments model (Allais, 1979; Hagen, 1979). These and other comparisons suggest that UWS is the first mathematical theory to provide a unifying explanation for availability biases in frequency judgment, memory, decisions from experience, and decisions from description.

3.6.3   Resource-rationality

We derived utility-weighted sampling by resource-rational analysis (Griffiths et al., 2015): We first defined the function of decision-making. Second, we modeled people's cognitive capacities by an abstract computational architecture that can simulate outcomes by sampling, evaluate their utility, combine the simulated utilities into an estimate of each action's expected utility by a weighted average, and choose the action with the highest utility estimate. In addition, we assumed that time constraints and cognitive capacity severely limit the number of simulations the mind can perform. Third, we derived an approximately optimal strategy for allocating the architecture's computational resources. Finally, we evaluated our original proposal (Lieder, Hsu, & Griffiths, 2014) against empirical data and alternative models of decision-making under uncertainty and refined it by making the utility-function context sensitive. Consistent with previous results (Vul et al., 2014), we also found that people appear to perform more simulations for high-stakes decisions (see Section "Deal or No Deal: Overweighting of extreme events in real-life high-stakes economic decisions" of the Appendix B) than for low-stakes decisions (Technion choice prediction tournament). Furthermore, simulations reported in the Appendix B showed that UWS captures that people's decision quality approaches optimality as the difference between their options increases. Overall, we found that the availability biases and inconsistent risk preferences modeled in this chapter can be reconciled with the rational use of cognitive resources (Griffiths et al., 2015).

Our rational analysis assumed that people's judgments and decisions are based on sampling. We

Table 3.5: Comparison of the range of phenomena explained by UWS versus previous theories.

Category	Phenomenon	UWS	Availability-by-recall	Regressed-frequency	Value assessment	Instance-based learning	Exploratory-sampler with recency	CAT	Decision-by-Sampling	Salience theory	Priority Heuristic	Regret theory	Prospect theory	SCPT	Dynamic Prospect Theory	Disappointment theory	3-moments model
Memory & Judgment	Memory bias for extreme events	✓															
	Overestimation of rare extreme events	✓	✓	✓													
	Overestimation of frequent extreme events	✓	✓														
Decisions from experience	Reversed reflection effect	✓															
	Temporal dynamics of risk preferences	✓			✓	✓	✓	✓							✓		
	Underweighting of rare events	✓			✓	✓	✓	✓									
	Payoff variability effect	✓			✓	✓	✓	✓									
	Wavy recency effect						✓										
Decisions from description	Reflection Effect	✓							✓	✓	✓	✓	✓	✓	✓	✓	✓
	Allais paradox	✓							✓	✓	✓	✓	✓	✓	✓	✓	✓
	Preference reversals	✓								✓		✓					
	Intransitivity	✓								✓	✓	✓					
	Common-ratio effects								✓	✓	✓	.5	✓	✓	✓	.5	✓
	Gradual effect of choice difficulty	✓							✓						✓	✓	
	Description-experience gap	✓															

Note: A checkmark means that theory can qualitatively account for the phenomenon, and '.5' means that theory can qualitatively account for a subset of the phenomena.

view sampling as a rational computational mechanism for approximating the expected utilities in decision problems with many possible outcomes whose probabilities have to be estimated from experience. This characterization holds for most everyday decisions. This suggests that utility-weighted sampling might be a resource-rational strategy for the decisions people make in real life. By contrast, when choosing between simple gambles with numerically stated outcome probabilities and payoffs, people could, in principle, compute each gamble's expected value and choose the gamble with the highest expected value. When the stakes are high enough to offset the additional time and effort required to compute expected values then the expected value strategy would become resource-rational and participants should apply it. Is it therefore a sign of irrationality when people use utility weighted sampling in decisions from description? On the one hand, it appears suboptimal that people use sampling in simple decisions from description instead of relying on arithmetic. On the other hand, decisions from description are very rare outside the laboratory and resource-rationality is defined with respect to the distribution of problems in the agent's natural environment. Furthermore, the payoffs used in the decisions from description paradigm are usually small or hypothetical, and people's application of mathematical procedures is often error prone, slow, and effortful. We therefore believe that people's use of utility-weighted sampling in the simulated decisions from description is not necessarily inconsistent with resource-rationality.

Our results should be taken with a grain of salt, because there is no guarantee that the parameter estimates for which our model captures empirical phenomena accurately reflect the resource-limitations of the human brain. We cannot rule out that the actual opportunity cost of simulating an outcome is so low that it would be resource-rational for people to generate so many samples that their availability biases should be much smaller than they are. Hence, without independent measurements of the available cognitive resources we cannot conclude that people are resource-rational but only that the simulated cognitive biases could be resource-rational *in principle*. To complete our theory of resource-rational decision-making, future work will have to provide a precise specification of the available cognitive resources and their costs as well as a mechanism that determines the optimal number of samples. We will discuss these limitations and future directions in more detail below.

Pushing our abstract computational model further towards the algorithmic and implementational level (Marr, 1982), we have shown that utility-weighted sampling can emerge from reward-modulated associative learning during repeated decisions from experience. Our learning rule assumes that synaptic plasticity is modulated by the absolute value of the reward prediction error (Equation 3.34) which can be interpreted as surprise or emotional salience. The success of the utility-weighted learning model might suggest that people gradually learn to make more rational use of

their finite cognitive resources and that emotion contributes to the emergence of resource-rational decision-making. A recent neuroimaging study discovered a neural correlate of the absolute reward prediction error in the basolateral amygdala (Roesch, Esber, Li, Daw, & Schoenbaum, 2012) – an area known to mediate the impact of emotional salience on associative learning in the dorsal and ventral striatum (Cador, Robbins, & Everitt, 1989; McGaugh, 2004; McGaugh, McIntyre, & Power, 2002). This suggests that the learning mechanism of our UWL model could be implemented via the amygdala's control over the neuromodulation of synaptic plasticity. Our work on utility-weighted sampling thereby illustrates how resource-rational analysis can be used to connect the computational level of analysis to the algorithmic and the implementation level (Griffiths et al., 2015; Marr, 1982). Future work might be able to leverage insights from neuroscience to quantify the resource-constraints and cost of computation in models of rational information processing (Lieder, Goodman, & Griffiths, 2013).

3.6.4 Connection to fast-and-frugal heuristics and ecological rationality

Interestingly, our resource-rational analysis led to simple and psychologically plausible decision strategies that resemble two *fast-and-frugal* heuristics (Gigerenzer, 2008b). Biased mental simulation (stochastically) considers the most important consequence first – like *take-the-best* – and binary choices are made by tallying if there are more positive than negative simulated outcomes – as in the *tallying heuristic*. The fact that we derived this strategy as a resource-efficient approximation to normative decision-making (resource-rational analysis) sheds light on why fast-and-frugal heuristics work and how they can be generalized to harder problems (Lieder et al., 2012, cf.).

Pleskac and Hertwig (2014) point out that natural decision environments often exhibit and inverse relationships between probability and reward, such as power-law distributions. It is these reward structures for which representative sampling fails and utility-weighted sampling becomes necessary. This suggests that utility-weighted sampling is an ecologically rational heuristic, and this might be why it is so effective and predictive of people's decisions and biases. Although we derived utility-weighted sampling for complex real-life decisions with infinitely many possible outcomes, we found that it also captures the simpler two-outcome choices people make in laboratory experiments that could be solved by computing and maximizing expected value. This is consistent with the view that people's heuristics are adapted to the structure of the problems they face in real-life rather than those posed in the laboratory (Gigerenzer, 2015). This highlights the value of deriving theories from an analysis of the problems people have to solve in real life instead of building them in a bottom-up

fashion from empirical findings in artificial laboratory experiments.

Importantly, utility-weighted sampling works not despite its bias but because of it (Gigerenzer & Brighton, 2009)). The underlying principle is the bias-variance tradeoff (Hastie et al., 2009). Fast-and-frugal heuristics tolerate bias to make good inferences from incomplete, noisy observations, and utility-weighted sampling tolerates bias to make good decisions based on incomplete, noisy simulations of possible outcomes. Thus, biased minds can not only make better inferences but also better decisions. However, our results highlight a tension between good inference and good decision-making: To make good decisions bounded sample-based agents should over-sample extreme events even though this leads to bad inferences such as the overestimation of event frequencies, and people appear to do the same. In more general terms, the human mind should, and appears to, sacrifice the rationality of its beliefs (*theoretical rationality*) for the rationality of its actions (*practical rationality*,), because limited computational resources necessitate tradeoffs. Concretely, our analysis suggested that the availability bias is a manifestation of resource-rational decision-making. Being biased can be resource-rational.

3.6.5 Limitations and future work

In addition to the many phenomena that our model captures there are others that it does not capture. For instance, UWS with the parameters estimated from the Technion choice prediction competition for decisions from description does not capture the common-ratio effects observed by Starmer and Sugden (1989)(Starmer & Sugden, 1989). Consistent with the failure of UWS to capture these effects, Starmer and Sugden (1989) demonstrated that at least some common-ratio effects are partly driven by a distortion of stated probabilities that is independent of the outcome. Furthermore, UWS with the parameters estimated from the Technion choice prediction competition for decisions from description also cannot capture the violation of weak stochastic transitivity demonstrated by Tversky (1969) as this effect appears to be driven by people's limited sensitivity to small differences in outcome probability. For both experiments, UWS predicted that people would always choose the gamble with higher expected value. These discrepancies highlight that probability weighting in decisions from description is impacted not only by the extremity of the associated outcomes but also by the probabilities themselves. UWS fails to capture these effects because it cannot account for outcome-independent distortions of probability. Incorporating this distortion into the UWS model of decisions from description is a potential direction for future research.

It is important to keep in mind that our goal was not to test a specific computational mechanism

but rather to explore the implications of finite time and limited cognitive resources for decision-making under uncertainty. We explored these implications under specific simplifying assumptions about people's utility function, resources, and cognitive operations that may have to be revised in the future. The empirical data we examined supported the conclusion that the neural mechanisms of decision-making share some of the abstract properties of utility-weighted sampling, but there additional intricacies that remain to be captured. The following discrepancies between our models' predictions and human behavior could be a starting point for making the utility-weighted sampling mechanism more realistic: Although our model's predictions of the Allais paradox are qualitatively correct, the predicted effect was much smaller than the one observed by Kahneman and Tversky (1979). Furthermore, despite its large number of parameters, the utility-weighted learning model does not fully capture the experimental data of Ludvig et al. (2014); in particular, our model could not predict that participants in Experiment 3 were more risk seeking for non-extreme loss trials than for non-extreme gain trials. Another avenue towards identifying the computational mechanism that underly availability biases could be to investigate their neural implementation. Although the utility-weighted learning model is inspired by neuroscientific findings, our hypotheses about the neural basis of utility-weighted learning remain to be tested.

Unlike most laboratory experiments, many real-world decisions involve many possible alternatives. This makes extending UWS to multi-alternative decisions an important direction for future research. One way to extend UWS to multi-alternative choice is to apply the UWS mechanism defined in Equation 3.12 to efficiently estimate the expected utility gain of each option separately and choose the alternative whose utility estimate is highest:

$$\Delta \hat{U}_{\tilde{q},s}^{\mathrm{IS}}(a) = \frac{1}{\sum_{j=1}^{s} 1/\left|\Delta u(o_j^{(a)}) - \overline{\Delta u}\right|} \cdot \sum_{j=1}^{s} \frac{\Delta u(o_j^{(a)})}{\left|\Delta u(o_j^{(a)}) - \overline{\Delta u}\right|}, \qquad (3.38)$$

$$a^\star = \arg\max_{a \in \mathcal{A}} \hat{U}_{\tilde{q},s}^{\mathrm{IS}}(a), \qquad (3.39)$$

where $a \in \mathcal{A}$ ranges from the first to the last alternative, and $\overline{\Delta u}$ can be thought of as the average utility gain obtained in past decisions or the reward expectancy conveyed by dopamine, as discussed above. This mechanism could be very efficient for decisions from experience because it allows multiple alternatives to be evaluated in parallel. Given the resulting estimates of the expected utility gain, the brain could read out the preferred action with a winner-take-all network (Maass, 2000). Alternatively, it is conceivable that decision-makers sometimes reduce multi-alternative decisions into a series of binary choices and make those choices with the UWS heuristic for binary decisions

(Equations 3.18–3.21). Finally, it is also conceivable that decision-makers would first identify which alternatives are most promising by evaluating them separately according to Equation 3.38 and then apply the UWS heuristic for binary decisions (Equations 3.18–3.21) to choose between the two actions with the highest estimated utility gains. Future work should evaluate which of these alternative extensions best predicts people's multi-alternative decisions from experience.

Our resource-rational analysis assumed that the limited resource is the number of samples that can be generated. This assumption appears justified for memory-based decisions where sampling by memory retrieval is the primary cognitive operation. But in decisions from description other cognitive and perceptual operations, such as inspecting the probabilities, or gauging the differential utilities of pairs of outcomes also consume a non-negligible amount of time and cognitive resources. In particular, the cost of determining the differential utility of all pairs of outcomes becomes prohibitive as the number of outcomes increases. Since our analysis ignores these computational costs, the applicability of our original model of decisions from descriptions is limited to choices with a small number of possible outcomes. However, this limitation does not apply to our model of decisions from experience, and a recent resource-rational analysis of multi-alternative, multi-outcome decisions from description captured important aspects of people's adaptive decision strategies in the Mouselab paradigm (Lieder, Krueger, & Griffiths, 2017).

While the simulation and integration mechanisms of UWS were derived from first principles, the choice of the utility function in Equation 3.16 was less principled. We chose it because it is the simplest instantiation of the efficient coding theory proposed by Summerfield and Tsetsos (2015) that captures our findings. It thus remains to be validated independently. Consistent with this normalized representation of utility, there is neural evidence that the human brain encodes relative value rather than absolute value (Mullett & Tunney, 2013). Yet, this evidence equally consistent with a rank-dependent utility function. Neurophysiological data from animal studies (Louie et al., 2011) and psychophysical data from humans (Louie et al., 2013) speak to the encoding of normalized value, but further research is needed to determine the exact nature of the brain's relative utility representation and its variability.

While we focused on one particular strategy for mitigating resource constraints, namely adjusting the simulation distribution, the brain also appears to adjust the number of samples. Our own and other recent findings suggest that people draw more samples when the stakes are high (Vul et al., 2014) and when they are very uncertain (Hamrick, Smith, Griffiths, & Vul, 2015). The models presented here capture neither of these effects, but future versions of UWS will accommodate them according to the principle that people make rational use of their finite cognitive resources (Griffiths

et al., 2015). Recent work has developed a mechanism for determining the optimal number of samples (Tajima et al., 2016), and future work should integrate this mechanism into UWS.

Testing whether the magnitude of the simulated availability biases is resource-rational will additionally require independent measurements of people's cognitive resources. Therefore, measuring resource constraints independently and using these measurements to derive and test quantitative predictions of human performance as a function of incentives and time pressure is an important direction for future research. A first step towards deriving these predictions could be to measure how long it takes to generate a single sample using psychophysical methods (Lengyel, Koblinger, Popović, & Fiser, 2015). It might also be possible to measure how long it takes to generate a sample by investigating the relationship between the time available to make a choice and the resulting choice variability. Alternatively, a lower bound on how long it takes to generate a sample could be derived from spiking neural network models of how the brain generates samples (Buesing et al., 2011). This bound on how fast samples can be generated could then be translated into an upper bound on how much the availability biases simulated here can be reduced by financial incentives. The estimated time per sample could also be used to derive the cost of sampling in scenarios where people have to trade off how much computation to invest in a decision against the number of choices they can make (Vul et al., 2014). The resulting model of the cost of sampling could inform a rational mechanism for choosing the number of samples (Hay et al., 2012; Tajima et al., 2016; Vul et al., 2014) to be generated by utility-weighted sampling. Future experiments should also test the assumption that the number of mental simulations is a critical limiting factor to the quality of people's decisions. This assumption predicts that time pressure and cognitive load should make people's risk preferences more inconsistent between gains versus losses. Conversely, instructing or incentivizing participants to simulate their decision more often should reduce the impact of extreme events.

Another avenue for future research is to investigate whether people use utility-weighted sampling adaptively. Three mechanisms of adaptivity are conceivable: First, people might adapt the number of simulations to the decision problem's incentives for speed and accuracy. Second, people might use their current estimate of the expected utility gain to adapt their simulation distribution from one simulation to the next as in adaptive importance sampling (Oh & Berger, 1992):

$$\bar{u}_s = \hat{U}_{q,s-1}^{\mathrm{IS}} \tag{3.40}$$

Third, people might use utility-weighted sampling selectively only for those problems in which they expect it to work well (Lieder & Griffiths, 2015, 2017).

Finally, utility-weighted sampling makes a number of novel predictions that can be tested empirically: Because the predicted availability biases increase with the extremity of the event, the *probability-weighting function* (Tversky & Kahneman, 1992) should be monotonic in the outcome's payoff relative to other outcomes. According to our UWL model, the rate at which action-outcome associations are learned is proportional to the absolute value of the reward's utility. This assumption could be tested by measuring the temporal evolution of memory biases as a function of outcome extremity in a modified version of the paradigm by Madan et al. (2014). In addition, the utility-weighted learning model predicts that whether an outcome becomes overweighted and how strongly depends on what the decision-maker expected when they experienced that outcome: A person who expected a large reward will come to overweight a neutral outcome whereas a person whose reward expectation was zero would come to underweight it. Likewise, people with a negative reward expectation should come to overweight positive outcomes much more strongly than people with a positive reward expectation and vice versa. In terms of individual differences, UWL predicts that people with lower sensitivity to rewards and punishments (Corr, 2004) should be less susceptible to develop availability biases in memory recall, frequency estimation, and decision-making than people with higher reinforcement sensitivity. Furthermore, people who are more sensitive to punishment than to reward should be more prone to develop such biases for losses than for gains, and the opposite should be true for people who are more sensitive to reward than to punishment. Perhaps the most counterintuitive prediction of UWS is that for certain decisions, such as the one illustrated in the Appendix B, where people's risk preferences should become more biased the more people think about them.

3.6.6 Conclusion

Overall, the findings presented in this chapter show that utility-weighted sampling is a promising rational process model of judgment and decision-making that provides a unifying explanation for a wide range of cognitive biases in memory recall, learning, frequency estimation, decisions from experience, and decisions from description. According to our resource-rational analysis, all of these availability biases result from the rational use of limited time and bounded cognitive resources. From this perspective, cognitive biases are a window on resource-rational information processing rather than a sign of human irrationality.

4

A rational solution to the strategy selection problem*

To succeed in life we have to solve a wide range of problems that place very different demands on us: sometimes we have to think fast and sometimes we have to think slow (cf. Kahneman, 2011). For instance, avoiding a car accident requires a split-second decision, whereas founding a successful start-up requires investing a lot of time into anticipating the future and weighting potential outcomes appropriately. No single decision mechanism works well across all situations. To meet the wide range of demands posed by different decision problems, it has been proposed that the human brain is equipped with multiple decision systems (Dolan & Dayan, 2013) and decision strategies (Payne, Bettman, & Johnson, 1988). Dual-process theories are a prominent example of this perspective (Evans, 2003; Evans & Stanovich, 2013; Kahneman, 2011). The coexistence of multiple alternative strategies is not specific to decision making. People also appear to possess multiple strategies for inference (Gigerenzer & Selten, 2002), memory (Bjorklund & Douglas, 1997), self-control (Braver, 2012), problem solving (Fum & Del Missier, 2001), and mental arithmetic (Siegler, 1999) to name just a few.

---

*This chapter is based on Lieder and Griffiths (2017).

The availability of multiple strategies that are applicable to the same problems raises the question how people decide when to use which strategy. The fact that so many different strategies have been observed under different circumstances shows that people's strategy choices are highly variable and contingent on the situation and the task (Beach & Mitchell, 1978; Fum & Del Missier, 2001; Payne, 1982; Payne et al., 1988). Overall, the contingency of people's strategy choices appears to be adaptive. Even though under certain circumstances people have been found to use heuristics that cause systematic errors (Ariely, 2009; Sutherland, 1992), their strategies are typically well-adapted to the problems to which they are applied (Anderson, 1990; Braver, 2012; Bröder, 2003; Fum & Del Missier, 2001; Payne et al., 1993). For instance, Payne and colleagues found that when the probabilities of alternative outcomes fall off quickly, then decision makers employ frugal heuristics that prioritize the most probable outcomes at the expense of less probable ones. Similarly, decision makers select fast heuristics when they are under time pressure but more accurate ones when they are not (Payne et al., 1988). These and other studies (e.g., Siegler, 1999) have also documented that people's propensity to use one strategy rather than another changes over time.

The adaptiveness of people's strategy choices appears to increase with experience. For instance, as children gain more experience with mental arithmetic they gradually learn to choose effective and efficient strategies more frequently (Siegler, 1999). In adults, adaptive changes in strategy selection have been observed on much shorter time scales. For instance, adults have been found to adapt their decision strategy to the structure of their decision environment within minutes as they repeatedly choose between different investment based on multiple attributes (Rieskamp & Otto, 2006): In a decision environment where the better investment option is determined by a single attribute people learn to use a fast-and-frugal heuristic that ignores all other attributes. But when the decision environment does not have that structure, then people learn to integrate multiple attributes.

How can we explain the variability, task- and context-dependence, and change in people's strategy choices? Despite the previous work reviewed in the following section and some recent progress on how the brain decides how to decide (Boureau, Sokol-Hessner, & Daw, 2015) the strategy selection problem remains unsolved (Marewski & Link, 2014). Finally, while it is typically assumed that people's use of heuristics is irrational (Ariely, 2009; Marcus, 2009; Sutherland, 1992), there is increasing evidence for adaptive strategy selection (Boureau et al., 2015; Braver, 2012; Daw, Niv, & Dayan, 2005; Fum & Del Missier, 2001; Gunzelmann & Anderson, 2003; Keramati, Dezfouli, & Piray, 2011; Payne et al., 1988). This raises the additional question whether and to what extent people's strategy choices are rational.

In this chapter, we formalize the strategy selection problem, derive a rational strategy selection

mechanism, and show that it can explain a wide range of empirical phenomena including the variability, contingency, and change of strategy selection across multiple domains – ranging from decision-making to arithmetic – and time scales. Our theory adds an important missing piece to the puzzle of bounded rationality by specifying when people should use which heuristic, and our findings reconcile the two poles of the debate about human rationality by suggesting that people gradually learn to make increasingly more rational use of their fallible heuristics.

The next section situates our work in the debate about human rationality and previous research on strategy selection. We then develop an alternative, rational account of strategy selection based on the idea of *rational metareasoning* from artificial intelligence research (Russell & Wefald, 1991b). The following sections evaluate our theory against traditional theories of strategy selection and show that it provides a unifying explanation for a wide range of phenomena: We show that rational metareasoning can account for people's ability to adaptively choose the sorting strategy that works best for each individual problem based on limited experience, while traditional theories of strategy selection cannot. The subsequent sections show that this conclusion holds not only for behavioral strategies but is equally true of cognitive strategies for decision-making, and mental arithmetic that operate on internal representations. We conclude with the implications of these findings for the debate about human rationality and directions for future research.

4.1 BACKGROUND

4.1.1 THE DEBATE ABOUT HUMAN RATIONALITY

The results presented in Chapters 1–3 suggested that major cognitive biases in judgment and decision-making that have been interpreted as evidence against human rationality are consistent with the rational use of finite cognitive resources. Concretely, the anchoring bias that pervades human judgment appears to be the manifestation of a resource-rational strategy for drawing inferences under uncertainty (Chapter 2) and numerous cognitive biases in people's decisions under uncertainty are accurately predicted by a resource-rational decision strategy (Chapter 3). This line of work demonstrates that fallible heuristics can be resource-rational for certain problems under some circumstances. Similarly, Gigerenzer and colleagues have found that simple, fast-and-frugal heuristics perform very well when their assumptions match the structure of the environment (Gigerenzer, 2008a, 2008b; Gigerenzer & Brighton, 2009; Gigerenzer & Selten, 2002; Gigerenzer & Todd, 1999; Todd & Gigerenzer, 2012).

Scholars who view heuristics as irrational kluges that give rise to fallacies and biases (Ariely, 2009; Marcus, 2009; Sutherland, 1992) emphasize situations in which the chosen heuristics are maladaptive, whereas researchers who interpret heuristics as rational strategies point to situations where people use them adaptively (Griffiths et al., 2015; Todd & Gigerenzer, 2012). Arguably, most heuristics are neither rational nor irrational per se. Instead, their rationality depends on how well they fit the problem to which they are being applied. Hence, the degree to which people are rational depends on when they use which heuristic. The critical question thus becomes "Are heuristics chosen rationally?" In this chapter, we address this question by developing and testing a rational model of strategy selection.

4.1.2 Previous theories of strategy selection

Strategy selection was initially viewed as a metacognitive decision based on explicit metacognitive knowledge about which cognitive strategies are best suited for which purposes (Flavell, 1979). Consistent with this perspective, Beach and Mitchell (1978) proposed that people choose decision strategies by performing an explicit cost-benefit analysis. Although Beach and Mitchell (1978) did not formalize this process enough to make quantitative predictions, their qualitative predictions were later confirmed in the domain of decision-making (Payne et al., 1988). Payne and colleagues demonstrated that which decision process performs best is contingent on time pressure and the structure of the decision problem.

The participants in the experiments conducted by Payne et al. (1988) responded adaptively to task contingencies *as if* their strategy choices were based on a cost-benefit analysis. Yet, under most circumstances, performing a complete cost-benefit analysis would take substantially longer than executing the most accurate strategy. In order to be beneficial, people's strategy selection mechanism has to be efficient. Furthermore, it has to avoid the infinite regress that could potentially result from reasoning about reasoning. These considerations have led researchers to abandon the idea that strategies are selected by a metacognitive cost-benefit analysis in favor of simpler models that select strategies by learning directly from experience (Erev & Barron, 2005; Rieskamp & Otto, 2006; Shrager & Siegler, 1998; Siegler, 1988; Siegler & Jeff, 1984). Consistent with this emphasis on learning, multiple experiments have found that people's strategy choices become more adaptive with experience (Bröder, 2003; Payne et al., 1988; Rieskamp & Otto, 2006).

Previous learning-based accounts of strategy selection were based on simple associative learning (Shrager & Siegler, 1998) and learning from feedback (Erev & Barron, 2005; Rieskamp & Otto,

2006). These mechanisms can be interpreted as a form of model-free metacognitive reinforcement learning in the sense they update the decision-maker's propensity to choose a strategy directly without building a model of what will happen when the strategy is selected[†]. According to the SSL (Rieskamp & Otto, 2006) and RELACS (Erev & Barron, 2005) models (defined in detail in Appendix C.1), people solve the strategy selection problem by learning which strategy works best on average in a given environment. This learning mechanism does not exploit the fact that every problem has distinct characteristics that determine the strategies' effectiveness.

According to the SCADS model (Shrager & Siegler, 1998), people learn to associate strategies with problem types. Every time a strategy is applied to a problem the association between the problem's type and the strategy is strengthened, and this strengthening is strongest when the strategy was successful. Using the same mechanism, the SCADS model also learns a global association between each strategy and problems in general. When presented with a problem the SCADS model chooses the strategy for which the product of the problem type specific association strengths and the global association strength is highest. This learning mechanism presupposes that each problem has been identified as an instance of one or more problem types. If each problem belongs to exactly one category, then the SCADS model learns to use the same strategy for all problems of a given type, but each problem can belong to multiple categories.

In his rational analysis of problem solving (Anderson, 1990) developed a more sophisticated strategy selection mechanism according to which people probabilistically select strategies (productions) that yield a high value of $\hat{P} \cdot G - \hat{C}$ where G is the value of achieving the goal and $\hat{P}$ and $\hat{C}$ are Bayesian estimates of the success probability and the cost of achieving the goal. This mechanism has been implemented in ACT-R to simulate strategy selection learning in problem solving (Gunzelmann & Anderson, 2003). However, like the model-free reinforcement learning mechanisms of SSL and RELACS (Erev & Barron, 2005; Rieskamp & Otto, 2006) the learning mechanism of ACT-R does not exploit the fat that some problems are more similar than others.

The cognitive niche theory (Marewski & Schooler, 2011) complements theories points out that people need only choose between those strategies that are applicable in a given situation. It emphasizes that the affordances of most situations severely limit the number of applicable strategies, for instance because the information required by many strategies is unavailable or cannot be recalled.

Recent work in computational neuroscience has modeled how the brain arbitrates between the

---

[†]From a different perspective, all theories of strategy selection learning can be seen as model-based because each strategy corresponds to a certain model of the environment (Gluth, Rieskamp, & Büchel, 2013).

model-free (habitual) and the model-based (goal-directed) decision system as meta-decision-making using ideas from reinforcement learning (Boureau et al., 2015; Daw et al., 2005; Keramati et al., 2011). This approach is promising and the reinforcement-learning framework is very powerful. However, it has yet to be extended to the complexities of the more general problem of strategy selection. In the following section, we pursue this idea to provide a new rational analysis of strategy selection that overcomes the limitations of previous theories.

4.2 Strategy selection learning as metacognitive reinforcement learning

In this section we provide a computational-level theory (Marr, 1982) of the strategy selection problem and propose a learning and a selection mechanism through which people might solve this problem. The key idea is that people learn to predict the accuracy and execution time of each strategy from features of individual problems and choose the strategy with the best predicted speed-accuracy tradeoff.

4.2.1 The strategy selection problem

Each environment E can be characterized by the relative frequency P_E with which different kinds of problems occur in it. In most environments, these problems are so diverse that none of people's strategies can achieve the optimal speed-accuracy tradeoff on all of them. Optimal performance in such environments requires selecting different strategies for different types of problems. One way to achieve this would be to learn the optimal strategy for each problem separately through trial and error. This approach is unlikely to succeed in complex environments where no problem is exactly the same as any of the previous ones. Hence, in many real-world environments, learning about each problem separately would leave the agent completely unprepared for problems it has never seen before. This can be avoided by exploiting the fact that each problem has perceivable features $f_1, \ldots, f_K$ that can be used to predict the performance of candidate strategies from their performance on previous problems. For instance, the features of a risky choice may include the number of options, the spread of the outcome probabilities, and the range of possible payoffs.

How good it is to apply strategy s to problem[i] depends not only on the expected reward but also on the expected cost of executing the strategy. Building on the theory of rational metareasoning developed in artificial intelligence research (Russell & Wefald, 1991b), this can be quantified by the

165

value of computation (VOC):

$$\text{VOC}(s, \text{problem}^{(i)}) = \mathbb{E}[U(s(\text{problem}^{(i)}); \text{problem}^{(i)}) - \text{cost}(s, \text{problem}^{(i)})],$$

where $s(\text{problem}^{(i)})$ is the action the potentially stochastic strategy s selects on problem$^{(i)}$, $U(A)$ denotes the utility of taking action A, and $\text{cost}(s, \text{problem}^{(i)})$ is the computational cost of executing strategy s on that problem. In the following we will assume that the computational cost is driven primarily by the (cognitive) opportunity cost of the strategy's execution time $T(s, \text{problem}^{(i)})$, that is

$$\text{cost}(s, \text{problem}^{(i)}) = \gamma \cdot T(s, \text{problem}^{(i)}).$$

The problem of optimal strategy selection can be defined as finding a mapping $m : \mathcal{F} \mapsto \mathcal{S}$ from feature vectors ($\mathbf{f}^{(i)} = (f_1(\text{problem}^{(i)}), \dots, f_K(\text{problem}^{(i)})) \in \mathcal{F}$) to strategies ($s \in \mathcal{S}$) that maximizes the expected VOC of the selected strategy across all problems the environment might present. Hence, we can define the strategy selection problem as

$$\arg\max_{m} \sum_{\text{problem} \in \mathcal{P}} P_E(\text{problem}) \cdot \text{VOC}(m(\mathbf{f}(\text{problem})), \text{problem}),$$

where $\mathcal{P}$ is the set of problems that can occur.

Critically, the VOC of each strategy depends on the problem, but the strategy has to be chosen entirely based on the perceivable features $\mathbf{f}$ and the strategy selection mapping m has to be learned from experience. In machine learning, these kinds of problems are known as contextual multi-armed bandits (May, Korda, Lee, & Leslie, 2012). Two critical features of this class of problems are that they impose an exploration-exploitation tradeoff and require generalization. In the next section, we will leverage these insights to derive a rational strategy selection learning mechanism.

The experience gained from applying a strategy s to a problem with perceivable features $\mathbf{f}$ and observing an outcome with utility u after executing the strategy for t units of time can be summarized by the tuple $(\mathbf{f}, s, u, t)$. Hence, people's experience after the first n problems can be summarized by the history

$$h_n = ((\mathbf{f}^{(1)}, s^{(1)}, u^{(1)}, t^{(1)}), \dots, (\mathbf{f}^{(n)}, s^{(n)}, u^{(n)}, t^{(n)})),$$

where $\mathbf{f}^{(i)}, s^{(i)}, u^{(i)}, t^{(i)}$ are the feature vector of the ith problem, the strategy that was applied to it, and the resulting utility and execution time respectively. Strategy selection learning induces a sequence $m^{(1)}, m^{(2)}, \dots, m^{(N)}$ of strategy selection mappings that depends on the agent's experience

(h_n) and its strategy selection learning mechanism $l : \mathcal{H} \mapsto \mathcal{M}$ where $\mathcal{H}$ is the set of possible histories and $\mathcal{M}$ is the set of possible strategy selection mechanisms. With this notation, we can express the agent's performance on the nth problem by

$$\text{VOC}(m^{(n)}(\mathbf{f}^{(n)}), \text{problem}^{(n)}),$$

where $m^{(n)}(\mathbf{f}^{(n)})$ is the strategy the agent selects for the nth problem, and the strategy selection mapping $m^{(n)}$ is $l(h^{(n-1)})$. Since the problem is sampled at random, the expected performance at time step n is

$$V_n(l) = \mathbb{E}_{P_E}[\text{VOC}(m^{(n)}(\mathbf{f}^{(n)}), \text{problem}^{(n)}) \mid m^{(n)} = l(h^{(n-1)})].$$

If the agent solves N problems before it runs out of time, its total performance is

$$V_{total}(l) = \mathbb{E}\left[\sum_{n=1}^{N} V_n(l)\right].$$

Using this notation, we can define the optimal strategy selection learning mechanism l^* as the one that maximizes the agent's total expected value of computation across all possible sequences of problems, that is

$$l^* = \arg\max_l V_{total}(l).$$

This concludes our computational-level analysis of strategy selection and strategy selection learning. We will now use this analysis as a starting point for deriving a rational strategy selection learning mechanism.

4.2.2   A RATIONAL PROCESS MODEL OF STRATEGY SELECTION

Our computational-level analysis identified that a general strategy selection learning mechanism should be able to transfer knowledge gained from solving one problem to new problems that are similar. In the reinforcement learning literature generalization is typically achieved by parametric function approximation (Sutton & Barto, 1998). The simplest version of this approach is to learn the coefficients of a linear function predicting the value of a state from its features. Such linear approximations require minimal effort to evaluate and can be learned very efficiently. We therefore propose that people learn an internal predictive model that approximates the value of applying a

strategy s to a problem by a weighted average of the problem's features $f_1(\text{problem}), \ldots, f_n(\text{problem})$:

$$\text{VOC}(s, \text{problem}) \approx \sum_{k=1}^{n} w_{k,s} \cdot f_k(\text{problem}). \tag{4.1}$$

This approximation is easy to evaluate, but it is not clear how it can be learned given that the VOC cannot be observed directly. However, when the strategy s generates a decision, then the VOC can be decomposed into the utility of the decision's outcome and the cost of executing the strategy. Assuming that the cost of executing the strategy is proportional to its execution time, the VOC can be approximated by

$$\text{VOC}(s, \text{problem}) \approx \mathbb{E}[U \mid \text{problem}, s] - \gamma \cdot \mathbb{E}[T \mid \text{problem}, s], \tag{4.2}$$

where U is the utility of the outcome obtained by following strategy s, γ is the agent's opportunity cost per unit time and $\mathbb{E}[T \mid \text{problem}, s]$ is the expected execution time of the strategy s when applied to the problem.

Approximating the VOC thus becomes a matter of estimating three quantities: the expected utility of relying on the strategy, the opportunity cost per unit time, and the expected time required to execute the strategy. The agent can learn its opportunity cost γ by estimating its reward rate (Boureau et al., 2015; Niv, Daw, Joel, & Dayan, 2007), and the utility of applying the strategy and its execution time T can be observed. Therefore, when the reward is continuous, then it is possible to learn an efficient approximation to the VOC by learning linear predictive models of the utility of its decisions and its execution time and combining them according to

$$\text{VOC}(s, \text{problem}) \approx \sum_{k=1}^{n} w_{k,s}^{(R)} \cdot f_k(\text{problem}) - \hat{\gamma} \cdot \sum_{k=1}^{n} w_{k,s}^{(T)} \cdot f_k(\text{problem}).$$

This equation is a special case of the general approach specified in Equation 4.1. When the outcome is binary, then the predictive model of the reward takes the form

$$P(O = 1 \mid s, \text{problem}) = \frac{1}{1 + \exp\left(-\sum_{k=1}^{n} w_{k,s}^{(R)} \cdot f_k(\text{problem})\right)}.$$

We model the agent's estimate of its opportunity cost γ by the posterior mean $\mathbb{E}[\bar{r} \mid t_{total}, r_{total}]$ of its reward rate $\bar{r}$ given the sum of rewards r_{total} that the agent has experienced and the time since the beginning of the experiment (t_{total}). To do so, we assume that both the prior and the likelihood

function are Gaussian, that is

$$P\left(r_{total}/t_{total} \mid \bar{r}\right) = \mathcal{N}(\mu = \bar{r}, \tau = t_{total} \cdot 1/60s),$$
$$P(\bar{r}) = \mathcal{N}(1, 1).$$

In this model, the weight of the agent's experience increases linearly with its time spent in the environment, and the prior corresponds to 60 sec worth of experience.

Our theory covers learning and strategy selection. To simulate learning, the agent's belief about the reward rate and the feature weights in the predictive model of a strategy's accuracy and execution time are updated by Bayesian learning every time it has been executed: The belief about the reward rate $\bar{r}$ is updated to $P(\bar{r} \mid r_{total}, t_{total})$ as described in Section C.1.2 of Appendix C.1. The weights of the execution time model are updated by Bayesian linear regression (see Section C.1.3 of Appendix C.1). The weights of the reward model are updated by Bayesian logistic regression (see Section C.1.4 of Appendix C.1) if the reward is binary (i.e., correct vs. incorrect), or by Bayesian linear regression (see Section C.1.3 of Appendix C.1) when the reward is continuous (e.g., monetary). Lastly, our model learns which features are relevant for predicting the most recent strategy's execution time and reward by performing Bayesian model selection as described in Section C.1.5 of Appendix C.1.

The second component of our model is strategy selection. Given the learned predictive models of execution time and reward, the agent could predict the expected VOC of each available strategy and select the strategy with the highest expected VOC. While this approach works well when the agent has already learned a good approximation to the VOC of each strategy, it ignores the value of learning about strategies whose performance is still uncertain. Hence, always using the strategy that appears best could prevent the agent from discovering that other strategies work even better. Yet, on average, strategies that appear sub-optimal will choose worse actions than the strategy that appears best. This problem recapitulates the well-known exploration-exploitation dilemma in reinforcement learning. To solve this problem our model selects strategies by Thompson sampling (May et al., 2012): For each strategy s, our model samples estimates $\tilde{w} = (\tilde{w}_{k,s}^{(T)}, \tilde{w}_{k,s}^{(R)})$ of the weights $w = (w_{k,s}^{(T)}, w_{k,s}^{(R)})$ of the predictive models of execution time and reward from their respective posterior distributions, that is

$$\tilde{w}_{k,s}^{(T)} \sim P(w_{k,s}^{(T)} \mid h_{t-1,s}),$$
$$\tilde{w}_{k,s}^{(R)} \sim P(w_{k,s}^{(R)} \mid h_{t-1,s}),$$

where $h_{t-1,s}$ is the agent's past experience with strategy s at the beginning of trial t. From these weights $\tilde{w}$, our model predicts the VOC values of all strategies s by

$$\hat{V}_t(s, \text{problem}) = \sum_{k=1}^{n} \tilde{w}_{k,s}^{(R)} \cdot f_k(\text{problem}) - \mathbb{E}[\hat{\gamma} \mid h_t] \cdot \sum_{k=1}^{n} \tilde{w}_{k,s}^{(T)} \cdot f_k(\text{problem}),$$

where $\mathbb{E}[\hat{\gamma} \mid h_t]$ is the posterior expectation of the agent's reward rate given its past experience. Finally, our model selects the strategy s_t^* with the highest predicted VOC,

$$s_t^* = \arg\max_s \hat{V}_t(s, \text{problem}).$$

This concludes the description of our model.

Our proposal is similar to model-based reinforcement learning (Dolan & Dayan, 2013; Gläscher, Daw, Dayan, & O'Doherty, 2010) in that it learns a predictive model. However, both the predictors and the predicted variables are different. While model-based reinforcement learning aims to predict the next state and reward from the agent's action (e.g., "Go left!"), our model learns to predict the costs and benefits of the agent's deliberation from the agent's cognitive strategy (e.g., planning four steps ahead vs. planning only one step ahead). While model-based reinforcement learning is about the control of behavior, our model is about the control of mental activities that may have no direct effect on behavior. In brief, the main difference is that we have modeled metacognitive learning instead of stimulus-response learning. Despite this difference in semantics, the proposed learning mechanism is structurally similar to the semi-gradient SARSA algorithm from the reinforcement learning literature (Sutton & Barto, 1998).

As illustrated in Figure 4.1, our model's prediction mechanism could be approximated by a simple feed-forward neural network: The first layer represents the input to the strategy selection network. The subsequent hidden layers extract features that are predictive of the strategy's execution time and accuracy. The second last layer computes a linear combination of those features to predict the execution time and external reward of applying the strategy, and the final layer combines these predictions into an estimate of the VOC of applying the strategy in the current state. The weights of this network could be learned by a basic error-driven learning mechanism, and the features might emerge from applying the same error-driven learning mechanism to connections between the hidden layers (Mnih et al., 2015). With one such network per strategy a simple winner-take-all network (Maass, 2000) could read out the strategy with the highest VOC. This neural network formulation suggests that a single forward pass through a small number of layers may be sufficient to compute each strat-

egy's VOC. The action potentials and synaptic transmission required to propagate neural activity from one layer to the next happens in milliseconds. The winner-take-all mechanism for reading out the strategy with the highest VOC can be performed in less than one tenth of a second (Oster, Douglas, & Liu, 2009). Hence, the brain might be able to execute the proposed strategy selection mechanism within fractions of a second.

4.2.3 Summary

We have derived a rational process model of strategy selection as an efficient approximation to the optimal solution prescribed by rational metareasoning. In contrast to previous accounts of strategy selection, our model postulates a more sophisticated, feature-based representation of the problem to be solved and a learning mechanism that achieves generalization. Instead of just learning about the reward that each strategy obtains *on average* our model learns to predict each strategy's execution time and expected reward on each individual problem. Hence, while previous models learned which strategy works best on average, our model learns to predict which strategy is best for each individual problem. Whereas previous theories of strategy selection (Erev & Barron, 2005; Rieskamp & Otto, 2006; Siegler, 1988; Siegler & Jeff, 1984; Siegler & Shipley, 1995) rejected the ideal of a cost-benefit analysis as intractable, we propose that people learn to approximate it efficiently. Note, however, that the consideration of the cost of thinking (Shugan, 1980) is not the distinguishing feature of our model because costs can be incorporated into the reward functions of previous theories of strategy selection. Rather, the main innovation of our theory is that strategies are chosen based on the features of the problem to be solved. The remainder of this chapter shows that this allows our model to capture aspects of human cognition that were left unexplained by previous theories.

4.3 Experiment 1: Evaluating the model with sorting strategies

To test whether our rational model of strategy selection leaning can better account for how people's strategy choices change with experience than traditional context-free accounts, like RELACS, SSL and SCADS, we designed an experiment in which feature-based versus context-free strategy selection learning make qualitatively different predictions[‡]. To differentiate between these two accounts we chose a domain in which the performance of alternative strategies is well understood and differs dramatically depending on easily detectable features of the problem. Furthermore, we were looking

---

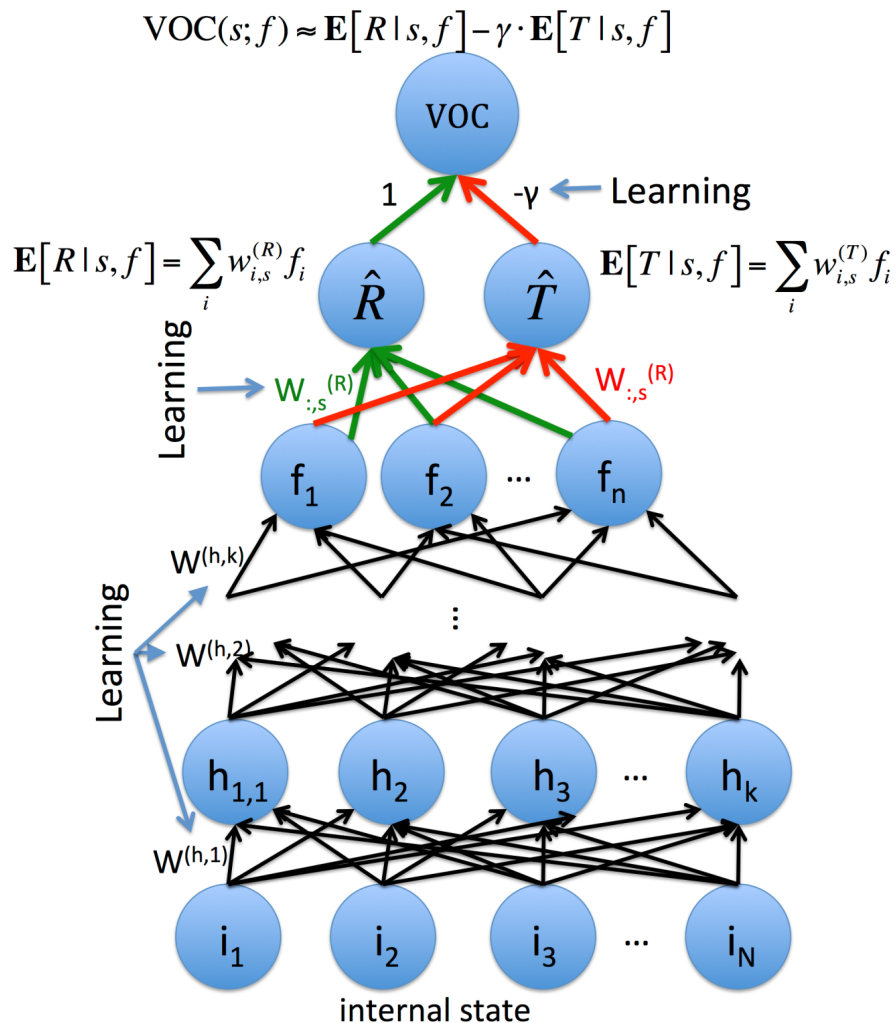[‡]A preliminary version of this study appeared in Lieder, Plunkett, et al. (2014).

Figure 4.1: Our rational process model of strategy selection learning could be implemented in a simple feed-forward neural network

for a domain in which we could directly observe people's strategy choices. These considerations led us to study how people learn to choose between two alternative strategies for sorting a list of numbers: *cocktail sort* and *merge sort* (Knuth, 1998). We chose these two sorting strategies because they have opposite strengths and weaknesses. Cocktail sort is very fast for short and nearly-sorted lists, but in the worst case its runtime increases quadratically with the length of the list ($O(n^2)$). Thus. for long, unsorted, or reversely sorted lists cocktail sort is extremely inefficient. By contrast, the execution time of merge sort does not depend on the degree to which the list is correctly or reversely

sorted and its execution time increases only log-linearly with the length of the list ($O(n \cdot \log(n))$). In the following we will assume that the task is to sort a list of numbers in ascending order.

To apply our theory to model how people learn to select between these two sorting strategies, we have to specify the features by which sorting problems are represented. For simplicity, we assume that the basic features are the length $\mid L \mid$ of the list $L = (e_1, e_2, \ldots, e_{|L|})$ and a measure of its presortedness:

$$f_1 = \mid L \mid,$$
$$f_2 = |\{m : e_m > e_{m+1}\}|,$$

where $\mid A \mid$ denotes the number of elements in the set or list A. The second feature estimates the number of pairs of elements that would have to be swapped in order to sort the list in ascending order from the number of times one element is larger than the next. Since it is well known that the runtimes of sorting algorithms are polynomials in the length of the list and its logarithm, we assume that the feature vector $\mathbf{f}$ includes all terms of the form

$$f_1^{k_1} \cdot \log(f_1)^{k_2} \cdot f_2^{k_3} \cdot \log(f_2)^{k_4},$$

where $k_1, k_2, k_3, k_4 \in \{0, 1, 2\}$ and $\sum_i k_i \leq 2$. As described above, our model will select a subset of these features and use them to predict the execution time and success probability of each sorting strategy.

4.3.1   Pilot studies and simulations

To ensure that our experiment would be able to discriminate between rational metareasoning, SSL, RELACS, and SCADS, we simulated a number of candidate experiments. These simulations required a model of each strategy's performance. To obtain this execution time model, we conducted two pilot experiments in which we measured the execution time characteristics of cocktail sort (Pilot Experiment 1) and merge sort (Pilot Experiment 2) on different types of lists. The results of these pilot experiments will also allow us to determine when each strategy should be used to achieve optimal performance.

We recruited 200 participants on Amazon Mechanical Turk: 100 for each pilot experiment. Each participant was paid 75 cents for about 15 minutes of work. In Pilot Experiment 1 participants were

required to follow the step-by-step instructions of the cocktail sort strategy (see Figure 4.2a). In Pilot Experiment 2 participants were required to follow the step-by-step instructions of the merge sort strategy (see Figure 4.2b). Participants were given detailed written instructions that precisely specified the strategy they had to execute. Furthermore, at each step the interface allowed only the correct next move of the required strategy and participants received feedback when they attempted an incorrect move. After completing four practice trials, participants were randomly assigned to sort a series of lists of varying lengths and presortedness. The lists were randomly generated so that each list was equally likely to be nearly sorted (1-20% inversions), unsorted (21-80% inversions), or nearly inversely sorted (81-100% inversions). Each list was equally likely to be very short (3-8 elements), short (9-16 elements), long (17-32 elements), or very long (33-56 elements). These lists were distributed across participants such that the total anticipated sorting time was between 10 and 20 minutes.

Figure 4.2: Interfaces used in Experiment 1 to train participants to perform (a) cocktail sort and (b) merge sort.

We used the measured sorting times to estimate how long comparisons and moves take for each strategy. For each list, we regressed the sorting times of each strategy on the number of comparisons and the number of moves that it performed on that list. The resulting model for the execution time

T_{CS} of cocktail sort (CS) was

$$T_{CS} = \hat{t}_{CS} + \epsilon_{CS},$$
$$\hat{t}_{CS} = 19.59 + 0.19 \cdot n_{\text{comparisons}} + 0.31 \cdot n_{\text{moves}}, \tag{4.3}$$
$$\epsilon_{CS} \sim \mathcal{N}(0, 0.21 \cdot \hat{t}_{CS}^2),$$

where $\hat{t}_{CS}$ is the expected execution time, ϵ_{CS} is the noise, $n_{comparisons}$ is the number of comparisons and n_{moves} is the number of moves. For merge sort (MS) our data showed that both comparisons and moves took substantially longer:

$$T_{MS} = \hat{t}_{MS} + \epsilon_{MS},$$
$$\hat{t}_{MS} = 13.98 + 1.10 \cdot n_{\text{comparisons}} + 0.52 \cdot n_{\text{moves}}, \tag{4.4}$$
$$\epsilon_{MS} \sim \mathcal{N}(0, 0.15 \cdot \hat{t}_{MS}^2).$$

According to these execution time models (Equations 3-4) and the number of comparisons and moves required by these sorting strategies, people should choose merge sort for long and nearly inversely sorted lists and cocktail sort for lists that are either nearly-sorted or short. We will therefore classify people's strategy choices as adaptive when they conform to these rules and as non-adaptive when they violate them.

The execution time models of the two strategies also allowed us to simulate 104 candidate experiments according to rational metareasoning, SSL, RELACS, and SCADS. To apply SSL, RELACS, and SCADS to sorting strategies, we had to specify the reward function. We evaluated three notions of reward: i) correctness ($r \in \{-0.1, +0.1\}$[§], ii) correctness minus time cost ($r - \gamma \cdot t$, where t is the execution time and $\gamma = 1$ is the opportunity cost), and iii) reward rate (r/t). Each of the three theories (SSL, RELACS, and SCADS) was combined with each of these three notions of reward leading to 9 alternative models in total. Since the SCADS model presupposes that each problem is characterized by a collection of binary features we designed the following categories: short lists (length ≤ 16), long lists (length ≥ 32), nearly sorted lists (less than 10% inversions), and random lists (more than 25% inversions). According to the SCADS model, each problem can belong to multiple categories or none at all. To obtain an upper bound on how well the SCADS model can select sorting strategies, we also considered three SCADS models with categories that were optimized for this experiment. These categories were short-and-presorted, long-and-presorted, short-and-inverted,

---

[§]These specific values were taken from the SCADS model (Shrager & Siegler, 1998).

long-and-inverted, short-and-inverted, long-and-disordered, and short-and-disordered. Each of these categories is the conjunction of one attribute based on length (short means $\leq$ 25 and long means $>$ 25) and one attribute based on presortedness (presorted means less than 25% inversions, inverted means more than 75% inversions, and disordered means 25–75% inversions). All associations between strategies and categories were initialized with a strength equivalent to one successful application, and the global strategy-reward associations were initialized in the same way. For consistency, the time cost parameter γ of the rational metareasoning model was also set to 1.¶

Our simulations identified several candidate experiments for which rational metareasoning made qualitatively different predictions than SSL, RELACS, and SCADS. We selected the experimental design shown in Table 4.1 because it achieved the best tradeoff between discriminability and duration. For this experimental design, rational metareasoning predicted that the choices of more than 70% of our participants would demonstrate adaptive strategy selection, whereas the SSL, RELACS, and SCADS models all predicted that people would consistently fail to select their sorting strategy adaptively (see Figure 4.4).

4.3.2 Method

We recruited 100 participants on Amazon Mechanical Turk. Each participant was paid $1.25 for about 20 minutes of work. The experiment comprised three blocks: the training block, the choice block, and the execution block.

In the *training block*, each participant was taught the cocktail sort strategy and the merge sort strategy. In each of the 11 training trials summarized in Table 4.1 participants were instructed which strategy to use. The interface shown in Figure 4.2 enforced that each of its step was executed correctly. Participants first practiced cocktail sort for five trials. Next, they practiced merge sort for four trials. These practice trials comprised nearly-reversely sorted lists of lengths 4, 8, and 16 and nearly-sorted lists of length 16 and 32. The nearly-sorted lists were created from ascending lists by inserting a randomly selected element at a random location. Nearly inversely sorted lists were created by applying the same procedure to a descending list. Finally, the last two trials contrasted the two strategies on two long, nearly-sorted lists (see Table 4.1).

¶The precise weighting of time cost versus error cost was irrelevant in these simulations because each sorting strategy was guaranteed to always generate a correct solution. Thus, there was no need to simulate how people estimate the time cost from experience.

Table 4.1: Design of Experiment 1

Training Block				Choice Block		
Trial Nr.	Strategy	Sequence Length	Inversions	Trial Nr.	Sequence Length	Inversions
1	Cocktail Sort	4	3	1	64	63
2	Cocktail Sort	8	7	2	61	60
3	Cocktail Sort	16	15	3	58	57
4	Cocktail Sort	16	1	4	55	54
5	Cocktail Sort	32	31	5	52	51
6	Merge Sort	4	3	6	49	48
7	Merge Sort	8	7	7	64	1
8	Merge Sort	16	15	8	61	1
9	Merge Sort	16	15	9	58	1
10	Cocktail Sort	32	1	10	55	1
11	Merge Sort	32	1	11	52	1
				12	49	1
				13	24	1
				14	21	1
				15	18	1
				16	15	1
				17	12	1

In the *choice block*, participants were shown 18 test lists and asked to choose which strategy (cocktail sort or merge sort) they would use if they had to sort it. To incentivize participants to choose the more efficient strategy, the instructions announced that in the following block one of their choices would be selected at random and they would have to execute it. The 18 test lists comprised six examples of each of three types of lists: long and nearly inversely sorted, long and nearly-sorted, and short and nearly-sorted (see Table 4.1). The order of these lists was randomized across participants.

In the *execution block*, one of the 12 short lists from the choice block was selected at random, and the participant had to sort it using the strategy they had selected for it.

4.3.3 Results

Our participants completed the experiment in 24.7 ± 6.7 minutes (mean $\pm$ standard deviation). In the training phase, the median number of errors per list was 2.45, and 95% of our participants made between 0.73 and 12.55 errors per list. The most important outcome was the relative frequency of adaptive strategy choices: On average, our participants chose merge sort for 4.9 of the 6 long and nearly inversely sorted lists for which it was optimal, that is 81.67% of the time. To quantify our uncertainty about this and subsequent frequency estimates we computed credible intervals based on a uniform prior (Edwards, Lindman, & Savage, 1963). According to this analysis, we can be 95% confident that the frequency with which people used merge sort on long nearly inversely sorted lists lies between 77.8% and 93.0%. By contrast, our participants chose merge sort for only 1.79 of the 6 *nearly-sorted* long lists for which it was inferior to cocktail sort (29.83% of the time, 95% credible interval: $[12.9\%, 32.4\%]$), and for only 1.62 of the 6 nearly-sorted short lists for which it was also inferior (27.00% of the time, 95% credible interval: $[16.7\%, 40.4\%]$); see Figure 4.3A. Thus, our participants chose merge sort significantly more often than cocktail sort when it was superior ($p < 10^{-10}$; Cohen's $w = 6.12$). But, when merge sort was inferior, they chose it significantly less often than cocktail sort ($p < 10^{-7}$, Cohen's $w = 6.33$). Overall, 83% of our participants chose merge sort more often when it was the superior strategy than when cocktail sort was the superior strategy and vice versa (95% credible interval: $[74.9\%; 89.4\%]$; see Figure 4.3). The high frequency of this adaptive strategy choice pattern provides strong evidence for the hypothesis that people's strategy choices are informed by features of the problem to be solved, because it would be extremely unlikely otherwise ($p < 10^{-11}$, Cohen's $w = 6.60$). This finding was predicted by our rational metareasoning model of strategy selection which achieved adaptive strategy selection in 70.5% of the simulations ($p < 10^{-14}$) and the SCADS model with optimized categories and the VOC-based reward function (performance minus time cost) which achieved adaptive strategy selection in 59.0% of the simulations ($p < 10^{-5}$) but not by any of the other SCADS, RELACS, and SSL models (all $p \geq 0.17$). Figure 4.3A compares the proportion of participants who chose their sorting strategy adaptively with the models' predictions. The non-overlapping credible intervals suggest that we can be at least 95% confident that people's strategy choices were more adaptive than predicted by SSL, RELACS, or SCADS and a series of t-tests confirmed this interpretation (all $p < 0.001$). While the frequency of adaptive strategy choices predicted by rational metareasoning ($70.5 \pm 3.2\%$) was also significantly higher than for any of the alternative models (all $p < 0.01$), our participants chose their strategies even more adaptively than our rational metareasoning model predicted (83.0% vs. 70.5%, $t(298) = 2.34, p = 0.01$). Like people, rational metareasoning se-

lected merge sort for significantly more than half of the lists that were long and inverted ($p < 10^{-6}$) but for significantly less than half of the lists that were long and presorted ($p < 10^{-15}$) or short and presorted ($p < 10^{-15}$). As shown in Figure 4.3B, none of the alternative models captured this pattern.

Our model has four components: i) choosing strategies based on their VOC by trading off expected performance versus expected cost, ii) learning to predict the performance of strategies from features of individual problems, iii) learning separate predictive models of computational effort and reward, and iv) meta-cognitive exploration by Thompson sampling. To determine which components of our model are critical to its ability to choose strategies adaptively, we created additional models by removing each of the four components in turn. This resulted in five additional models: one rational metareasoning model without features, one rational metareasoning model without exploration, two models that choose strategies based on criteria other than the VOC, and one model that approximated the VOC directly without learning to predict execution time and reward separately. The last three models use the same reward functions as the three instantiations of each of the previous theories of strategy selection learning: reward only, reward rate, and reward minus time cost; while the first two models choose strategies based on a criterion other than the VOC, the last model learns a model-free approximation to the VOC without learning to predict either deliberation time or accuracy.

To evaluate these "lesioned" models, we simulated the sorting experiment according to each of them and measured how often the resulting strategy choices were adaptive (see Figure 4.8 in the Appendix C). We found that the features and the VOC-based strategy selection mechanism were necessary to capture human performance. Exploration and learning separate predictive models for execution time and accuracy were not necessary to capture human performance in the sorting task, but they were necessary to capture human performance in the experiments simulated below; detailed statistical analyses are provided in the Appendix C.

4.3.4 DISCUSSION

We evaluated rational metareasoning against three existing theories of human strategy selection. We found that the predictions of rational metareasoning were qualitatively correct and that its choices came close to human performance. By contrast, the nine alternative models instantiating previous theories completely failed to predict people's adaptive strategy choices in our experiment: The RELACS and SSL models do not represent problem features and thus cannot account for people's
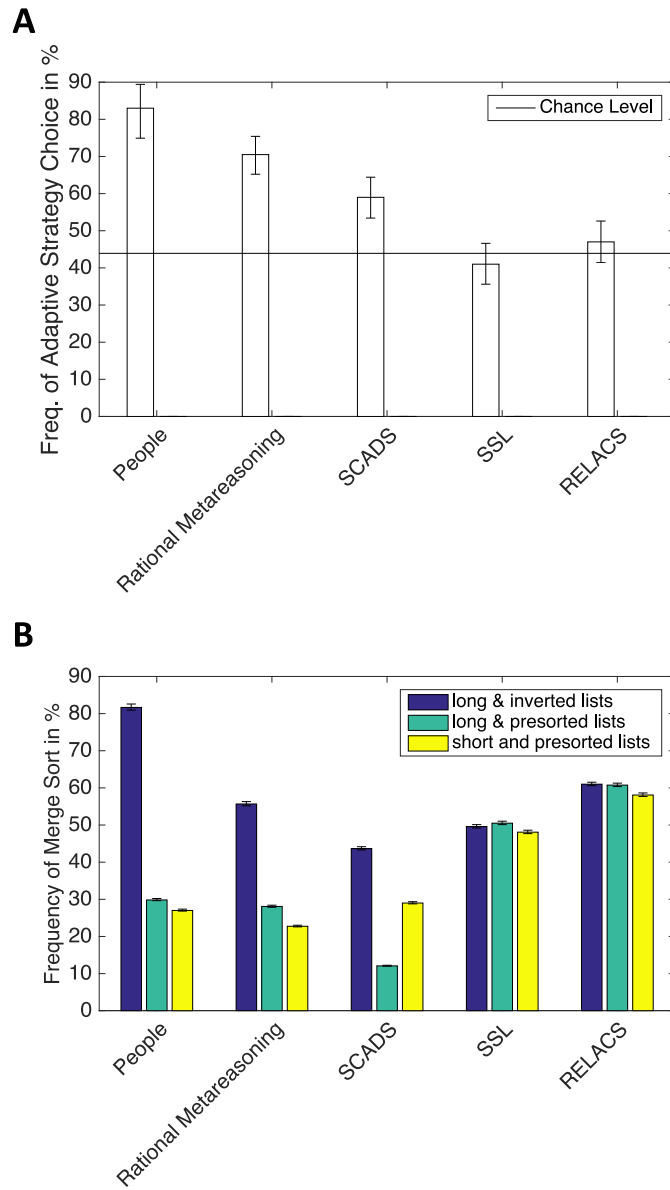
Figure 4.3: Pattern of strategy choices in Experiment 1. A: Percentage of participants who chose merge sort more often when it was superior than when it was not. Error bars indicate 95% credible intervals. The results for SCADS, SSL, and RELACS correspond to the version of the respective model that achieved the highest frequency of adaptive strategy selection. B: Relative frequency with which humans and models chose merge sort by list type.

ability to learn how those features affect each strategy's performance. The basic associative learning mechanism assumed by SSL and RELACS was maladaptive in Experiment 1 because cocktail sort was faster for most training lists but substantially slower for the long, nearly inversely sorted test lists.

The primary advantage allowing our model to perform better than SSL and RELACS is that it leverages problem features that distinguish the lists for which cocktail sort is superior from the lists for which merge sort is superior. If SSL and RELACS were applied two either set of lists separately, they would learn to identify the correct strategy for each of them. However, in the real world, problems rarely come with a single label that identifies the correct strategy. Instead, the correct strategy has to be inferred from a combination of multiple features (e.g., length and presortedness) none of which is sufficient to choose correct strategy on its own. Our rational metareasoning model handles this challenge but SSL and RELACS do not address it yet.

The SCADS model failed mainly because its associative learning mechanism was fooled by the imbalance between the training examples for cocktail sort and merge sort. Furthermore, the strategy selection component of the SCADS model can neither extrapolate nor capture the non-additive interaction between length and presortedness.

Our findings suggest that people leverage the features of individual problems to learn how to select strategies adaptively. The success of the rational metareasoning model and its evaluation against lesioned metareasoning models suggests that our hypothesis that people learn to predict the VOC of alternative strategies from the features of individual problems may be able to account for the adaptive flexibility of human strategy selection.

In contrast to the sorting strategies in Experiment 1, most cognitive strategies operate on internal representations. In principle, strategies that operate on internal representations could be selected by a different mechanism than strategies that operate on external representations. However, there are two reasons to expect our conclusions to transfer: First, people routinely apply strategies that they have applied to external objects to their internal representations of those objects. For instance, mental arithmetic is based on calculating with fingers. Thus, the strategies people use to order things mentally might also be based on the strategies they use to sort physical objects. Second, strategy selection can be seen as an instance of metacognitive control, and metacognitive processes tend to be domain general. In the following sections, we show that our conclusions do indeed transfer to cognitive strategies that operate on internal representations.

4.4 Cognitive flexibility in complex decision environments

People are known to use a wide repertoire of different heuristics to make decisions under risk (Payne et al., 1993). These strategies include fast-and-frugal heuristics which, as the name suggests, perform very few computations and use only a small subset of the available information (Gigerenzer & Gaissmaier, 2011). For instance, the lexicographic heuristic (LEX) focuses exclusively on the most probable outcome that distinguishes between the available options and ignores all other possible outcomes. Another fast-and-frugal heuristic that people might sometimes use is Elimination-By-Aspects (Tversky, 1972, EBA;). Here, we used the deterministic version of EBA described by (Payne et al., 1988). This heuristic starts by eliminating options whose payoff for the most probable outcome falls below a certain threshold. If more than one option remains, EBA repeats the elimination process with the second most probable outcome. This process repeats until only one option remains or all outcomes have been processed. After the elimination step EBA chooses one of the remaining outcomes at random. In addition to fast-and-frugal heuristics, people's repertoire also includes more time consuming but potentially more accurate strategies such as the weighted-additive strategy (WADD). WADD first computes each option's expected value, and then chooses the option whose expected value is highest.

In addition to gradually adapting their strategy choices to the structure of the environment (Rieskamp & Otto, 2006) people can also flexibly switch their strategy as soon as a different problem is presented. (Payne et al., 1988) provided a compelling demonstration of this phenomenon in risky choice: Participants chose between multiple gambles described by their possible payoffs and their respective probabilities. There was a fixed set of possible outcomes that occurred with known probabilities and the gambles differed in the payoffs they assigned to these outcomes. Participants were presented with four types of decision problems that were defined by the presence or absence of a time constraint (15 seconds vs. none) and the dispersion of the outcomes' probabilities (low vs. high); high dispersion means that some outcomes are much more probable than others, whereas low dispersion means that all outcomes are almost equally likely. Ten instances of each of the four problem types were intermixed in random order. The outcomes' payoffs ranged from $0 to $9.99, and their values and probabilities were stated numerically. (Payne et al., 1988) used process tracing to infer which strategies their participants were using: The payoffs and their probabilities were revealed only when the participant clicked on the corresponding cell of the payoff matrix displayed on the screen, and all mouse clicks were recorded. This allowed Payne and colleagues to measure how often people used the fast-and-frugal heuristics (LEX and EBA) for different types of problems by

the percentage of time people spent on the options' payoffs for the most probable outcome. For the expected-value strategy WADD this proportion is only 25%, but for the fast-and-frugal heuristics LEX and EBA it can be up to 100%. The experiment revealed that people adaptively switch decision strategies in the absence of feedback: When the dispersion of outcome probabilities was high, people focused more on the most probable outcome than when all outcomes were almost equally probable. Time pressure also increased people's propensity for such selective and attribute-based processing; see Figure 4.4. Thus, participants appeared to use fast-and-frugal heuristics more frequently when they had to be fast and when all but one or two outcomes were extremely improbable. This makes sense because the fast-and-frugal heuristics LEX and EBA are fast precisely because they focus on the most predictive attributes instead of integrating all attributes.

We investigated whether rational metareasoning can account for people's adaptive flexibility in this experiment. To do so, we simulated the experiment by applying our model to the selection between the ten decision strategies considered by (Payne et al., 1988) including WADD and fast-and-frugal heuristics such as LEX and EBA. To simulate each strategy's execution time we counted how many elementary operations (Johnson & Payne, 1985) it would perform on a given problem and assumed that each of them takes one second. This allowed us to simulate the effect of the time limit on a strategy's performance by having each strategy return its current best guess when it exceeds the time limit (Payne et al., 1988). For the purpose of strategy selection learning, our model represented each decision problem by five simple and easily computed features: the number of possible outcomes, the number of options, the number of inputs per available computation, the highest outcome probability, and the difference between the highest and the lowest payoff. Our model used these features to learn a predictive model of each strategy's relative reward

$$r_{\text{rel}}(s; o) = \frac{V(s(D), o)}{\max\limits_{a} V(a, o)},$$

where $s(D)$ is the gamble that strategy s chooses in decision problem D, $V(c, o)$ is the payoff of choice c if the outcome is o, and the denominator is the highest payoff the agent could have achieved given that the outcome was o. To choose a strategy, the predicted relative reward $\hat{r}_{rel}$ is translated into the predicted absolute reward $\hat{r}$ by the transformation

$$\hat{r} = \min\{r_{min} + (r_{max} - r_{min}) \cdot \hat{r}_{rel}, r_{max}\},$$

where r_{min} and r_{max} are the smallest and the largest possible payoff of the current gamble respec-

tively. The model then integrates the predicted absolute reward and the predicted time cost into a prediction of the strategy's VOC according to Equation 4.2 and chooses the strategy with the highest VOC as usual. The priors on all feature weights of the score and execution time models were standard normal distributions. The simulation assumed that people knew their opportunity cost and did not have to learn it from experience. Rather than requiring the model to learn the time cost as outlined above, the opportunity cost was set to \$7 per hour and normalized by the maximum payoff (\$10) to make it commensurable with the normalized rewards.

To compare people's strategy choices to rational metareasoning, we performed 1000 simulations of people's strategy choices in this experiment. In each simulation, we modeled people's prior knowledge about risky choice strategies by letting our model learn from ten randomly generated instances of each of the 144 types of decision problems considered by Payne et al. (1988). We then applied rational metareasoning with the learned model of the strategies' execution time and expected reward to a simulation of Experiment 1 from Payne et al. (1988). On each simulated trial, we randomly picked one of the four instances and generated the payoffs and outcome probabilities according to the problem type: Outcome distributions with low dispersion were generated by sampling outcome probabilities independently from the standard uniform distribution and dividing them by their sum. Outcome distributions with high dispersion were generated by sampling the outcome probabilities sequentially such that the second largest probability was at most 25% of the largest one, the third largest probability was at most 25% of the second largest one, and so on. Since the participants in this experiment received no feedback, our simulation assumed no learning during the experiment.

To evaluate our theory against alternative hypotheses, we also ran 1000 simulations according to SCADS. To evaluate our theory against alternative hypotheses, we also ran 1000 simulations according to the SCADS model. We did not evaluate SSL or RELACS because these theories do not distinguish different kinds of problems and hence cannot account for the phenomena observed by Payne et al. (1988).

The SCADS model was equipped with nine categories (time pressure, no time pressure, many options (> 3), few options (≤ 3), many possible outcomes (> 3), few possible outcomes (≤ 3), high stakes (range of payoffs $\geq 50\%$ of highest payoff), low-stakes (range of payoffs $\leq 10\%$ of highest payoff), and non-compensatory (largest outcome probability > 0.5)). As before, we considered three instances of SCADS whose reward functions were either the relative payoff, the relative payoff minus the opportunity cost of the strategy's execution time, or the reward rate. The SCADS model's category-specific and global strategy-reward associations were initialized with strengths equivalent to one observation per strategy.

We found that rational metareasoning correctly predicted that time-pressure and probability dispersion increase people's propensity to use the fast-and-frugal heuristics LEX and EBA; see Figure 4.4. Time pressure increased the predicted frequency of fast, attribute-based processing by $29.69\%(t(1998) = 9.70,\ p < 10^{-15})$, and high dispersion of the outcome probabilities increased the predicted frequency of fast, attribute-based processing by $44.11\%(t(1998) = 14.41,\ p < 10^{-15})$. Furthermore, their strategy choices only change in response to reward or punishment but the experiment provided neither. The SCADS model can choose strategies adaptively in principle, but in our simulations its strategy choices were dominated by the global, problem-independent associations. Consequently, our SCADS models always converged onto a single strategy during the training phase and continued to do so in the test trials. Hence, the predicted effects of time pressure (-0.1 to 0%, all $p \geq 0.4955$) and dispersion (0% to 0.05%, all $p \geq 0.4978$) were not significantly different from zero. In conclusion, rational metareasoning can account for adaptive flexibility in decision-making under risk but SSL, RELACS, and SCADS cannot. These results suggest that rational metareasoning can capture the adaptive flexibility of people's strategy choices not only for behavioral strategies that manipulate external representations but also for cognitive strategies that operate on internal representations.

To evaluate which components of rational metareasoning are critical to capture people's adaptive decision-making, we lesioned our model by separately removing each of its four components. We found that the feature-based problem representations and exploration were critical to the model's adaptive strategy choices but learning separate models of the costs and benefits and choosing strategies based on the VOC was not; for more detail see Appendix C. Although learning about the time cost was not necessary to perform well in the experiment by Payne et al. (1988), there are other scenarios, such as the sorting experiment, where this is critical.

Having evaluated our rational theory of strategy selection learning in the domain of decision-making, we now illustrate its generality by showing that it can also capture how people learn to solve complex problems and the development of mental arithmetic skills in children.

4.5 Learning to use the right strategy in problem solving

To solve complex problems, people employ some general and many domain-specific cognitive strategies. Hence, learning when to use which problem-solving strategy may be an important component of learning how to solve challenging problems. Our models of strategy selection learning may thus
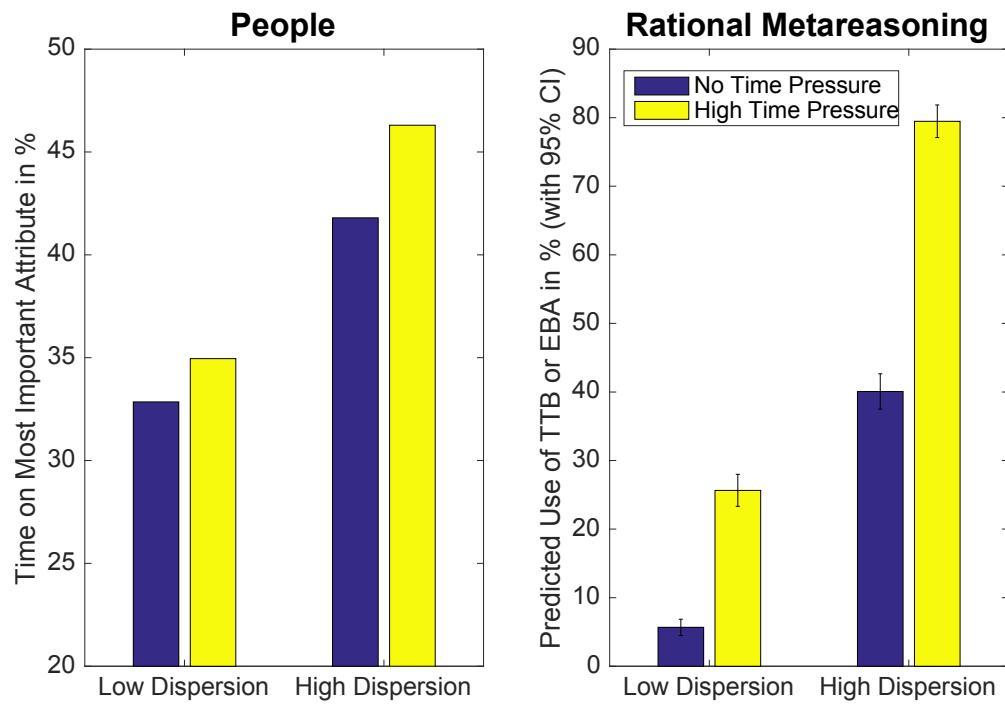
Figure 4.4: Rational metareasoning predicts the increase in selective attribute-based processing with dispersion and time pressure observed by (Payne et al., 1988).

be useful for answering this important question and finding better ways to teach problem-solving.

Previous experiments suggest that people choose problem-solving strategies adaptively depending on task and context (Fum & Del Missier, 2001). This ability to choose problem-appropriate strategies appears to be learned from experience. Gunzelmann and Anderson (2001, 2003) found that people learn to solve Tower of Hanoi problems in increasingly fewer steps by transitioning from in-effective guessing strategies to increasingly more effective planning strategies. In the classic Tower of Hanoi task the participant is presented with a large disk, a medium disk, and a small disk that are distributed over three pegs. Their task is to move the disks from their initial configuration into a target configuration under the constraints that they can move only one disk at a time and that a larger disk must not be placed on top of a smaller disk.

In this section, we investigate whether our rational metareasoning model can explain how people learn that they can solve problems more effectively by planning more. To do so, we simulate the experiment by Gunzelmann and Anderson (2001) that posed people problems that were structurally equivalent to the Tower of Hanoi task. Each participant solved a series of 12 problems in which both the initial configuration and the target configuration were flat states – configurations in which each of the three disks is on a different peg. The target state was always five moves away from the starting state. They found that each of their participants' moves followed one of three strategies: the *subgoal strategy*, the *flat-to-flat strategy*, or the *guessing strategy*. The subgoal strategy solves the problem in three steps: Its first subgoal is to moves the large disk to its target location. If this cannot be accomplished in a single move, then it adds subgoals to vacate the target peg and/or free the large disk. If one of these sub-tasks cannot be achieved by a single move, it adds another subgoal to make enable it. Once the largest disk is in the correct place, the subgoal to move the medium disk to its target location will be pursued, and so on. The flat-to-flat strategy transforms one flat state into another by swapping the small disk with either the large disk or the medium disk depending on which of the resulting flat states is closer to the goal. Finally, the guessing strategy chooses a legal move at random. The process of choosing a strategy, executing it, and updating the model is repeated until the problem is solved.

The experiment by Gunzelmann and Anderson (2001) comprised two blocks of planning problems separated by a filler task. There were 2×2 conditions defined by the difficulty of the problems in the first block (easy vs. difficult) and the difficulty of the problems in the second block (easy vs. difficult). In the *hard* problems moving the large disk to its target location requires two nested sub-goals: the subgoal to remove the medium disk from the target location and the sub-subgoal to vacate the peg to which the medium disk should be moved. In the easy problems, by contrast, the target

location of the large disk can be vacated directly without setting a sub-subgoal. The sequences of moves taken by the participants were analyzed to diagnose which of the three strategies were used on each trial. These analyses revealed that the frequency of the random-strategy and the flat-to-flat strategy decreased over time while the frequency of the subgoal strategy increased. These effects occurred in all four conditions of the experiment. The decrease in the frequency of the random strategy indicates that people learned to plan more, and the increase in the frequency of the subgoal strategy relative to the flat-to-flat strategy indicates that people learned to plan more effectively. According to the ACT-R model by Gunzelmann and Anderson (2001)(Gunzelmann & Anderson, 2001), these adaptive changes result from people learning that planning is more effective than initially assumed whereas guessing is less effective than initially assumed.

Next, we investigate whether our rational metareasoning model can explain these phenomena as well as the ACT-R model by Gunzelmann and Anderson (2001) and which components of our model are necessary to capture them.

4.5.1   Model and Methods

To simulate the experiment by Gunzelmann and Anderson (2001), we applied our model to selection between the three strategies defined above: the guessing strategy, the flat-to-flat strategy, and the subgoal strategy. Our simulation assumed that feedback is binary: the reward is one if the strategy reached the target state and zero otherwise. According to our model, people learn to predict the performance of each strategy from three simple features of Tower-of-Hanoi problems: the number of disks that are in the wrong position, whether or not the current state is flat, and the height of the tallest tower. In addition, the Bayesian regression models of the strategies' performance and execution time included a constant term that captures all features that do not vary between three-disk, flat-to-flat Tower of Hanoi problems. The model's parameters specified how long it takes to execute a move (τ_{move}), the time required for planning how to place a disk in the right place measured in moves (τ_{plan}), the expected performance of guessing (ρ_{guess}), the expected performance of planning (ρ_{plan}), and the precision of the prior distributions of the execution time models and the reward models (π_{prior}).

The prior on the coefficients of the predictive models of the strategies' execution time and reward were set to $P(\beta_T^{(s)}) = \mathcal{N}(\mu_T^{(s)}, \Pi_T^{-1})$ and $P(\beta_R^{(s)}) = \mathcal{N}(\mu_R^{(s)}, \Pi_R^{-1})$ respectively where $s \in$ {guess,flat-to-flat,subgoal} indexes the strategy:

$$
(\mu_T^{\text{(guess)}} \mu_T^{\text{(flat-to-flat)}} \mu_T^{\text{(subgoal)}}) = \tau_{move} \cdot \begin{pmatrix} 0 & \tau_{\text{plan}} & \tau_{\text{plan}} \\ 0 & -1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},
$$

$$
(\mu_R^{\text{(guess)}} \mu_R^{\text{(flat-to-flat)}} \mu_R^{\text{(subgoal)}}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \rho_{\text{guess}} & \rho_{\text{plan}} & \rho_{\text{plan}} \end{pmatrix},
$$

$$
\Pi_T = \Pi_R = \pi \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.
$$

The rows of the matrices above correspond to features listed above and columns correspond to strategies (guessing, flat-to-flat, subgoal strategy). The prior on the error variance of the execution time model was Gamma(1,1).

Following Gunzelmann and Anderson (2001), we assumed that executing a move takes 4 seconds $\tau_{\text{move}} = 4$. Gunzelmann and Anderson (2001, 2003) also assumed that the relative cost of time is such that achieving the goal is worth 50 seconds. We therefore fixed the agent's estimate of its opportunity cost to $\frac{1}{50\text{sec}}$ and set the reward of achieving the goal to 1. We modeled the time taken by the random strategy as the time per move times number of random moves. The subgoal strategy and the flat-to-flat-strategy are associated with an additional time cost of planning. We modeled the cost of planning by the number of subgoals times the time it takes to set a subgoal and plan how to achieve it (i.e. τ_{subgoal}). Our model thereby captures that executing the subgoal strategy takes longer on problems that require more subgoals (*difficult problems*) than on problems that require fewer subgoals (*easy problems*). For the flat-to-flat strategy the planning time was set to τ_{subgoal} to reflect an intermediate amount of planning.

To determine which components of our model are supported by the data, we also evaluated the five lesioned metareasoning models described above and three SCADS models with the reward functions described above (i.e., $r = \text{correctness}$, $r = \text{correctness}\,\text{cost}$, and $r = \text{correctness}/\text{cost}$

respectively). The SCADS models were equipped with four categories. The first category encoded whether the task was a flat-to-flat problem. The three remaining categories encoded whether the large disk was on the correct peg, whether the medium disk was on the correct peg, and whether the small disk was on the correct peg respectively. Each SCADS model had four free parameters. The first two parameters encoded the relative strengths of the initial associations between the categories and the three problem solving strategies. The third parameter encoded the total strength of the initial associations, and the fourth parameter encoded the subgoaling time.

The model parameters were estimated by maximum-likelihood estimation from the average number of exploratory moves by problem number, the average length of the final path by problem number, and the strategy choice frequencies for the subgoal strategy and the flat-to-flat strategy in the experiment by Gunzelmann and Anderson (2001). Our likelihood model assumed that the standard error of the number of moves was 1. We maximized the likelihood functions using a state-of-the art Bayesian optimization algorithm known as Infinite Metric Gaussian Process Optimization (IMGPO)(Kawaguchi, Kaelbling, & Lozano-Pérez, 2015). Since the likelihood function was not available analytically, we estimated the likelihoods of the data given each set of parameters considered by IMGPO by simulating the experiment 500 times and smoothing the simulated strategy choice frequencies according to Equation 5. The optimization algorithm was run for about 48 hours per model which corresponded to about 64 iterations; this was probably sufficient to find decent parameter estimates because IMGPO converges exponentially fast.

4.5.2 Results

The parameter estimates for the rational metareasoning model were $\tau_{\mathrm{plan}} = 3.39$, $\rho_{\mathrm{plan}} = 0.56$, $\rho_{\mathrm{guess}} = 0$, $\pi_{\mathrm{prior}} = 10.5$, and $\tau_{\mathrm{subgoal}} = 8.5$. According to these estimates, people initially assume that planning is more effective than guessing but takes much longer. The strength of this initial belief corresponds to 5-6 observations per strategy and problem type, and that setting a subgoal and planning how to achieve it takes about as much time as executing 8.5 moves, that is 34 seconds. Thus, according to our model, people initially believe that the planning strategy would take at least 108 seconds whereas achieving the goal is worth only 50 seconds. Based on this, it appears that people learn to plan more by realizing that planning takes less time and is more effective than they thought whereas guessing is less effective than they thought.

Rational metareasoning achieved a very good fit to people's strategy choices. Our model captured that people learn to plan more with practice: it correctly predicted people's increasing use of
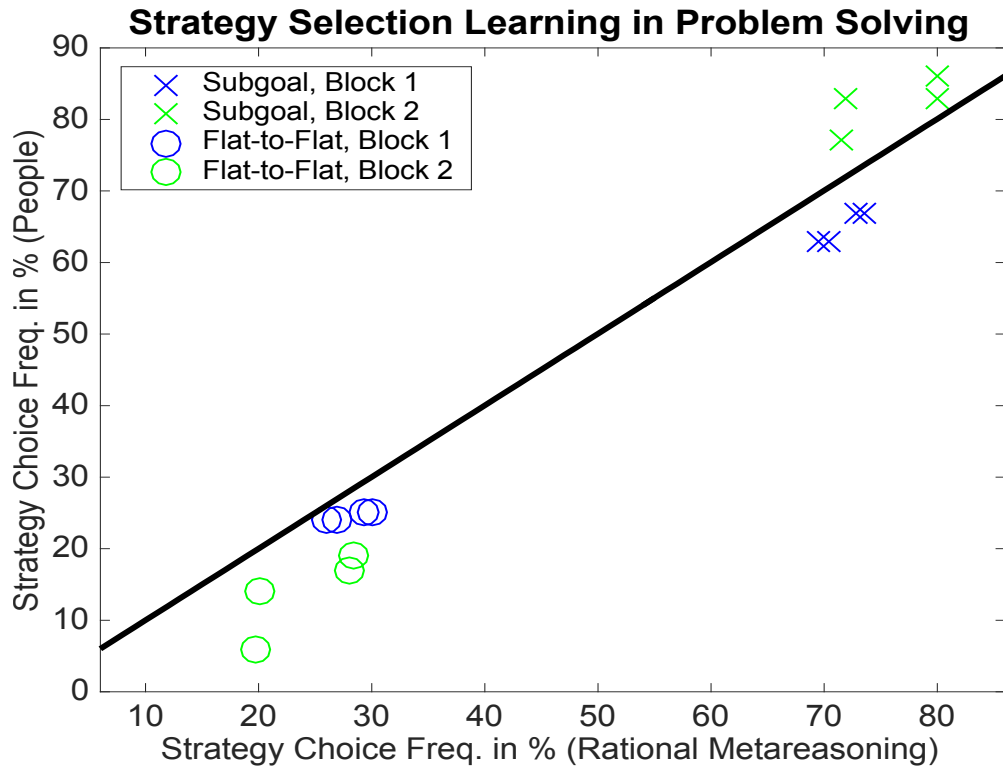
Figure 4.5: Strategy selection learning by rational metareasoning captures changes in people's use of the sub-goals strategy in the experiment by Gunzelmann and Anderson (2001). Rational metareasoning captures the finding that people learn to plan more from the first block to second block.
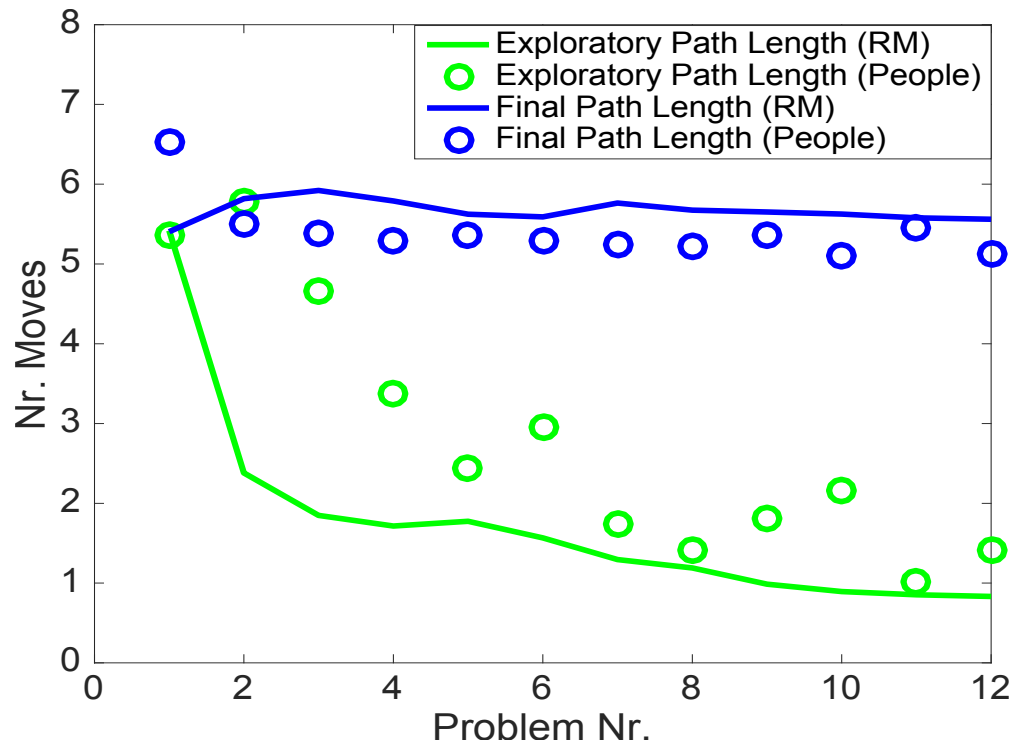
Figure 4.6: Predicted versus measured exploratory and final path lengths as a function of the trial number. Rational metareasoning captures the gradual shortening of path exploratory and final paths.

planning strategies from the first block to the second block (see Figure 4.1) and the decreases in the number of exploratory moves before people launch into a planned-out solution (Figure 4.6). As shown in Figure 4.1, our model explained 96% of the variance in the frequency with which people chose the sub-goal strategy or the flat-to-flat strategy in the different blocks of the experiment $(r(14) = 0.9818, p < 0.001)$. This is very close to the 97% explained by Gunzelman and Anderson's model $(r(14) = 0.9869, p < 0.001)$ even though their model had at least twice as many parameters as ours. This includes the difference in the frequencies of the sub-goal strategy versus the flat-to-flat strategy and the variability of each strategy's usage frequency across blocks and conditions. Our model explained about 52% of the variance in the average frequency with which people chose the subgoal strategy $(r(6) = 0.7244, p = 0.04)$ and about 66% of the variance in the average frequency with which they chose the flat-to-flat strategy $(r(6) = 0.8145, p = 0.01)$. Hence, our model explained the use of the subgoal strategy equally well as the model by Gunzelmann and Anderson $(r(6) = 0.88, p = 0.004, R^2 = 0.78)$ and predicted the use of the flat-to-flat strategy more accurately than their model $(r(6) = 0.67, p = 0.07, R^2 = 0.45)$.

The average number of moves that rational metareasoning predicts as a function of time, difficulty, and condition was highly correlated with the number of moves in people's solutions $(r(14) = 0.8791, p < 0.001$; see Figure 4.7). Our model's fit $(R^2 = 0.77)$ was not as good as that of the ACT-R model by Gunzelmann and Anderson $(r(14) = 0.95, R^2 = 0.91)$ which was specifically designed for this task. The moves that participants made comprised an initial exploratory phase and a final solution phase. As shown in Figure 4.6, our model captures that the number of exploratory moves decreases substantially $(r(10) = 0.7855, p = 0.003)$ whereas the observed and predicted lengths of the final paths decreased very little. Finally, our model also captures the differential effects of hard versus easy problems (see Figure 4.7). When people were presented problems that required fewer subgoals in the first block ("easy problems") and problems that required more subgoals in the second block ("hard problems"), then the average number of moves decreased initially but increased again in the second block. According to theory, the subgoal strategy requires more planning time for hard problems than for easy problems. Thus, when people encounter hard problems their estimate of the execution time of the planning strategy increases. Consequently, they choose the subgoal strategy less often and will more often choose the flat-to-flat strategy that requires more moves but less planning time.

Bayesian model comparisons revealed that the learning effects reported by Gunzelmann and Anderson (2001) provide strong evidence for the full rational metareasoning model (BIC $= 213.4$) over the lesioned metareasoning model without exploration (BIC ≥ 221.1) and very strong

evidence for the full rational metareasoning model over all other lesioned metareasoning models (BIC $\geq$ 292.4) and all SCADS models (BIC $\geq$ 293.2). This suggests that feature-based learning, exploration, and model-based strategy selection are necessary to capture people's capacity to learn how to solve problems; Figure 4.8 provides a more detailed summary of these results.

4.5.3 Discussion

In this section, we have shown that our rational metareasoning model of strategy selection learning can capture how people's choice of problem solving strategies changes as they practice the Tower of Hanoi task (Gunzelmann & Anderson, 2001, 2003). Concretely, our model captures that people learn to plan more (Figure 4.5) and come to solve the Tower of Hanoi task in fewer and fewer moves by avoiding unsystematic, exploratory moves (Figure 4.6). According to our model, these changes occur because people learn that planning takes less time and is more accurate than initially assumed whereas guessing performs worse than initially expected. This might be the case because people's prior beliefs about the effort of planning and its benefits reflect the difficulty of the much more complex problems they face in their everyday life where planning takes longer and plans often do not work out as expected. Alternatively, planning and setting subgoals might be rather time consuming initially but become faster and more efficient with practice in the task. Figure 4.5 shows that our model also captures that as people practice they come to prefer the more effective but more effortful subgoal strategy over the easier but less effective flat-to-flat strategy. Finally, our model also captured the differential effects of practicing on easy versus difficult Tower of Hanoi problems (Figure 4.7).

The ACT-R model by Gunzelmann and Anderson (2001, 2003) achieved a slightly better fit to the participants' numbers of moves but our model predicted the use of the flat-to-flat strategy more accurately. Hence, a simple application of our general rational metareasoning framework yielded a model that explains Gunzelmann and Anderson's data about as well as the model that they tailored to this data set.

The main advantage of our theory over the model by Gunzelmann and Anderson is its ability to explain a much wider range of phenomena, as the other sections of this chapter illustrate. For instance, the strategy selection learning mechanism of the ACT-R model (Gunzelmann & Anderson, 2001, 2003) presupposes that rewards are binary (goal achieved vs. not achieved) whereas rational metareasoning is also applicable when the feedback is a continuous reward signal such as a payoff. In addition, our theory can explain people's adaptive flexibility in strategy selection and the transfer of strategy selection learning effects between different types of problems. The ACT-R model by Gun-
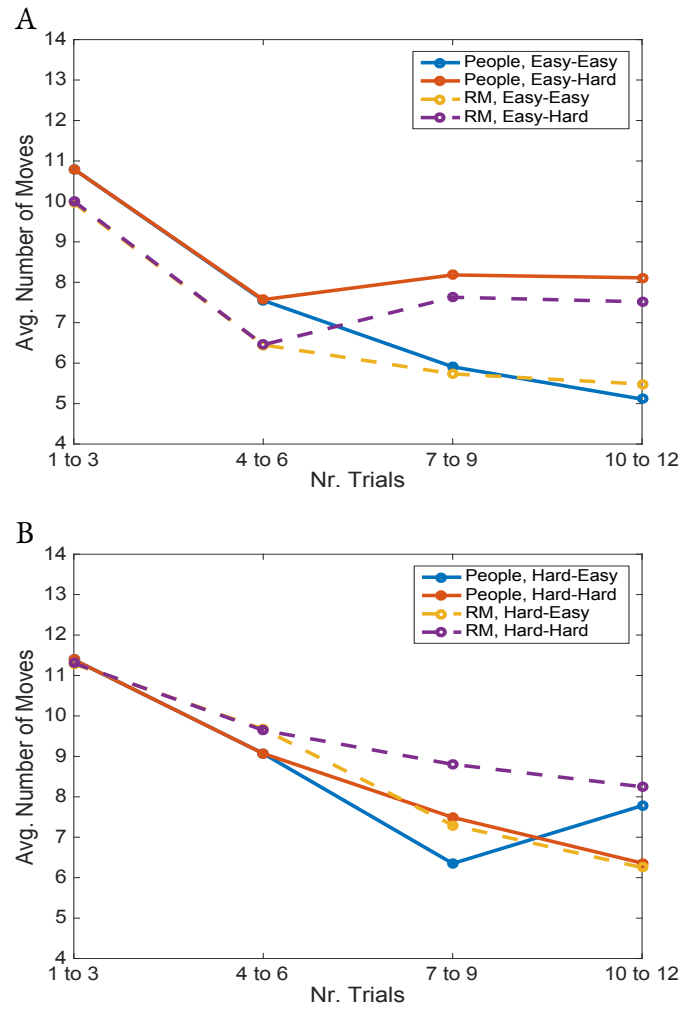
Figure 4.7: Strategy selection learning according to rational metareasoning (RM) captures how the length of people's solutions of Tower of Hanoi problems changes with experience.
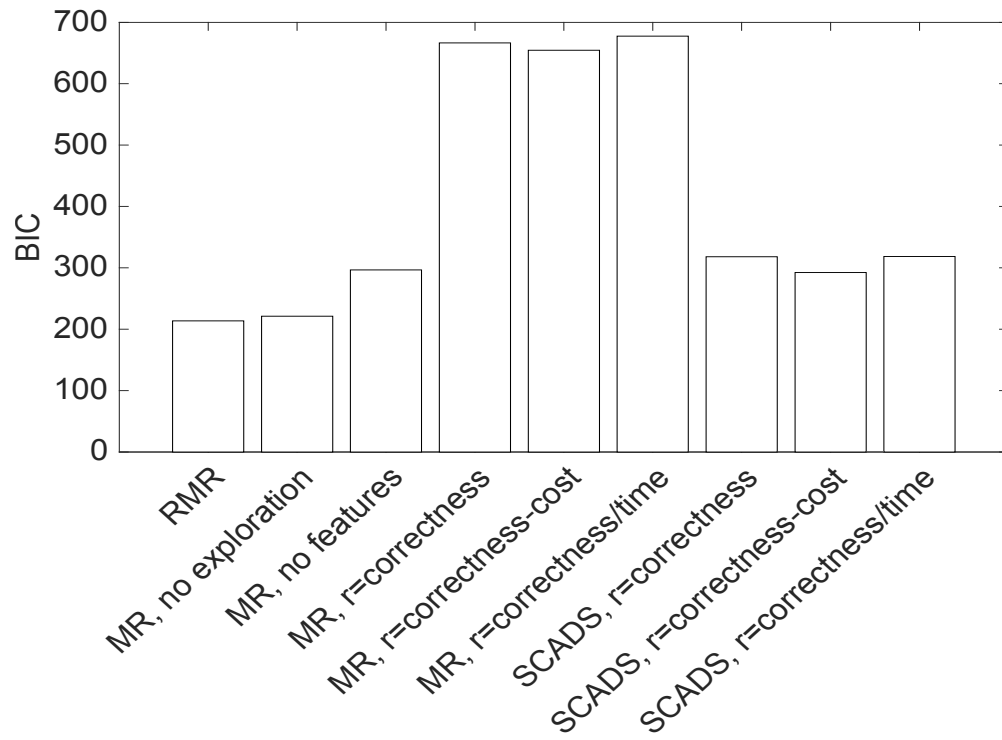
Figure 4.8: Model selection applied to the problem solving data from Gunzelmann and Anderson.

zelmann and Anderson (2001, 2003) cannot capture these phenomena, because its learning mechanism does not consider the context in which the strategy was applied. Therefore, it cannot achieve graded or selective generalization. It can only learn the value of applying a rule that is associated with certain conditions. By contrast, our rational metareasoning model learns about the relationship between features of individual problems and expected performance, and this allows it to account for graded and selective transfer to similar versus dissimilar problems. This difference does not manifest in scenarios with only a single type of task, namely flat-to-flat Tower-of-Hanoi problems with three disks and three pegs that require five moves to be solved. Yet, this difference might become visible if we simulated how people learn to solve different kinds of problems that required different strategies. Designing and running such experiments to discriminate between the ACT-R model of strategy selection and our rational metareasoning model is an interesting direction for future research.

Interestingly, the strategy selection mechanism of ACT-R applies not only to entire strategies but also to their sub-strategies and elementary operations. Thus, ACT-R models could, in principle, learn to chain multiple strategies none of which would be able to solve the problem on its own. To capture such phenomena rational metareasoning models will be extended from learning about the immediate reward and cost of executing a strategy to learning to predict the sum of immediate and future returns according to Q-learning (Watkins & Dayan, 1992). We will revisit this idea in the General Discussion.

4.6 STRATEGY SELECTION AND COGNITIVE DEVELOPMENT

So far, we have found that adults' strategy choices in sorting, decision-making, and planning become increasingly more rational through learning within minutes. Since learning is an important driving force of cognitive change our theory predicts similar phenomena should also occur on the much longer time-scales of cognitive development.

A substantial literature on the development of children's arithmetic competencies suggests that cognitive development does not proceed in a sequence of discrete stages characterized by a progression of beliefs, representations, and cognitive strategies as proposed by Piaget (Piaget, 1952) but rather as a gradual shift in the frequency with which children use each of multiple coexisting cognitive strategies (Siegler & Shipley, 1995). According to Siegler's *overlapping waves* theory of cognitive development (Siegler, 1996, see Figure 4.9A) children of every age use a variety of strategies, and over time strategies that are both effective and efficient come to be used more frequently.

To give just one example of such *strategic development* (Siegler, 1999) we consider the development of children's strategies for mental addition shown in Figure 4.9B. (Svenson & Sjöberg, 1983) found that the Retrieval strategy becomes increasingly more prominent as children get older, while the frequency of not providing an answer drops rapidly. The frequency of the *Sum* strategy rises initially making it the most common strategy at the beginning of second grade, but afterwards its frequency drops again. The frequency of the *Min strateg*y rises initially, and then it stays roughly constant until the Min strategy is overtaken by the *Retrieval strategy*.

Children's strategic development raises the question of how they learn to use effective strategies more and less effective strategies less. Learning to use effective strategies is complicated by the fact that each strategy's effectiveness differs from one problem to the next: a strategy that works excellently for one type of problem may fail miserably on a different kind of problem. According to the SCADS model by Shrager and Siegler (1998), children solve this problem by gradually strengthening the association between the type of the problem solved and the strategy used after every correct answer. However, this model presupposes that children already know how to categorize problems in such a way that problems within the same category require the same strategies. Furthermore, the SCADS model presumes that learning is driven solely by whether or not the strategy produced the correct answer. This ignores the effort and time required to execute the strategy, and the mechanism is difficult to apply when performance feedback is continuous, as in economic decisions, rather than binary. Furthermore, even when those limitations are overcome the specific learning mechanism of the SCADS model appears to fail in some situations in which humans succeed (Lieder, Hsu, & Griffiths, 2014).

Our rational metareasoning model overcomes these limitations of the SCADS model. It could thus be used to model strategic development in domains that do not comply with the assumptions of the SCADS model. However, the applicability of our model to cognitive development remains to be evaluated. In this section we provide a proof of concept that our model can capture the developmental progression of children's cognitive strategies in the domain of mental arithmetic. To do so, we simulate the development of children's strategies for mental addition (Svenson & Sjöberg, 1983) according to rational metareasoning.

We recreated Shrager and Siegler's simulation of the development of children's strategy use for single-digit addition problems in which both summands lie between 1 and 5 (Shrager & Siegler, 1998; Svenson & Sjöberg, 1983) with our strategy selection model. To make the model predictions as comparable as possible, we retained all of the assumptions that Shrager and Siegler made about children's strategies. Concretely, we assumed that children use the following four strategies for mental addi-
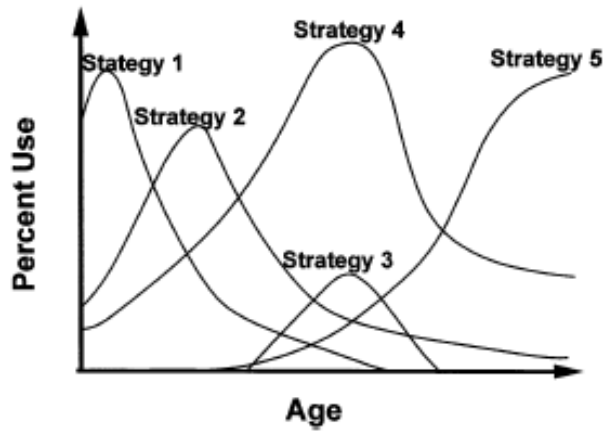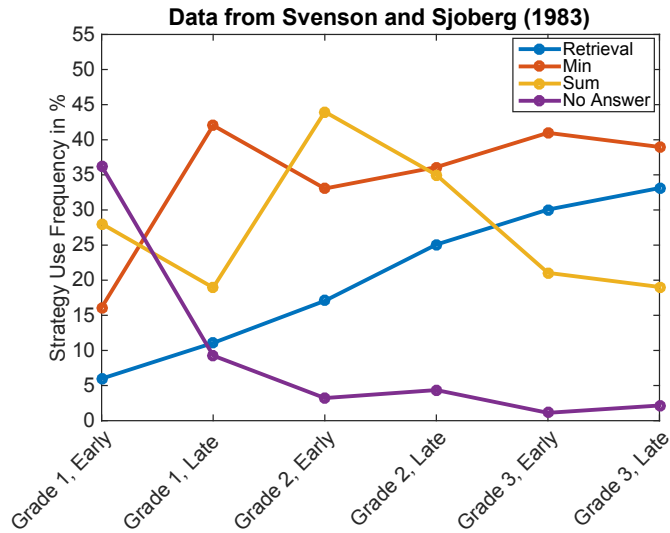
Figure 4.9: Overlapping waves theory of cognitive development. (A) Illustration of the theory from (Siegler, 1999). (B) Empirical support for overlapping waves in the development of children's strategy use in mental addition according to (Svenson & Sjöberg, 1983).

tion (illustrating them for the example of calculating $3 + 5$):

1. *Retrieval*: retrieve the answer from memory ("8").

2. *Sum*: First, use the fingers of one hand to count up to the first summand ("1-2-3"). Then use the fingers of the other hand to count up the second summand ("1-2-3-4-5"). Finally, count

the total number of raised fingers on either hand ("1-2-3-4-5-6-7-8").

3. *Shortcut Sum*: First, use the fingers of one hand to count up to the first summand ("1-2-3"). Then continue counting from the first summand while raising the fingers of the second hand one-by-one until it shows the second summand ("4-5-6-7-8").

4. *Min*: Start counting upwards from the larger summand ("6-7-8").

These four strategies differ in how many counting operations they require to solve any given problem. To account for the discovery of the Shortcut Sum strategy and the Min strategy, our metareasoning models start out with only the Retrieval strategy and the sum strategy. The Shortcut Sum strategy and the Min strategy are added after 90 and 95 trials respectively because this is how long it took children to discover those strategies in a study by Siegler and Jenkins (1989). To simulate reaction times, we assumed that each counting operation takes about half a second as indicated by the findings of Geary, Brown, and Samaranayake (1991). Following Shrager and Siegler (1998), errors were modeled by assuming that each counting step is incorrectly executed with probability $p_{\text{error}} = 0.04$. We generated the number of incorrectly executed steps by drawing from the binomial distribution Binomial (#steps,p_{error}). The effect of each error was to either omit a counting operation, for example "3,3" instead of "3,4", or to skip a number, for example "3,5" rather than "3,4".

To model the Retrieval strategy, we modeled children's memory for arithmetic facts by the associative memory model used in the SCADS model (Shrager & Siegler, 1998; Siegler & Jeff, 1984; Siegler & Shipley, 1995) with the same set of parameters. This model characterizes memory for arithmetic facts by how strongly each possible answer a is associated with each problem $x + y$. The state of a child's long-term memory for arithmetic facts can therefore be described by a three-dimensional matrix $A(a, x, y)$ of associative strengths. For the most familiar addition problems whose first or second summand was 1 the associative strength was initialized with 0.05. For all other addition problems, the associative strengths were initialized by $1/(10 \cdot \#\text{values})$. Each time a strategy produced an answer the strength of the association between the answer and the pair of summands was increased by 0.06 if the answer was correct or by 0.03 when the answer was wrong. Each time the Retrieval strategy is used it samples a confidence criterion between 0 and 1 uniformly at random. The probability that a potential answer will be sampled is its associative strength divided by the sum of the associative strengths of all possible answers. If the associative strength of the sampled answer exceeds the confidence criterion, then the answer is reported. Otherwise the sampling process is repeated. If no answer's associative strength exceeded the confidence threshold after 10 attempts, then the

Retrieval strategy fails to answer the question. The execution time of the Retrieval strategy was modeled as 0.5 seconds times the number of retrieved answers.

To apply our rational strategy selection learning model to mental addition, we have to specify how problems are represented, the form of the meta-level model, and children's prior knowledge about the performance of addition strategies. We assume that children represent the addition problem $x + y =$? by three simple features

$$\mathbf{f} = (f_1, f_2, f_3) = (s_1, s_2, \max_{\mathrm{a}} A(a, x, y)),$$

where the third feature is the associative strength of the answer that is most strongly associated with the problem in memory. Since the feedback that children receive in mental addition is binary ("right" or "wrong"), the meta-level model is

$$P(R = 1 \mid \mathrm{f}, S = i) = \frac{1}{1 + \exp(\beta_i + \sum_{k=1}^{3} \alpha_{k,i}^{(R)} \cdot f_k)},$$

where the bias term b_i captures influences on the strategy's performance that are not captured by the features of the problem to be solved. We model children's prior knowledge about the performance of addition strategies by the model's prior on the bias weights. The simulations by Shrager and Siegler (1998) and Siegler and Shipley (1995) assumed that children initially know only the *Retrieval strategy* and the *Sum* strategy but have to discover the more efficient strategies on their own, since parents teach the *Sum* strategy first and memory retrieval is a domain general capacity that precedes knowledge of arithmetic. To capture these assumptions our simulations assume that children's prior expectation about the strategies' performance is positive for the *Sum* strategy $(P(b_2) = \mathcal{N}(5, 1))$, neutral for the familiar *Retrieval strategy* $(P(b_2) = \mathcal{N}(0, 1))$, but negative for the other strategies that are still unfamiliar $(P(b_3) = P(b_4) = P(b_5) = \mathcal{N}(-5, 1))$. As in all previous applications of our model, the meta-level model uses Bayesian linear regression to predict each strategy's execution time from each problem's features. The relative cost of time was set such that finding the correct answer was worth 100 seconds. Since this corresponds to each child's subjective utility of being correct, this simulation assumed that the opportunity cost is known and does not have to be learned. To determine the predictions of our rational metareasoning model, we simulated the 200 virtual participants' choices of addition strategies across 100 blocks of 10 addition problems each. Addition problems were independently generated by randomly sampling the first and the second summand from two independent uniform distributions over their possible values, that is 1, 2, 3, 4, or 5.

A
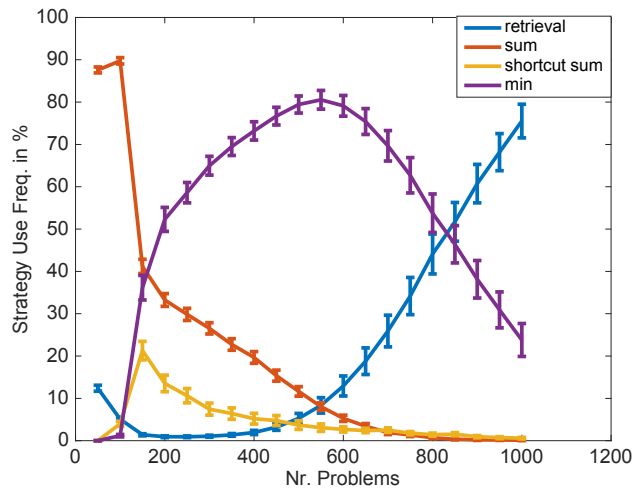
B
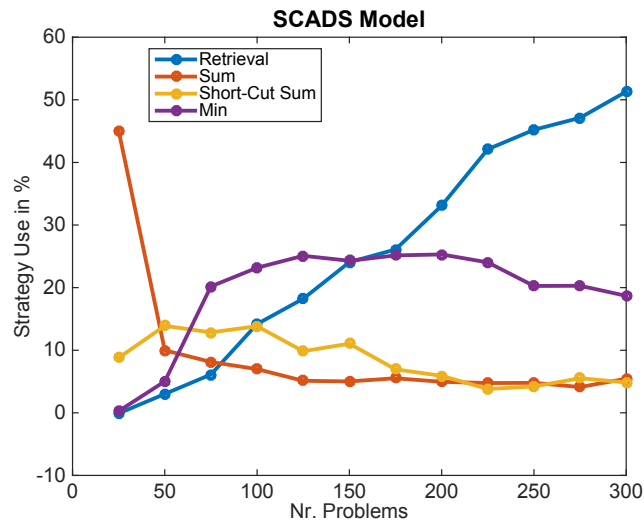
Figure 4.10: Comparison of models of how children learn to select arithmetic strategies. Error bars enclose 95% confidence intervals. (A) Predictions of rational metareasoning. (B) Predictions of the SCADS model according to Shrager and Siegler (1998).

The simulation results shown in Figure 4.10 suggest that our rational theory of strategy selection learning can capture the qualitative changes in children's use of addition strategies observed by Svenson and Sjöberg (1983): Our simulation captures the transient rise and fall of accurate but effortful

A
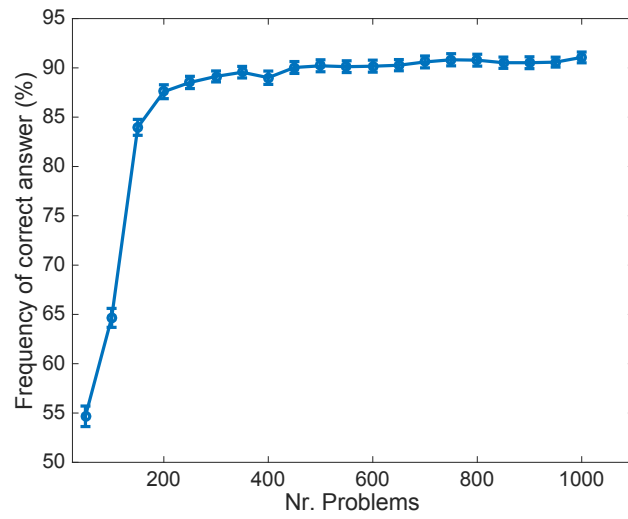
B
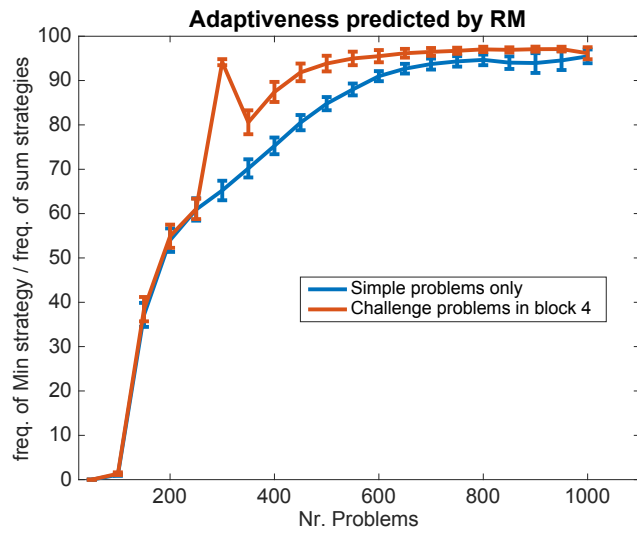
Figure 4.11: Learning curves of rational metareasoning simulations of children's strategic development in mental arithmetic. Error bars enclose 95% confidence intervals. (A) Gradual increase in performance predicted by rational metareasoning (RM). (B) Generalization of the Min strategy with versus without challenge problems.

addition strategies, the shift toward the more efficient Min strategy, and the eventual transition towards the predominant use of the Retrieval strategy. Comparing the predictions of our model with

those of the SCADS model (Figure 4.10) suggests that both models capture the same developmental trends about equally well. Furthermore, like the SCADS model, rational metareasoning also captures the gradual increase in children's performance (see Figure 4.11A) and the transfer from simple addition problems with summands ranging from 1 to 5 to more challenging addition problems with one addend above 20 and the other addend below 5. As shown in Figure 4.11B, rational metareasoning predominantly selected the most accurate and the most efficient approach, namely the *Min strategy*, to solve the challenge problems even though it had never encountered any of those problems before.

Like the SCADS model, our model captures the increasingly adaptive strategy choices that children make: our model learned to use the Retrieval strategy more often for easy problems than for hard problems. This is adaptive because the Retrieval strategy is less accurate on hard problems because, due to past mistakes, hard problems are more strongly associated with wrong answers than easy problems. In our simulation, the correlation between a participant's average performance on a problem and the frequency with which they used the *Retrieval strategy* increased from $r(4998) = -0.11$ ($p < 0.001$) in the first 500 problems to $r(4998) = 0.28$ ($p < 0.001$) in problems 501 to 1000. In addition, our model learned to choose the *Min strategy* over less efficient and more error-prone addition strategies when the *Retrieval strategy* appeared inapplicable. Furthermore, the model learned to choose the *Min strategy* adaptively: The advantage of the *Min strategy* over alternative addition strategies increases with the sum and the difference between the addends. Across all simulated trials, the model's choice of the *Min strategy* was significantly correlated with the sum ($r(141609) = 0.30$, $p < 0.001$) and the absolute value of the difference between the addends ($r(141609) = 0.24$, $p < 0.001$). Furthermore, the correlation with the sum or the difference was stronger than the correlation with other factors such as the product ($r(141609) = 0.20$, $p < 0.001$). In addition, the model's choices of the *Min strategy* became more adaptive: Shortly after the discovery of the Min strategy (trials 100-150 to be precise) its use was less well predicted by the difference between the two summands ($r(4998) = 0.17$, $p < 0.001$) than by their product ($r(4998) = 0.32$, $p < 0.001$), but ten blocks later the difference between the two summands predicted the choice of the *Min strategy* ($r(4998) = 0.23$, $p < 0.001$) better than their product ($r(4998) = 0.09$, $p < 0.001$) as in Siegler and Shipley (1995).

As shown in Figure 4.11B, the proportion of applications of the Min strategy out of all addition strategies increased steadily from 37.2% in the first 50 trials after its discovery towards 100%. The learning curve shows that the process by which the Min strategy is generalized from one problem on which it worked well to all other problems is gradual and takes more than 1000 examples. This

is consistent with the empirical finding that children are slow to generalize the Min strategy to other problems upon its discovery. Siegler and Jenkins (1989) found that the generalization of the Min strategy proceeds much more rapidly when children who have recently discovered the Min strategy are posed challenge problems such as $4 + 25$. Shrager and Siegler (1998) modeled the experiment by Siegler and Jenkins (1989) by replacing the 50 simple problems presented after the discovery of the Min strategy by 50 challenge problems in which one of the addends is larger than 20 and the other addend is smaller than 5. We performed the equivalent simulation with our rational metareasoning model by replacing the 50 problems following the first five blocks by 50 challenge problems. As shown in Figure 4.11B, feature-based strategy selection learning captures the empirical finding that challenge problems boost children's transfer of the Min strategy to challenging as well as simple problems (Siegler & Jenkins, 1989). To test if the observed differences were significant, we performed one t-test for each of the 20 simulated blocks of 50 problems with the Bonferroni-corrected significance level of $0.05/20 = 0.0025$. We found that the average adaptivity was not significantly different before the challenge problems (all $p \geq 0.20$) but became highly significant once the challenge problems were introduced ($p < 0.001$) and remained statistically significant until block 19 ($all p \leq 0.02$) after which the performance of both groups reached its asymptote (all $p \geq 0.50$).

To determine which components of our model were critical to capture the development of children's choice of addition strategies, we reran the simulation with the five lesioned metareasoning models. We found that exploration is necessary for strategic development, because without exploration the rational metareasoning model never discovered the Shortcut Sum strategy or the Min strategy, and it failed to switch to the Retrieval strategy even after it had plenty of experience to rely on (Figure C.6). Feature-based strategy selection was also critical, because the metareasoning model without features predicted that children would transition directly from the Sum Strategy to the Retrieval strategy without using the Shortcut Sum or the Min strategy in between (Figure C.7). This might be because the features are necessary to learn that the Retrieval strategy works only when the problem is familiar whereas the Min Strategy is superior for unfamiliar problems where one of the addends is small. Likewise, the lesioned metareasoning model that maximized accuracy regardless of time cost never discovered the Min strategy or the Shortcut Sum strategy but transitioned directly from the standard Sum strategy to memory retrieval (Figure C.8). Model-free metacognitive reinforcement learning of the VOC (r = reward-cost) predicted that the Sum strategy would fade much faster than it has been observed in children and failed to predict children's eventual transition to the Retrieval strategy (cf. Figure 4.9B vs. Figure C.9) Finally, model-free learning of the reward rates predicted an almost instantaneous shift to the Min strategy and also failed to predict the subsequent

transition to the Retrieval strategy (see Figure C.10). These findings suggest that maintaining separate representations of execution time, opportunity cost, and expected reward enables faster learning and adaptation to changes in the strategies' performance or the reward rate.

In this section, we have demonstrated that rational metareasoning can explain several qualitative features of the shifts in children's choice of addition strategies. Most importantly, feature-based strategy selection learning formalizes the overlapping waves theory of cognitive development (Siegler, 1996) by a powerful, general learning mechanism. This suggests that our model should be able to capture similar phenomena in other domains of cognitive development as well. However, the change in children's strategy choices explained by our model is only one of three parts of strategic development, which also includes the discovery of new strategies and the change of existing strategies. To overcome this limitation, future work should combine our model of strategy selection learning with models of strategy discovery and strategy change. We will revisit this future direction in the General Discussion.

Feature-based strategy selection learning is more widely applicable than the basic SCADS model. Unlike the SCADS model our model can also learn from continuous feedback, as well as execution time or mental effort, and it does not presuppose that problems can be categorized appropriately. On the other hand, the SCADS model captures an important mechanism that is not yet included in our resource-rational account of strategic development: strategy discovery. Both mechanisms play an important role in strategic development. Therefore, our contributions are more complementary than they are in competition. Formalizing the computational mechanisms of strategy discovery and the formation of mental habits within the rational metareasoning framework is a promising direction for future research. To apply rational metareasoning to the strategy discovery problem, future research might combine learning to predict the VOC of individual computations from features of the current mental state with techniques from hierarchical reinforcement learning (Barto & Mahadevan, 2003; Barto, Singh, & Chentanez, 2004; Botvinick, Niv, & Barto, 2009; Sutton, Precup, & Singh, 1999).

4.7   General Discussion

How do we know when to think fast and when to think slow? Do we use our heuristics rationally or irrationally? How good are we at selecting the right strategy for the right problem? To answer these questions, we derived a rational solution to the strategy selection problem and evaluated it against

human behavior and previous theories of strategy selection.

The results presented in this chapter support the conclusion that people gradually learn to use their cognitive strategies increasingly more rationally. According to our rational metareasoning model, these adaptive changes result from a rational metacognitive learning mechanism that builds a predictive model of each strategy's execution time and accuracy.

Jointly, the experiments, simulations, and model comparisons reported in this chapter provided very strong evidence for all four components of our model: strategy selection based on an approximate cost-benefit analysis, feature-based metacognitive reinforcement learning, separate predictive models of accuracy and execution time, and the exploration of alternative strategies.

Our model's predictions captured the variability, contingency, and change of people's strategy choices in domains ranging from sorting to decision-making, and mental arithmetic as well as problem solving. Our model provides a unifying explanation for a number of phenomena that were previously explained by different models. Overall, the dependence of people's strategy choices on task and context variables was consistent with a rational strategy selection mechanism that exploits the features of each problem to achieve an optimal cost-benefit tradeoff. Likewise, the change in people's strategy choices over time was consistent with rational learning of a predictive model of each strategy's performance and choosing strategies rationally with respect to the model learned so far. This learning mechanism simultaneously accounts for the developmental progression of children's arithmetic competence on a time scale of years and the adaptions of adults' decision strategies on a time scale of minutes. The remaining variability of people's strategy choices was consistent with the near-optimal exploration-exploitation tradeoff of Thompson sampling.

Critically, our new experiments and simulations showed that our model captures people's capacity to adapt to heterogeneous environments where each problem is unique and may require a different strategy than the previous one. Previous theories were unable to account for this adaptive flexibility but our rational account of strategy selection does. When we consider all of these phenomena jointly, our findings support the view that people choose cognitive strategies rationally subject to the constraints imposed by their finite time, limited information, and bounded cognitive resources. Its rational cost-benefit analysis allows our model to capture that people allocate their time and cognitive resources strategically so as to maximize their expected reward rate across multiple decisions rather than just their immediate reward on the current problem.

4.7.1 IMPLICATIONS FOR THE DEBATE ABOUT HUMAN RATIONALITY

Our theory reconciles the two poles of the debate about human rationality by suggesting that people gradually learn to make increasingly more rational use of fallible heuristics. Our emphasis on metacognitive learning provides a fresh alternative to previous accounts that viewed rationality as a fixed, static ideal, and irrationality as a pervasive trait. Instead, our theory suggests that we are constantly learning to think, learn, and decide more resource-rationally with respect to the problems, rewards, and costs we experience. Hence, if we engage seriously with the environments we want to master, then metacognitive learning should propel us towards bounded rationality as we learn to choose the strategies that achieve the best possible cost-benefit tradeoff. Thus, although we might never reach the ideals of (bounded) rationality, we can become a little more resource-rational every time we use a cognitive strategy. Whether these improvements depend on deliberate reflection is an interesting question for future research.

The strategy selection problem is a critical missing piece in the puzzle of what it means to be boundedly rational. Our proposal for a rational solution to the strategy selection problem might therefore be an important step towards a unifying theory of bounded rationality. Indeed, recent work suggests that rationally choosing among a small number of cognitive strategies is optimal for bounded agents (Milli, Lieder, & Griffiths, 2017, 2018). Our model solves the riddle how a bounded agent can possibly optimize the use of its limited resources by investing some of them into solving the computationally intractable and potentially recursive problem of optimizing the use of its limited resources. We have proposed that the mind side-steps the computational complexity and infinite regress of this problem by *learning*–rather than computing–the value of investing time and cognitive resources into one strategy versus another. We show that good strategies can be selected very efficiently once an approximation to the value of computation has been learned and that the learning process can be implemented very efficiently as well (see Figure 4.1). Despite its simplicity this mechanism can adaptively choose between complex and extremely time- and resource-consuming strategies. It may thereby enable the mind to save a substantial amount of cognitive resources and find good approximate solutions to intractable problems. Our model can therefore be used to complete dual-process theories of bounded rationality (Evans, 2003; Evans & Stanovich, 2013; Kahneman, 2011) by a rational, yet tractable, mechanism for determining when to employ which system. Our strategy selection mechanism could be integrated into dual-process theories to predict exactly when people think fast and when they think slow. Likewise, our mechanism could also be integrated into adaptive toolbox theories of bounded rationality (Todd & Brighton, 2015; Todd & Gigerenzer, 2012) to predict exactly which heuristic people will use in a given situation. This line of research

would lead to mathematically precise, falsifiable theories of bounded rationality that could be quantitatively evaluated against empirical data and each other.

Our perspective emphasizes the importance of metacognitive values for human rationality. This emphasis is consistent with the view that individual differences in rationality reflect people's dispositions towards different cognitive styles ("the reflective mind") rather than their cognitive abilities per se (Stanovich, 2011, "the algorithmic mind"). Our theory suggests that the disposition towards rational versus heuristic thinking is not fixed and innate but malleable and learned from experience. Yet, our theory also suggests that a person's propensity for rational thinking can be highly situational because the mind estimates the value of deliberation from contextual features.

4.7.2   FUTURE DIRECTIONS

Future work should extend the proposed model to capture additional aspects of human cognition. One such extension could be a more realistic model of the cost of strategy execution which captures that some strategies are more effortful than others. This could be achieved by modeling how much cognitive resources, such as working memory, each strategy consumes at each point in time. With this extension, the total cost of executing a strategy could be derived by adding up the opportunity costs of its consumed resources over the time course of its execution.

While our model comparisons show that strategy selection learning requires some form of exploration, it is silent about how this exploration is accomplished. The Thompson sampling mechanism evaluated here is one of the best solutions to the exploration-exploitation tradeoff known to date (Chapelle & Li, 2011; Kaufmann, Korda, & Munos, 2012), but many alternative exploration mechanisms have been proposed in the reinforcement learning literature. These proposals range from simple mechanisms like epsilon-greedy action selection and the soft-max decision rule (Sutton & Barto, 1998) to more sophisticated mechanisms including upper-confidence bound algorithms (Auer, 2002) and other exploration bonuses (Brafman & Tennenholtz, 2002). At this point, each of these algorithms is a viable hypothesis about human strategy selection, and designing experiments to test them is an important direction for future research.

While our simulations and model comparisons favored learning separate predictive models of execution time and accuracy over learning the VOC directly, this advantage might reflect specific, auxiliary assumptions of our model. A more definitive answer will require experiments that systematically disambiguate these two learning mechanisms. Based on how model-free and model-based

control over behavior are usually disambiguated (Dickinson, 1985), strategy selection experiments that devaluate speed or accuracy (but not both) after people have learned to achieve the optimal speed-accuracy tradeoff might be a fruitful direction for future research.

Since our model is agnostic about the set of strategies people choose from, future work should determine which strategies are available to people. This could be done by comparing rational metareasoning models with different sets of strategies using Bayesian model selection (Scheibehenne, Rieskamp, & Wagenmakers, 2013).

People's decision mechanisms likely include strategies with continuous parameters, such as sequential sampling models with decision thresholds and attentional biases (P. L. Smith & Ratcliff, 2004), satisficing strategies with aspiration levels (Simon, 1955), and simulation-based decision mechanisms that can perform varying numbers of simulations (e.g., Lieder, Griffiths, & Hsu, 2017). Furthermore, the proposed process model only learns about a small subset of all possible cognitive strategies. To select among all possible sequences of elementary information processing operations, our process model has to be extended to learning the VOC of individual computations instead of only learning the VOC of complete strategies that always generate an action yielding reward. Current work is extending the proposed model to overcome these limitations (Krueger, Lieder, & Griffiths, 2017; Lieder, Krueger, & Griffiths, 2017; Lieder, Shenhav, Musslick, & Griffiths, 2018).

To capture people's ability to plan sequences of cognitive operations, future work might add predictive models for features of the agent's future internal states alongside the predictive models of the expected reward and execution time. This extension would correspond to learning option models (Sutton et al., 1999)–a form of model-based hierarchical reinforcement learning (Barto & Mahadevan, 2003; Sutton et al., 1999) that holds promise for explaining the complex hierarchical structure of human behavior (Botvinick & Weinstein, 2014). Both extensions could be combined with ideas from hierarchical reinforcement learning to capture how people discover novel, more effective strategies by flexibly combining elementary operations with partial.

The third major limitation of the current model is that it presupposes domain-specific problem features. A complete account of strategy selection would have to specify where those representations come from. To provide such an account, our model could be implemented as a hierarchical neural network with several layers in-between the perceptual input and the representation of the features as illustrated in Figure 4.1. In such a network the features could emerge from the same error-driven learning mechanism used to learn the weights between the feature layer and the layers representing the network's predictions (Mnih et al., 2015).

Future experiments might also investigate whether the proposed feature-based strategy-selection mechanism coexists with a more basic, automatic strategy selection mechanisms based on context-free RL. If so, then our framework could be used to model the arbitration between them as rational meta-strategy-selection.

One important open theoretical question is under which, if any, conditions the proposed strategy selection mechanism is boundedly optimal (Russell & Subramanian, 1995). While it is possible to prove the optimality of a program for a particular computational architecture, such proofs have yet to be attempted for computational models of the human mind.

5

People gradually learn to make increasingly more rational use of their cognitive resources[*]

Chapter 4 presented a model according to which people's ability to adaptively select between alternative cognitive strategies is acquired through metacognitive learning. This chapter presents two series of experiments that investigate the implications of the proposed learning mechanism. The first section tests the model's prediction that resource-rationality increases with learning. The second section investigates how the rate of this improvement depends on the reward structure of the environment and how such learning can be accelerated.

5.1   Rational strategy selection is learned from experience

According to the rational metareasoning model presented in Chapter 4, people acquire their capacity for adaptive strategy selection by learning an internal predictive model of each strategy's perfor-

---

[*]This chapter is based on Lieder and Griffiths (2017) and Krueger et al. (2017).

mance. This model predicts that people should gradually learn to perform more valuable computations and fewer computations whose costs outweigh their benefits. In other words, people should learn to make increasingly more rational use of their finite time and computational resources. This hypothesis makes four predictions:

1. People learn to perform fewer computations whose time cost outweighs the resulting gain in decision quality.

2. People learn to perform more computations whose expected gain in decision quality outweighs their time cost.

3. Ecological rationality increases with learning: people gradually learn to adapt their strategy choices to the structure of their environment.

4. Adaptive flexibility increases with learning: people learn to use different strategies for different kinds of problems.

The following four sections test each of these four predictions in turn.

5.1.1   Experiment 1: When people think too much they learn to think less

The goal of Experiment 1 was to test our model's prediction that people will learn to deliberate less and decide more quickly when they are placed in an environment where the cost of deliberation outweighs its benefits.

Methods

We recruited 100 adult participants on Amazon Mechanical Turk. Participants were paid $0.75 for 15 minutes of work and could earn a bonus of up to $2 for their performance on the task; the average bonus was $1.15 and its standard deviation was $0.73. The experiment was structured into three blocks: a pretest block, a training block, and a posttest block. Participants received feedback about the outcomes of their choices in the training block but not in pretest or the posttest block. Each block lasted four minutes, and the participants' task was to win as many points as they could.

Figure 5.1 shows a screenshot of an example trial in the pretest phase. In each trial, participants were shown a number of gambles. They could either choose one of the gambles or skip the decision and move on to the next trial without receiving a payoff. As soon as the participant responded the
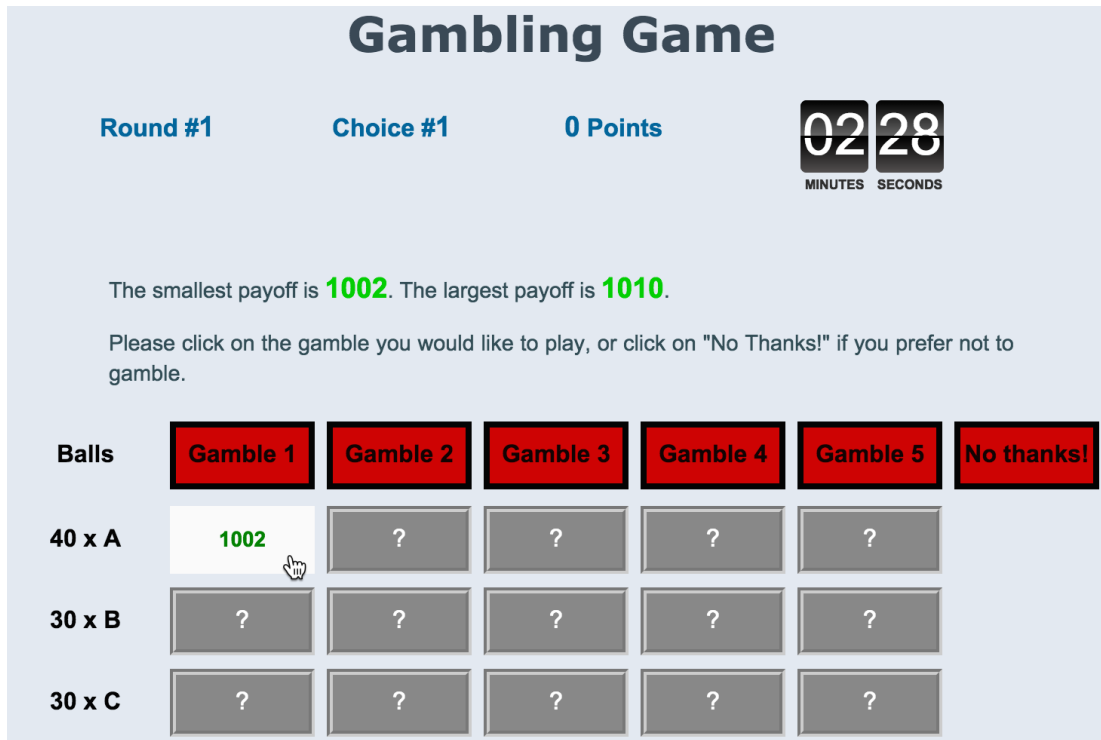
Figure 5.1: Screenshot of example trial in the pretest phase of Experiment 1.

next trial was shown. The number of trials was solely determined by how quickly the participant responded on each trial. On each trial, the decision problem was equally likely to belong to either of the four types summarized in Table 5.1. The four problem types differed primarily in the range of possible payoffs (low stakes, vs. high stakes, vs. all positive, vs. all negative), and on each trial this range was shown as a cue (see Figure 5.1). Critically, as shown in Table 5.1, the problem types and their frequencies were chosen such that the best approach was to skip trials where all outcomes were negative, choose randomly on trials where all outcomes were positive, and minimize the time spent on the high-stakes and the low-stakes problems by choosing randomly or skipping them altogether.

The number of outcomes was 3, 4, or 5 with probability 0.25, 0.50, and 0.25 respectively. The number of gambles was either 4 or 5 with equal probability. Given the number of outcomes and gambles, the payoffs were sampled uniformly from the problem type's range of payoffs given in Table 5.1. The outcome probabilities were sampled independently from the payoffs. Concretely, if there were k outcomes, then the first $k - 1$ outcome probabilities were sampled by a stick-breaking process where the relative length of each new stick was sampled from a uniform distribution. The

Table 5.1: Frequency and properties of the four types of decision problems used in Experiment 1.

| Problem Type	Frequency	Worst Outcome	Best Outcome	Optimal Strategy
All great	25%	990	1010	random choice
All bad	25%	-1010	-1000	Disengagement
High Stakes	25%	-1000	1000	Disengagement
Low Stakes	25%	-10	10	Disengagement

Note: All gambles were compensatory.

probability of the k-th outcome was set to 1 minus the sum of the first $k - 1$ probabilities.

MODEL PREDICTIONS.   To simulate people's choice of decision strategies and how it changes with learning, we combined our rational process model of strategy selection learning with the 10 decision strategies considered by Payne et al. (1988): the weighted-additive strategy, the equal weight strategy, satisficing, choosing at random, the majority of confirming dimensions strategy, the lexicographic heuristic (take-the-best), the semi-lexicographic heuristic, elimination-by-aspects, as well as two hybrid strategies that combine elimination-by-aspects with the weighted-additive strategy and the majority of confirming dimensions strategy respectively. Two additional strategies allowed the decision-maker to choose at random and skip the trial without deliberation respectively. The model's prior on the reward rate was a normal distribution with a mean of 1 point per second and a precision equivalent to 1 minute's worth of experience in the task. The priors on the regression coefficients and the error variance of the agent's predictive model of the strategies' performance were the same as in the simulations of the experiment by Payne et al. (1988). The features of the agent's predictive model combined those used to simulate the experiment by Payne et al. (1988) with four indicator variables signaling the presence or absence of the cues associated with the four types of gambles. Using these parameters, we ran 200 simulations of the experiment according to each model.

As shown in Figure 5.2A, our rational model predicted that participants should learn to decide more quickly and thereby win increasingly more points per second by engaging in deliberation less often and acquiring fewer pieces of information. Since the simulated decision-maker estimates its reward rate by Bayesian inference as defined above, it gradually realizes that its opportunity cost is very high. In addition, the simulated decision-maker learns that deliberate strategies are slow, and that the random strategy performs about as well as deliberation when all outcomes are similar. Hence,
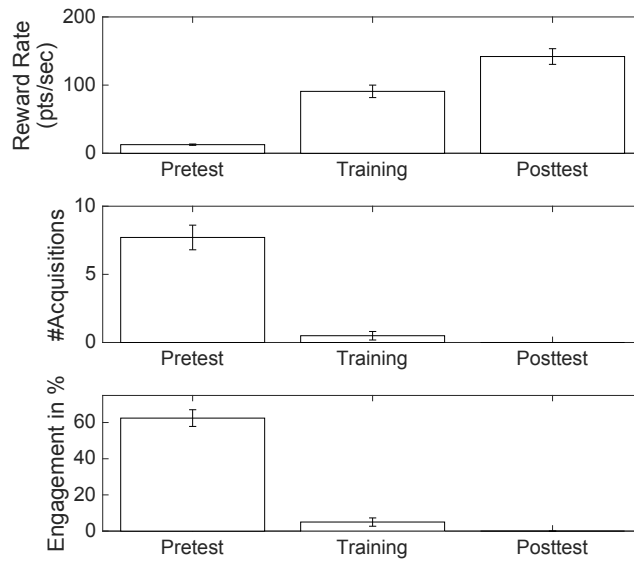
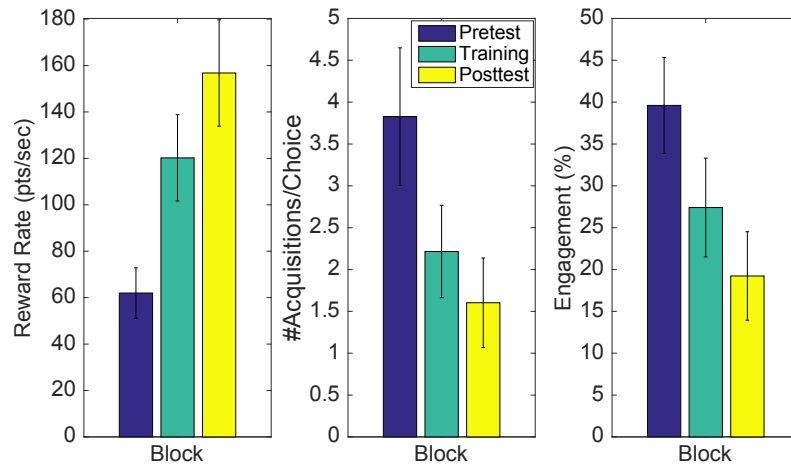**Figure 5.2:** Experiment 1: Learning when not to engage in effortful decision-making. A: Predictions of rational metareasoning for Experiment 1. B: The empirical findings of Experiment 1 confirmed the three qualitative model predictions.

the simulated decision-maker eventually learns to avoid deliberating, to skip problems with negative payoffs, and to apply the random strategy when all outcomes are great.

Results

To test our hypothesis that people learn to deliberate less, we classified the participants' response patterns into three categories: The response strategy on a trial was categorized as *random choice* if the participant chose one of the gambles without inspecting any of the outcomes. If the participant chose "No thanks!" without inspecting the outcomes, then the response strategy was classified as *disengaged*. Finally, if the participant clicked on at least one of the outcome boxes, then the response was categorized as *engaged*. We measured our participants' performance on the task by three metrics: *engagement*, *reward rate*, and *adaptive randomness*. Engagement was defined as the proportion of trials on which participants were engaged; the reward rate is the number of points earned per second; and adaptive randomness was measured by the frequency of random choice in problem type 1 (*all great*) minus the frequency of random choice on problems of types 2 (*all bad*) and 3 (*high stakes*); see Table 5.1. Our model predicted that participants' reward rate and adaptive randomness would increase significantly from the pretest to the posttest while their engagement decreases.

As shown in Figure 5.2B, we found that the learning induced changes in our participants' strategy choices were consistent with our theory's predictions. There was a significant increase in the participants' average reward rate ($t(99) = 9.98$, $p < 10^{-15}$; Cohen's $d = 1.00$) as they learned to process less information ($t(99) = -4.80$, $p < 10^{-5}$; Cohen's $d = -0.48$) and their engagement decreased significantly ($t(98) = -7.89$, $p < 10^{-11}$; Cohen's $d = -0.79$). Even though participants acquired increasingly less information, their average reward per decision did not change significantly from the first block to the last block ($t(98) = 0.69$, $p = 0.49$; Cohen's $d = 0.07$).

To examine whether the effect of learning on the number of computations performed by our participants depended on the problem type we ran a 2x2 mixed-effects, repeated-measures ANOVA with the average number of information acquisitions for a given problem type in a given block as the dependent variable and the problem type and the block number as independent variables. The main effect of the problem type was significant ($F(3, 1184) = 23.01$, $p < 10^{-13}$) suggesting that participants' information acquisition strategies differed significantly between the four types of decision problems (see Figure 5.3A): In high-stakes decisions, participants inspected 2.95 ± 0.55 outcomes on average, but on the trials where all outcomes were equally bad they inspected only about 0.5 potential payoffs (Cohen's $d = 1.96$). For low-stakes decisions and decisions in which all pos-

sible outcomes were great participants inspected an intermediate number of outcomes (about 1.5 inspected outcomes on average, Cohen's $d = 1.21$ and $d = 1.18$ respectively). The number of information acquisitions changed significantly across the three blocks of the experiment ($F(1, 1184) = 23.64$, $p < 10^{-5}$). Concretely, information acquisition decreased by 1.4 pieces of information per block ($t(1184) = -4.86$, $p < 10^{-6}$; Cohen's $d = -0.14$). There was a statistically significant interaction between problem type and block number ($F(3, 1184) = 4.74$, $p = 0.003$) indicating that the number of information acquisitions decreased more strongly for some problem types than for others. This decrease was statistically significant for problems in which all outcomes are great ($t(99) = -3.30$, $p < 0.001$, Cohen's $d = -0.33$), problems in which all outcomes are bad ($t(99) = -5.15$, $p < 10^{-6}$, Cohen's $d = -0.52$), and the high-stakes decision problems ($t(99) = -5.06$, $p < 10^{-6}$, Cohen's $d = -0.51$). But for the low-stakes problems the decrease was weaker and not statistically significant ($t(99) = -1.50$, $p = 0.07$, Cohen's $d = -0.15$).

The observed decrease in the number of information acquisitions was partly driven by a decrease in the frequency with which people engaged with the decision problems by inspecting at least one of their payoffs. As shown in Figure 5.3A, the proportion of decision problems in which people inspected at least one of the payoffs dropped from 37% in the pretest to 19% in the posttest. To test whether learning decreased the number of computations that people perform above and beyond the effect of disengagement, we repeated the analysis of variance described above for only those trials on which people engaged with the decision problem (see Figure 5.3B). We found that the main effect of the block number was still highly significant ($F(1, 659) = 8.08$, $p = 0.005$). The estimated decrease in information acquisition on trials on which people engaged with the decision problem was 1.1 pieces of information per block (95% CI: $[-1.80, -0.33]$, $t(659) = -2.84$, $p = 0.005$, Cohen's $d = -0.11$) and this value was not significantly different from the average decrease across all trials (1.4 acquisitions/block, 95% CI: $[-1.60, -0.68]$). There was also a significant interaction between the block number and problem type ($F(3, 659) = 2.61$, $p = 0.05$).

Furthermore, we found a significant increase in adaptive randomness ($t(97) = 7.21$, $p < 10^{-10}$, Cohen's $d = 0.73$). This means that our participants learned to selectively apply the random choice strategy to the *all great* problems (see Figure 5.3C). Consistent with this finding, the frequency of random choice increased on the *all great* trials ($t(97) = 6.61$, $p < 10^{-8}$, Cohen's $d = 0.67$) but decreased on all other trial types ($t(98) = -2.77$, $p = 0.003$, Cohen's $d = -0.28$).

Finally, we investigated whether people learn to prioritize the most probable outcome over less probable outcomes. To do so, we recorded the rank of the probability of the outcome participants inspected first and averaged it by block. The rank of the most probable outcome is one, the rank

of the second most probable outcome is two, etc. On average, people inspected the second most probable outcome first. This is consistent with the interpretation that our participants sometimes used strategies that prioritize the most probable outcomes and sometimes used strategies that do not. There was a very small and almost statistically significant decrease in the rank of the probability of the outcome inspected first from 2.33 ± 0.05 in the pretest to 2.15 ± 0.08 in the posttest ($t(59) = -1.67$, $p = 0.05$; Cohen's $d = 0.22$).

In summary, Experiment 1 placed participants in an environment where maximizing the reward rate required choosing without deliberation, and the participants learned to reap increasingly higher reward rates by acquiring increasingly fewer pieces of information, choosing at random when all outcomes were great and to skipping all other problems. There was also a trend towards learning to prioritize the most probable outcome. All of these effects are consistent with the hypothesis that people learn to make increasingly more rational use of their finite time and computational resources.

MODEL COMPARISONS.   While our findings were qualitatively consistent with the model predictions there were quantitative differences: People tended to outperform the model in terms of the reward rate in the pretest block, and their average number of acquisitions and frequency of engaging in deliberation changed less than predicted by rational metareasoning (compare Figure 5.2A vs. Figure 5.2B, and see the Appendix C for a more detailed comparison).

To evaluate our rational metareasoning model against the 14 alternative models described above, we ran 200 simulations of Experiment 1 according to each of the models. For each model, we performed six one-sample t-tests to determine whether it captured the increase in reward rate, the decrease in the number of acquisitions, and the decrease in the frequency of engagement from block 1 to block 2 and from block 2 to block 3, and one t-test to evaluate whether the model captured that people acquired more pieces of information on high-stakes problems than on other kinds of problems. We found that while our rational metareasoning model captured all of these effects, none of the SCADS, RELACS, or SSL models were able to capture all four effects simultaneously. The only component of the metareasoning model that was not necessary to capture human performance in Experiment 1 were the features. The reason why the lesioned metareasoning model without features could perform well is that the explicitly stated payoff ranges were sufficient for choosing strategies adaptively. Critically, none of the other lesioned metareasoning models were able to capture human performance. This suggests that all other components of our rational metareasoning model– choosing strategies based on the VOC, exploration, and learning separate predictive models of execution time and reward–are necessary to capture people's ability to adapt to the decision environment
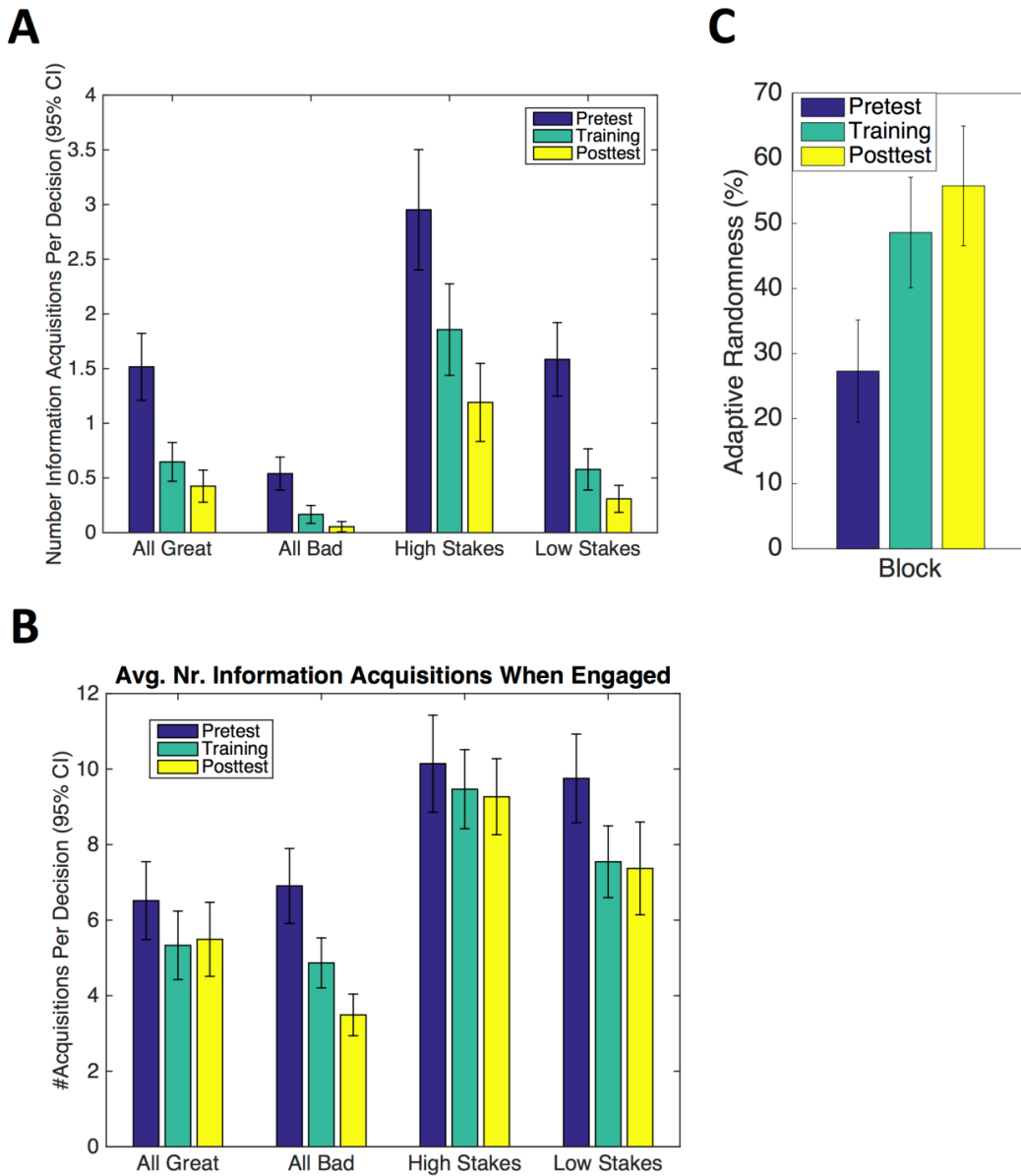
Figure 5.3: Adaptive disengagement in Experiment 1. A: Average number of information acquisitions by block and problem type. B: Number of information acquisitions when engaged. C: Adaptive randomness increased as participants learned to apply the random choice strategy more often to problems where all outcomes were great and less often to other problems.

of Experiment 1. For a more detailed summary of these simulation results, please see Appendix C.

Discussion

The observation that sometimes people are cognitive misers poses a challenge to most rational models, but our model predicted it correctly. According to our model, people become faster and less accurate at a challenging task when the difference between the rewards for good versus bad performance is small compared to how much time it would take to perform better. The observation that over time participants came to engage less with all four types of problems could also be interpreted as a general disengagement from the experiment rather than a rational adaptation to the structure of the decision environment. To disambiguate rational adaptation from disengagement we designed an additional experiment in which our theory predicts that people should learn to invest increasingly more time and effort.

5.1.2 Experiment 2: Learning to deliberate more

The goal of Experiment 2 was to test our model's prediction that people learn to deliberate more when they initially think too little. To create a situation where people think too little, we first put them in an environment whose reward rate was so high that deliberating on low-stakes problems was a waste of time and then changed the environment so that low-stakes problems became the only opportunity to earn money.

Method

We recruited 201 adult participants on Amazon Mechanical Turk. Participants were paid $0.75 for participation and could earn a performance-dependent bonus of up to $2. After performing the task participants completed an attention check that required them to estimate the highest possible payoffs of the different types of games they played in the experiment. Participants were excluded if they reported a positive number for the gamble that had only negative outcomes, if their estimate for the high-stakes gamble (±100) was less than twice their estimate for the low-stakes gamble (±10), or if any of their estimates was larger than 500. Based on these criteria, we had to exclude 57 participants (28.36%). In the experiment, participants visited a virtual casino that offered three different kinds of games: In *Blue Mountain Games* the stakes were high (±100). In *Purple*

Sun Games the stakes were low (± 10), and in *Orange Diamond* games all outcomes were negative ([$-100; -90$]). Each type of game was associated with a logo. The instructions informed participants that there were three kinds of games and what their payoffs were. In contrast to Experiment 1, the range of possible outcomes was not stated explicitly on every trial; instead they had to be inferred from the game's logo. Figure 5.4 shows a screenshot from Experiment 2.

Figure 5.4: Screenshot from Experiment 2.

The experiment was structured into five blocks lasting 2.5 minutes each. The first and the sec-

ond block had a high reward rate. They comprised 50% high-stakes problems and 50% low-stakes problems. Blocks 3-5 were the pretest, the training, and the posttest block respectively, and they all had the same structure. In each of these blocks, the first four trials were low-stakes problems (± 10 points) and the remaining trials comprised 75% trials with only negative outcomes ($[-100, -90]$) and 25% low-stakes problems. Hence, starting with the pretest block, low-stakes decisions became the only opportunity to win points and the opportunity cost for engaging in them became negligible. In all decision problems presented in Experiment 2, identifying the optimal choice required integrating multiple attributes. The number of gambles was always five, and the number of outcomes was always four. The payoffs were sampled uniformly from the range associated with the problem and their probabilities were determined as in Experiment 1. In contrast to Experiment 1, participants could *not* skip trials but always had to choose a gamble to advance.

We ran 200 simulations of this experiment using the same strategies and parameters as for Experiment 1 except that the agent did not have the option to skip trials. Rational metareasoning predicted that starting from the pretest (block 3), participants will learn to reap increasingly higher reward rates by engaging more often in the now worthwhile low-stakes problems and acquire increasingly more information to make those choices (see Figure 5.5).

Results and Discussion

First, we quantified learning by the change in our participants' average reward rate from the pretest to the posttest. The increase in people's average reward rate from -2.00 ± 0.24 in the pretest to -1.16 ± 0.23 in the posttest was statistically significant according to a one-sided t-test ($t(143) = 2.87$, $p = 0.002$; Cohen's $d = 0.26$). The reward rate depends on two factors: the reward per decision and the number of decisions per minute. To determine which of the two factors was responsible for the increase, we analyzed the learning induced changes in each factor separately. First, we analyzed how the reward per decision changed from the pretest to the posttest. For low-stakes problems the reward increased significantly from about 1.93 points per decision to about 2.42 points per decision ($t(143) = 1.92$, $p = 0.03$; Cohen's $d = 0.16$). By contrast, for problems on which all outcomes were negative the average reward did not change significantly (-14.43 vs. -14.22, $t(96) = 1.08$, $p = 0.14$; Cohen's $d = 0.11$). Next, we analyzed potential changes in the second factor: the number of decisions per unit time. We found that participants slowed down significantly from 12.86 ± 1.42 decisions per minute in the pretest to 8.32 ± 1.45 decisions per minute in the posttest ($t(143) = 2.99$, $p = 0.003$; Cohen's $d = 0.25$). Hence,
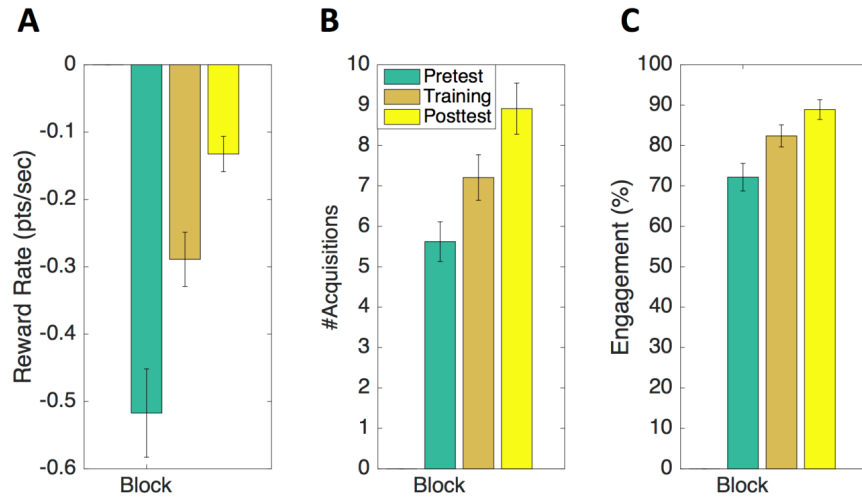
Figure 5.5: Rational metareasoning predictions of strategy selection learning in Experiment 2. A: Rational metareasoning predicted a significant increase in the reward rate from the pretest (block 3) to the posttest (block 5). B: Rational metareasoning predicted a significant increase in the number of information acquisitions on the worthwhile low-stakes problems. C: Rational metareasoning predicted a significant increase in people's engagement with the worthwhile low-stakes problems.

participants learned to reap a higher reward rate by deliberating more to make better decisions.

To test the hypothesis that deliberation increased with learning more rigorously, we analyzed the number of information acquisitions as a proxy for the number of computations performed by our participants. Concretely, we tested our model's prediction that people should learn to invest more computation into low-stakes decisions. As shown in Figure 5.6A, participants learned to allocate their time adaptively. Starting from the pretest (block 3)–where low-stakes problems became worthwhile solving–there was a significant increase in the number of information acquisitions on the low-stakes problems from 4.97 ± 0.34 to 6.42 ± 0.43 ($t(2798) = 5.19$, $p < 0.001$; Cohen's $d = 0.10$). This increase was specific to the low-stakes problems: It did not occur for problems with only negative outcomes. To the contrary, on problems with only negative outcomes the number of information acquisitions decreased from 2.95 ± 0.24 to 2.51 ± 0.26 ($t(5112) = -2.38$, $p = 0.02$; Cohen's $d = 0.03$). This suggests that people learned to allocate their computation more adaptively from the pretest to the posttest. The number of information acquisitions was particularly high on the first four trials of the three last blocks: the number of information acquisitions increased from 8.19 ± 0.51 in the pretest to 9.27 ± 0.50 in the posttest ($t(143) = 2.14$, $p = 0.02$; Cohen's $d = 0.18$).
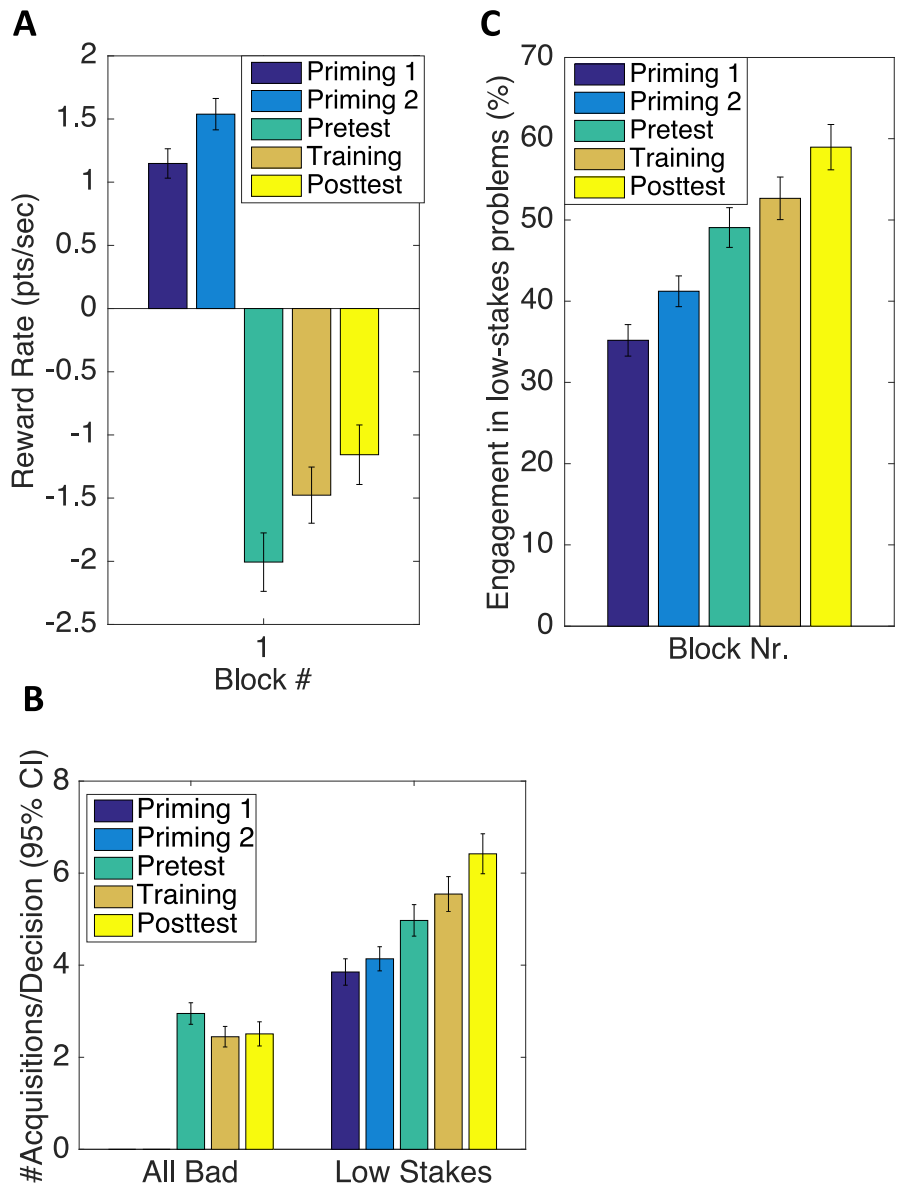
Figure 5.6: Empirical results of Experiment 2. A: Reward rate by block. Error bars denote plus and minus one SEM. B: Average number of information acquisitions. C: Engagement in low-stakes decisions.

The observed increase in the number of information acquisitions on low-stakes problems might be caused by an increase in the frequency with which people engaged with them, an increase in the number of computations they invested into solving those they engaged with, or both. We found that the increase in the number of information acquisitions per problem was mostly driven by an increase in the frequency with which people engaged in effortful decision making on low-stakes problems (see Figure 5.6B): the frequency of engagement in low-stakes problems increased from only $49.1 \pm 0.2\%$ in the pretest to $58.95 \pm 2.8\%$ in the posttest ($\chi^2(2) = 26.9$, $p < 0.001$; Cohen's $w = 5.19$). This increase was accompanied by a decrease in the frequency with which participants chose randomly which was the only way to avoid engaging with the problem. Importantly, we also found that the number of inspected outcomes increased even on the low-stakes problems that participants engaged with (10.14 in the pretest versus 10.89 acquisitions in posttest, $t(1490) = 2.07$, $p = 0.04$; Cohen's $d = 0.05$). On the problems with only negative outcomes, by contrast, there was a significant decrease in the number of information acquisitions ($t(1291) = -8.52$, $p < 0.001$; Cohen's $d = -0.24$). In conclusion, the increase in the number of information acquisitions on low-stakes problems was driven by both factors: our participants learned to engage in low-stakes decisions more frequently and to deliberate more when engaged. Both changes are consistent with learning to become more resource-rational. Finally, we also found that people gradually learn to prioritize the most probable outcomes. The average rank of the outcome that participants inspected first significantly decreased from 2.36 ± 0.07 in the first block to 2.18 ± 0.11 in the last block ($t(133) = 3.96$, $p < 0.001$; Cohen's $d = 0.34$). This learning process occurred even though identifying the optimal decision always required inspecting multiple outcomes.

As for Experiment 1, the predictions of our rational model were qualitatively correct, but the observed learning effects were slightly smaller than expected. The model achieved a slightly higher reward rate than people (cf. Figure 5.5A vs. Figure 5.6A), acquired about 0.5–3 additional pieces of information (cf. Figure 5.5B vs. Figure 5.6B), and engaged in 20%–30% more problems than people (cf. Figure 5.5C vs. Figure 5.6C). In summary, we found that people learn to deliberate more and gather more information when the reward structure of their environment calls for it. This result complements the finding from Experiment 1 where people learned to invest less computation because the return on investing deliberation was less than their opportunity cost. In conclusion, our results suggest that strategy selection learning makes people more resource-rational by tuning strategy choices towards the optimal speed-accuracy tradeoff.

Model Comparisons.   For the purpose of model comparison, we ran 200 simulations of Experiment 2 according to each of the 14 alternative models described above. For each model, we performed six one-sample t-tests to determine whether it correctly predicted the increases in reward rate, information acquisitions, and the frequency of engagement that occurred from block 3 to block 4 and from block 4 to block 5, as well as one t-test to evaluate whether the model captured that people gathered more information on high-stakes problems than on other kinds of problems. We found that while our rational metareasoning model captured all of these effects, none of the SCADS, RELACS, or SSL models was able to capture these four effects simultaneously. Among the lesioned metareasoning models, only the one approximating the VOC by model-free reinforcement learning from the difference between reward and time cost captured all four phenomena. Critically, none of the other lesioned metareasoning models were able to do so. This suggests that choosing strategies based on the VOC, exploration, and feature-based learning are necessary to capture the adaptive strategy selection learning our participants demonstrated in Experiment 2. Hence, only the full rational metareasoning model can capture the findings from Experiments 2 and 3 simultaneously. For a more detailed summary of these simulation results, please see Appendix C.

According to our rational theory of strategy selection, the reason why some people are cognitive misers in certain tasks (Toplak, West, & Stanovich, 2013) is that their metacognitive model predicts that the reward for normative performance is just not worth the effort it would require. The results of Experiment 2 suggest that cognitive misers will often learn to deliberate more when the returns of deliberation justify its cost.

5.1.3 Ecological rationality increases with learning

The third prediction of our model is that people adapt their strategy choices to the structure of their environment. To evaluate this prediction, we examined a concrete example where people can use two different strategies to choose between two options with multiple attributes[†] the comprehensive Weighted-Additive-Strategy (WADD) versus the fast-and-frugal heuristic Take-The-Best (TTB). There are different variants of the WADD strategy. Since we will be modeling a multi-attribute binary choice task, we use the version of WADD that sums up the weighted differences between the first option's rating and the second option's rating across all attributes (Tversky, 1969). For each attribute this strategy compares the two ratings (1 operation). If the attribute values disagree, then it reads and adds or subtracts the attribute's validity (2 operations). Finally, it chooses the first at-

---

[†]A preliminary version of these simulations appeared in (Lieder & Griffiths, 2015).

tribute if the sum is positive or the second attribute if the sum is negative (1 operation). TTB is the equivalent of the lexicographic heuristic for multi-attribute decisions: it chooses the option that is best on the most predictive attribute that distinguishes between the options and ignores all other attributes. Our implementation of Take-The-Best first searches for the most predictive attribute by sequentially reading the validities of unused attributes (1 operation per attribute), comparing them to the highest validity found so far (1 operation per attribute), and memorizes the new validity if it exceeds the previous maximum (1 operation). Once the most predictive attribute has been identified, TTB compares the options' ratings on that attribute (1 operation), and then either makes a choice (1 operation), or continues with the next most predictive attribute.

TTB works best when the attributes' predictive validities fall off so quickly that the recommendation of the most predictive attribute cannot be overturned by rationally incorporating additional attributes; environments with this property are called *non-compensatory*. By contrast, TTB can fail miserably when no single attribute reliably identifies the best choice by itself; and environments with this property are called *compensatory*. Thus, to adapt rationally to the structure of their environment, that is to be *ecologically rational*, people should select TTB in non-compensatory environments and avoid it in compensatory environments.

Bröder (2003) found that people use TTB more frequently when their decision environment is non-compensatory. Rieskamp and Otto (2006) found that this adaptation might result from reinforcement learning. In their experiment participants made 168 multi-attribute decisions with feedback. In the first condition, all decision problems were compensatory, whereas in the second condition all decision problems were non-compensatory. To measure people's strategy use over time, Rieskamp and Otto (2006) analyzed their participants' choices on trials where TTB and WADD made opposite decisions. Participants in the non-compensatory environment learned to choose in accordance with TTB increasingly *more often*, whereas participants in the compensatory environment learned to do so increasingly *less* often.

These findings raise the question of how people learn when to use TTB. One hypothesis is that people learn how well TTB works *on average*, as postulated by the SSL model (Rieskamp & Otto, 2006). Our alternative hypothesis is that people learn to predict how fast and how accurate TTB and alternative strategies will be on *individual problems* based on problem features, as postulated by rational metareasoning. To test these two hypotheses against each other, we simulated Experiment 1 from Rieskamp and Otto (2006) according to rational metareasoning and SSL and compared how well the models' predictions explained the data. The experiment was divided into seven blocks. Each block comprised 24 trials, and each trial presented a choice between two investment options with

five binary attributes. The attributes' predictive validities were constant and explicitly stated. Both models assumed that participants in this experiment always choose between Take-The-Best ($s_1 =$ TTB) and the weighted-additive strategy ($s_2 =$ WADD). Our rational metareasoning model of this paradigm assumed that strategy selection in binary multi-attribute decisions relies on three features $\mathbf{f} = (f_1, f_2, f_3)$: the predictive validity of the most reliable attribute that discriminates between the two options (f_1), the gap between the validity of the most reliable attribute favoring the first option and the most reliable attribute favoring the second option (f_2), and the absolute difference between the number of attributes favoring the first option and the second option respectively (f_3). Our model assumes that people first inspect the validities of all cues and extract the three features $f_1, \ldots, f_3$ from them, then select a strategy based on these features, and finally execute that strategy to reach a decision.[‡]

The probability that a strategy s makes the correct decision ($R = 1$) was modeled by

$$P(R = 1 \mid s, \mathbf{f}) = \frac{1}{1 + \exp(-(b_s + \sum_i w_{s,i} \cdot f_i))}.$$

We modeled people's knowledge about the feature weights w_s by the prior distribution

$$P(w_s) = \mathcal{N}\left(\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \sigma^{-1} = \tau \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right),$$

$$P(b_s), = \mathcal{N}(\mu = b_s^{(0)}, \sigma^{-2} = \tau),$$

where the expected value of the offset b_s (i.e., $b_s^{(s)}$) and the strength τ of the prior belief are free parameters.

To simulate the first experiment from Rieskamp and Otto (2006), we created a compensatory environment and a non-compensatory environment. In the compensatory environment, WADD always makes the Bayes-optimal decision and TTB disagrees half of the time. Conversely, in the non-compensatory environment TTB always makes the Bayes-optimal decision and WADD disagrees half of the time. To determine the optimal choices, we computed the probability that option A is

---

[‡]This entails that all cue validities are inspected on all trials even when a fast-and-frugal heuristic like TTB is chosen. This makes the number of information acquisitions on trials where TTB is used more similar to the number of information acquisitions on trials where WADD is used. This diminishes the relative number of information acquisitions saved by TTB. However, this does *not* affect the number of inspected cue values.

superior to option B given their ratings by Bayesian inference under the assumption that positive and negative ratings are equally common. First, we randomly generated a set of candidate decision problems. For each of these problems, we computed the posterior probability that the first option is superior to the second option given their attributes' values and their validities. We then used these posterior probabilities to select which candidate decision problems to present in the compensatory environment and which to present in the non-compensatory environment. To match the reward probabilities of Experiment 1 by Rieskamp and Otto (2006), the feedback was determined solely based on the environment and the chosen strategy: the probability of positive feedback was 92% whenever the strategy matched the structure of the environment (e.g., WADD in the compensatory environment) and only 58% when it did not (e.g., TTB in the compensatory environment). Positive feedback meant winning \$0.15 whereas negative feedback meant losing \$0.15.

To simulate the experiment, we let our rational metareasoning models learn the agent's opportunity cost from experience; the prior mean of the opportunity cost was initialized with \$7/h and the prior precision corresponded to one minute's worth of experience. For simplicity, we assumed that people perform one step of TTB or WADD per second. To estimate which strategy people considered more effective a priori, we set the prior expectation of the problem-independent performance of TTB (b_{TTB}^0) to zero and fit the model's prior expectation of the problem independent performance of WADD $(b_{WADD}^{(0)})$ and the strength of the agent's prior beliefs about the strategies' performance and execution time (τ) to the data. Specifically, we determined these parameters by maximum-likelihood estimation from the frequencies with which Rieskamp and Otto's participants used TTB in each block using grid search. The likelihood function was estimated by running at least 10 simulations of the experiment for each point on the grid of potential parameter values. Rieskamp and Otto (2006) estimated that participants made accidental errors in about 5% of the trials. To capture these errors and avoid numerical problems, we modelled people's apparent strategy choice frequencies by

$$\hat{\theta}_{\text{strategy}}^{(b)} = \frac{0.9 \cdot n_{\text{strategy}}^{(b)} + 0.1 \cdot 0.5 \cdot n_{\text{total}}^{(b)}}{n_{\text{total}}^{(b)}} \tag{5.1}$$

where strategy is a placeholder for either TTB or WADD and $n_{\text{total}}^{(b)} = n_{\text{TTB}}^{(b)} + n_{\text{WADD}}^{(b)}$ is the total number of trials in block b.[§]

[§]This assumption is not a model of the underlying psychological processes. Instead, it serves as a placeholder for all unknown and known influences on strategy selection that the model does not capture. The frequency of trials in which the strategy is chosen at random was selected so as to generate 5% of trials in which the chosen strategy disagrees with the one prescribed by the model. We assumed random choice because it is the weakest assumption we could make.

The resulting parameter estimates captured that people initially preferred WADD to TTB ($\hat{b}_{\text{WADD}}^{(0)} = +0.32$) and required many decisions' worth of experience to revise their beliefs ($\hat{\tau} = 88.59$). Our simulation showed that rational metareasoning can explain people's ability to adapt their strategy choices to the structure of their environment (see Figure 5.7): When the decision environment was non-compensatory, then our model learned to use TTB and avoid WADD. But when the decision environment was non-compensatory, then our model learned to use WADD and avoid TTB. In addition, rational metareasoning captured that people adapt their strategy choices gradually.

We also estimated the parameters of the SSL model and the three SCADS models introduced above. The SCADS models were equipped with two categories for compensatory versus non-compensatory problems respectively. The free parameters of the SCADS models determined the initial associations between each category and the two strategies. The first parameter was the sum of the two strategies' association strengths, and the second parameter was the relative strength of the association with the WADD strategy. The global association strengths were the sums of the category-specific associations. For the SSL model, we estimated the relative reward expectancy of the WADD strategy (β_{WADD}) and the strength of the initial reward expectancy (w) by the simulation-based maximum-likelihood method described above (Equation 5.1).

The maximum-likelihood estimates of the SSL model's parameters were $\beta_{\text{WADD}} = 0.35$ and $\hat{w} = 30$. The mean squared error of the fit achieved by the SSL model was about half the MSE of the rational metareasoning model (0.0018 vs. 0.0043); see Figure 5.7. Consequently, the Bayesian information criterion provided strong evidence for the SSL model over the full rational metareasoning model (Kass & Raftery, 1995, $\text{BIC}_{\text{SSL}} = 60.70$ vs. $\text{BIC}_{\text{RM}} = 68.94$;) and all lesioned metareasoning models ($\text{BIC} \geq 70.52$). The BIC of the full rational metareasoning model was slightly higher than the BIC for the lesioned metareasoning model without features ($\text{BIC} = 70.52$), and the data provided strong or very strong evidence for the full metareasoning model over all other lesioned metareasoning models ($\text{BIC} \geq 75.94$). The fit of the SCADS models was comparable to the fit of the SSL model and significantly better than the fits of the metareasoning models ($\text{BIC}_{\text{SCADS1}} = 61.09$, $\text{BIC}_{\text{SCADS2}} = 62.38$, and $\text{BIC}_{\text{SCADS3}} = 62.01$). Finally, we repeated our model comparison for both environments separately. Consistent with the original model comparison results, we found that SSL provided a better explanation for the data from the compensatory environment ($\text{BIC}_{\text{SSL}} = 32.64$ vs. $\text{BIC}_{\text{RM}} = 34.29$) and the data from noncompensatory environment ($\text{BIC}_{\text{SSL}} = 36.22$ vs. $\text{BIC}_{\text{RM}} = 37.81$) than the rational metareasoning models. The performance of the SCADS models was close to the performance of the SSL models ($\text{BIC}_{\text{SCADS}} = 32.72$ for the compensatory environment and $\text{BIC}_{\text{SCADS}} = 36.35$ for the noncompensatory environ-
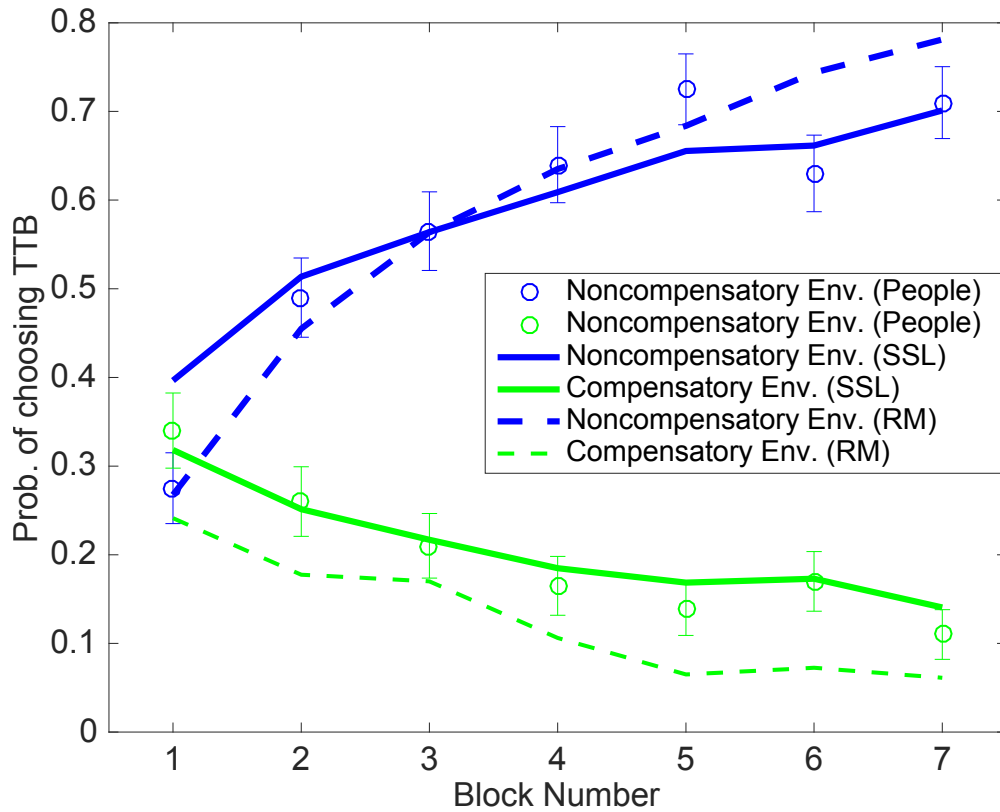
Figure 5.7: Fit of rational metareasoning model and SSL to the empirical data by (Rieskamp & Otto, 2006).

ment).

The quantitative differences between the model fits should be taken with a grain of salt because they depend on the auxiliary assumption that people use the exact TTB and WADD strategies available to the models and no other strategies. This assumption is questionable for at least two reasons: First, TTB and WADD are merely placeholders for the class of non-compensatory strategies and the class of compensatory strategies respectively (Rieskamp & Otto, 2006). Second, previous work suggests that the human mind is equipped with a much larger repertoire of decision strategies (Payne et al., 1988). If the rational metareasoning model was also equipped with a larger repertoire of strategies, then it would learn more gradually and probably achieve a better fit to the human data. Due to these caveats, we focus on the models' qualitative predictions because they are less sensitive to different auxiliary assumptions.

The feature-based learning mechanism of the SCADS model and the context-free learning mechanism of the SSL model captured the human data equally well ($\mathrm{BIC_{SSL}} - \mathrm{BIC_{SCADS1}} = 0.04 \ll 2$), and the feature-based learning mechanism of the rational metareasoning model also captured the qualitative changes in people's strategy choices. Since the data by Rieskamp and Otto (2006) can be explained by either feature-based or context-free strategy selection learning, we designed a new experiment to determine which mechanism is responsible for people's adaptive strategy choices.

5.1.4   Experiment 3: Adaptive flexibility increases with learning

The fourth prediction of our model is that people learn to flexibly switch their strategies on a trial-by-trial basis to exploit the structure of individual problems. An alternative hypothesis embodied by SSL and RELACS is that strategy selection learning serves to identify the one strategy that works best on average across all problems in a given environment. To design an experiment that can discriminate these hypotheses, we evaluated the performance of context-free versus feature-based strategy selection learning in 11 environments with 0%, 10%, 20%, . . . ,100% compensatory problems and 100%, 90%, 80%, . . . ,0% non-compensatory problems respectively. Critically, all compensatory problems were designed such that TTB fails to choose the better option and WADD succeeds, and all non-compensatory problems were designed such that TTB succeeds and WADD fails. For each of the 11 decision environments, we compared the average performance predicted by rational metareasoning with the parameters $b_{\mathrm{WADD}}^{(0)} = 0$ and $\tau = 1$, against the predictions of the five lesioned metareasoning models with the same parameters, SSL with parameters $\beta_1 = \beta_2 = 0.5$ and $w = 1$, RELACS with parameters $\alpha = 0.1$ and $\lambda = 1$, and the three SCADS models with an association strength of 0.5 between each strategy and two categories corresponding to compensatory and non-compensatory problems respectively.¶

Our simulations revealed that feature-based and context-free strategy selection learning predict qualitatively different effects of the relative frequency of compensatory versus non-compensatory decision problems; see Figure 5.9A. Concretely, the performance of model-free strategy selection learning drops rapidly as the decision environment becomes more heterogeneous: As the ratio of compensatory to non-compensatory problems approaches 50/50 the performance of context-free strategy selection learning (SSL, RELACS, and the lesioned metareasoning model without features)

¶These parameters were chosen to give each model a weak, initial bias towards using both strategies equally often. The exact value of this bias is not critical because it is quickly overwritten by experience.

and SCADS‖ drops to the level of chance. By contrast, the performance of feature-based strategy selection learning (rational metareasoning) is much less susceptible to this heterogeneity and stays above 60%. The reason is that rational metareasoning learns to use TTB for non-compensatory problems and WADD for compensatory problems, whereas SSL and RELACS learn to always use the same strategy. We can therefore determine whether people rely on context-free or feature-based strategy selection with the following experiments that puts participants in a heterogeneous environment.

Investment Decision 1/30

Please determine which of the two unnamed companies should be given the loan. There is only one correct answer. If you are right you earn $50,000, else you lose $50,000.

Criteria	Probability of Success	Rating of A	Rating of B
Efficiency	85%	-	+
Qualifications of Employees	60%	+	-
Capital Structure	78%	-	-
Management	75%	+	-
Own Financial Resources	70%	+	-
Financial Flexibility	90%	-	+
I invest in ...		A	B

Outcome:	**Wrong! + $0**
Handling Fee:	**- $ 50,000**
Balance:	**- $50000**

Next

Figure 5.8: Interface of Experiment 3: Strategy selection in multi-attribute decision-making.

‖The problem preventing the SCADS model from choosing the best strategy for each category is that the category-specific association strengths are multiplied by a category-unspecific association strength.

Methods

We recruited 100 participants on Amazon Mechanical Turk. The experiment lasted about 25-30 min, and participants were paid $1.25 plus a performance-dependent bonus of up to $1.25. The experiment instantiated the decision environment with 50% compensatory problems and 50% non-compensatory problems from the simulations above. Participants played a banker deciding between giving a loan to company A versus company B based on their ratings on multiple attributes with explicitly stated predictive validities (see Figure 5.8). There were 12 attributes in total. Half of these attributes were reliable (predictive validity $\geq 85\%$) whereas the other half was unreliable (predictive validity $\leq 63\%$). Concretely, the attributes *Financial Flexibility*, *Efficiency*, *Capital Structure*, *Management*, *Own Financial Resources*, and *Qualifications of Employees* had predictive validities of 95%, 93%, 90%, 87%, 85%, and 83% respectively, whereas the attributes *Investment Policy*, *Business History*, *Real Estate*, *Industry*, *Reputation*, and *Location* had predictive validities of 63%, 60%, 57%, 55%, 53%, and 51% respectively. Each trial presented either 3, 4, 5, or 6 attributes with equal probability. On non-compensatory trials, exactly one of the attributes was reliable and all other attributes were unreliable. By contrast, on compensatory trials all attributes were reliable or all attributes were unreliable. Reliable and unreliable attributes were selected randomly and their order was randomized. The two options always had opposite ratings on the most predictive attribute, and 75% of the ratings on other attributes were opposite to the rating on the most predictive attribute while 25% agreed with it. After choosing Company A or Company B, participants received stochastic binary feedback: $\$ + 50,000$ versus $\$ - 50,000$. On compensatory trials, the probability of positive feedback was 95% when the participant's choice agreed with the choice of WADD and 5% when it disagreed with WADD. On non-compensatory trials the probability of positive feedback was 95% when their choice agreed with TTB and 5% otherwise.

Each participant made 100 binary choices, earning a bonus of 1.25 cents for each correct decision and losing 1.25 cents for each incorrect decision. Critically, the ratio of compensatory to non-compensatory problems was 50/50: The problems were chosen such that TTB and WADD make opposite decisions on every trial. In half of the trials, the decision of TTB was correct and in the other half WADD was correct. Therefore, always using TTB, always using WADD, choosing one of these two strategies at random, or context-free strategy selection would perform at chance level; see Figure 5.9A.

Results and Discussion

To determine the quality of people's strategy choices, we compared their decisions on each trial to those of the strategy appropriate for the problem presented on that trial. For compensatory trials, we evaluated people's choices against those of WADD and for non-compensatory trials we evaluated them against TTB. People's decisions agreed with those of the appropriate strategy on 76.2% of the trials (see Figure 5.9A). To quantify our uncertainty about this estimate, we computed its credible interval assuming a uniform prior (Edwards et al., 1963). We found that the 99% highest-posterior density interval ranged from 75.1% to 77.3%. We can thus be 99% confident that people's average performance in the mixed decision environment was at least 75% and conclude that they performed significantly better than chance ($p < 0.001$, Cohen's $w = 52.36$). As shown in Figure 5.9B, people's performance increased significantly from 70.4% in the first ten trials to 80.4% in the last ten trials ($\xi^1(1) = 26.96, p < 0.001$, Cohen's $w = 5.19$). To gain a better understanding of this effect, we performed a logistic regression of the agreement between people's choices and those of the appropriate strategy; the regressors were the trial number, a constant, and the decision's compensatoriness. We found that people's performance increased significantly over trials ($t(9996) = 9.46, p < 0.001$). Consistent with the finding that people initially prefer compensatory strategies (Rieskamp & Otto, 2006), people performed better on compensatory trials than on non-compensatory trials overall ($t(9996) = 9.46, p < 0.001$) and this effect dissipated over time ($t(9996) = -7.20, p < 0.001$). Analyzing compensatory and non-compensatory trials separately with logistic regression revealed that our participants' performance on non-compensatory trials improved significantly over time ($t(4998) = 9.46, p < 0.001$) while their performance on compensatory trials remained constant ($t(4998) = -0.92, p = 0.36$). Interestingly, people performed significantly above chance already on the first trial (73% correct; $p < 0.001$). This suggests that people either entered the experiment with applicable expertise in when to use compensatory versus non-compensatory decision strategies, as suggested by the results of Payne et al. (1988) or possess general purpose strategies that work well on both kinds of problems. Both factors might also explain why people performed systematically better than all of our models (Figure 5.9A).

This level of performance could not have been achieved by context-free strategy selection, which performed at chance, but it is qualitatively consistent with feature-based strategy selection which performed significantly better than chance; see Figure 5.9A. We also simulated the experiment with three SCADS models that were equipped with two categories corresponding to compensatory versus non-compensatory problems and differed in their reward function ($r = $ correctness, vs. $r = $ correctness $-$ cost, vs. $r = $ correctness/time). We found that the performance of the SCADS

models was very similar to the performance of the SSL model. Most importantly, its performance dropped to the chance level as the environment became increasingly more heterogeneous. This happened because the global association strength interfered with category-specific strategy choices. Additional simulations with the five lesioned metareasoning models revealed that feature-based learning was indispensable to capture human performance. For more information, see Figure C.5 in Appendix C.

These results should be taken with a grain of salt because the model comparisons presuppose that TTB and WADD are the only decision strategies that people are equipped with even though people's repertoire most likely includes many additional strategies. It is conceivable that participants succeeded in Experiment 3 by relying on a single strategy that succeeds on both compensatory and non-compensatory problems. Because of this possibility, Experiment 3 does not provide definite evidence for feature-based strategy selection. However, Experiment 1, Experiment 2, and the simulations of mental arithmetic presented in the following sections also support feature-based strategy selection. Taken together these experiments and simulations provide very strong evidence for feature-based strategy selection learning.

5.1.5 Conclusion

The experiments presented in this section confirmed the predictions of our resource-rational theory of strategy selection learning: The first experiment showed that people learn to think less when they think too much. The second experiment showed that people learn to think more when they think too little. Thirdly, we showed that people learn to adapt not only how much they think but also *how* they think to the structure of the environment. Finally, Experiment 3 demonstrated that adaptive flexibility also increases with learning, and this enables people's strategy choices to exploit the structure of individual problems. Most importantly, in all four cases, the underlying learning mechanisms made people's strategy choices increasingly more resource-rational. Hence, the empirical evidence presented in this section supports our hypothesis that the human brain is equipped with learning mechanisms that make it more resource-rational over time. Even though people may not be resource-rational when they first enter a new environment, the way in which they process information appears to converge to the rational use of their finite time and bounded computational resources. This perspective replaces the static view that people are either rational or irrational with a dynamic view according to which people can become more rational over time. According to this dynamic view human rationality should be measured by people's ability to improve their reasoning

A

B
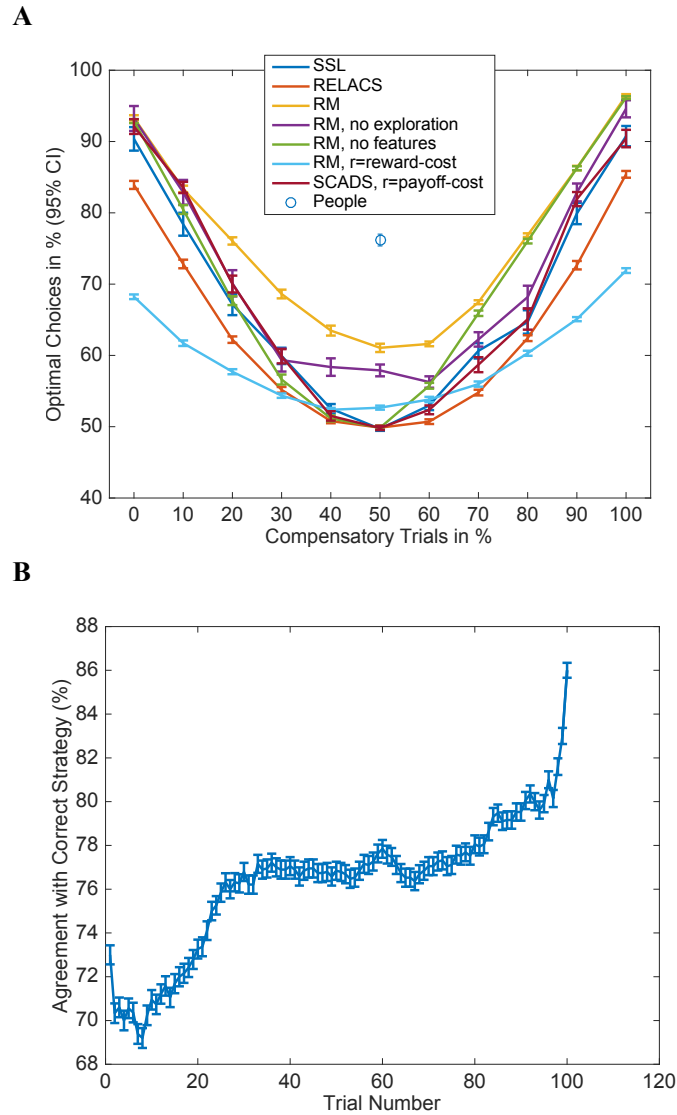
Figure 5.9: Model predictions and findings of Experiment 3. A: People and rational metareasoning perform significantly above chance in heterogeneous environments but context-free strategy selection mechanisms do not. B: People's performance increased with experience. The trial-by-trial frequencies were smoothed by a moving average over 20 trials. The error bars enclose 95% confidence intervals.

and decision-making based on their experience.

In addition to its contributions to the debate about human rationality and its utility for future

basic research, our model of strategy selection learning might also have potential practical applications in education and cognitive training. In terms of education, our model could be used to optimize the problem sets used to teach students when to use which approach–for instance in mathematical problem solving or high school algebra. In terms of cognitive training, our model could be used to investigate which training regimens increase cognitive flexibility by promoting adaptive strategy selection. According to our theory, people's ability to (learn to) represent problems by general features that are predictive of the differential efficacy of alternative strategies would be a critical prerequisite for such training to succeed.

In conclusion, our findings paint an optimistic picture of the human mind by highlighting metacognitive learning and the resulting cognitive growth. This perspective highlights that our rationality is not fixed but malleable and constantly improving. We hope that specifying what people's metacognitive learning mechanisms might be, our model will give us a handle on how to leverage them to promote cognitive growth.

5.2 ENHANCING METACOGNITIVE REINFORCEMENT LEARNING WITH REWARD STRUCTURES AND FEEDBACK[**]

The results presented in the previous section suggest that our metacognitive reinforcement learning model of strategy selection learning captures at least one of the mechanisms through which people can learn to make better decisions. This section explores the potential implications of this theory for brain training (Anguera et al., 2013; Bavelier, Green, Pouget, & Schrater, 2012; C. S. Green & Bavelier, 2008; Morrison & Chein, 2011; Owen et al., 2010) and promoting cognitive growth more generally. Concretely, if cognitive plasticity is driven by metacognitive reinforcement learning then it might be possible to leverage methods for accelerating reinforcement learning in robots (Ng et al., 1999) to design feedback structures for cognitive training in humans.

This section evaluates this approach in the domain of planning. As a first step, we developed a metacognitive reinforcement learning model of how people learn how many steps to plan ahead in sequential decision problems, and we test its predictions empirically. The results of a first experiment suggested that the model can discern which reward structures are more conducive to metacognitive learning. A second experiment found that feedback structures designed based on the metacog-

---

[**]This section is based on Krueger, Lieder, and Griffiths (2017). Paul Krueger conducted the experiments reported in this section, analyzed the data, and contributed to writing and experimental design.

nitive reinforcement learning model can accelerate learning to plan.

5.2.1   Background: Planning and reinforcement learning

We will model each sequential decision problem we pose to our participants as a Markov decision process (MDP)

$$M = (\mathcal{S}, \mathcal{A}, T, \gamma, r, P_0),\tag{5.2}$$

where the states $\mathcal{S}$ correspond to locations, the actions $\mathcal{A}$ correspond to moves, $T(s, a, s')$ is the probability of the next location s' given the previous location s if the participant takes action a, $0 \leq \gamma \leq 1$ is the discount factor on future rewards, $r(s, a, s')$ is the reward generated by this transition, and P_0 is the probability distribution of the initial location S_0 (Sutton & Barto, 1998). We can thus describe the participant's strategy for solving the task as a *policy* $\pi : \mathcal{S} \mapsto \mathcal{A}$ that specifies which action to take in each of the locations. The expected sum of discounted rewards that a policy π will generate in the MDP M starting from a state s is known as its *value function*

$$V_M^\pi(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r\left(S_t, \pi(S_t), S_{t+1}\right)\right].\tag{5.3}$$

The optimal policy $\pi_M^\star$ maximizes the expected sum of discounted rewards, that is

$$\pi_M^\star = \arg\max_\pi \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r\left(S_t, \pi(S_t), S_{t+1}\right)\right].\tag{5.4}$$

Solving large planning problems is often intractable because the number of possible action sequences grows exponentially with the number of steps one plans ahead. When the state space $\mathcal{S}$ is discrete and relatively small, dynamic programming can be used to find optimal plans in polynomial time (Littman, Dean, & Kaelbling, 1995). But the high-dimensional, continuous state spaces people have to plan with in real life are too large for these methods. Instead, people seem to rely on approximate planning strategies (Huys et al., 2015) and often decide primarily based on immediate and proximal outcomes while neglecting the long-term consequences of their actions (Myerson & Green, 1995). Despite its fallibility, looking only a few steps ahead can drastically simplify the planning problem, and this may often be a necessity for bounded agents with imperfect knowledge of the environment (Jiang, Kulesza, Singh, & Lewis, 2015). Since cutting corners in the decision process is both necessary and problematic, good decision-making requires knowing when that is admissi-

ble and when it is not. Knowing how much to plan is therefore an important metacognitive skill to learn.

The results presented in Chapter 4 suggest that this metacognitive skill can be learned through trial and error. Learning through trial and error can be understood in terms of *reinforcement learning* (Sutton & Barto, 1998). While certain reinforcement learning algorithms can, in principle, learn to solve arbitrarily complex problems, reinforcement learning can also be very slow—especially when rewards are sparse and the optimal policy is far from the learner's initial strategy. A common approach to remedy this problem is to give the algorithm pseudo-rewards for actions that do not achieve the goal but lead in the right direction (Ng et al., 1999).While previous work has developed this idea to accelerate learning a direct mapping from states to actions, the work presented here leverages it to accelerate learning how to plan.

5.2.2 Deciding how to decide

People can use many different decision strategies. This poses the problem of deciding how to decide (Boureau et al., 2015). Here, we will formalize this problem within the framework of rational metareasoning (Russell & Wefald, 1991b) introduced in Chapter 1. Previous research on meta-decision-making has focused on the arbitration between habits versus planning (Dolan & Dayan, 2013; Keramati et al., 2011). While this is an important meta-control problem, it is only one part of the puzzle because people are equipped with more than one goal-directed decision-mechanism. Hence, when the model-based system is in charge, it has to be determined how many steps it should plan ahead. Ideally, the chosen planning horizon should achieve the optimal tradeoff between expected decision quality versus decision time (Vul et al., 2014) and mental effort (Shenhav et al., 2017).

Here, we make the simplifying assumption that people always choose the action that maximizes their sum of expected rewards over the next h steps, for some value of h that differs across decisions. A planning horizon of $h = 1$ entails looking only at the immediate outcome of each action (myopic one-step planning) whereas a planning horizon larger than one entails solving a sequential decision problem to form a multi-step plan. Under this assumption, the meta-decision problem is to select a planning horizon h from a set $\mathcal{H} = \{1, 2, \cdots, \}$, execute the plan, select a new planning horizon, and so on. More formally, this problem can be formalized as a meta-level MDP (Hay et al., 2012). In our task, the meta-level MDP is

$$M_{\mathrm{meta}} = \left(\mathcal{S}_{\mathrm{meta}}, \mathcal{H}, T_{\mathrm{meta}}, r_{\mathrm{meta}} \right),$$ (5.5)

where the meta-level state $m \in \mathcal{S}_{\text{meta}} = \{0, 1, 2, 3, 4\}$ encodes the number of remaining moves, and the meta-level action $h \in \mathcal{H} = \{1, 2, 3, 4\}$ is the planning horizon used to make a decision. The meta-level reward function r_{meta} integrates the cost of planning with the return of the resulting action:

$$r_{\text{meta}}(m_k, h_k) = -\text{cost}(h_k) + \sum_{t=1}^{h} r(s_t, \text{plan}_t^{(k, h_k)}), \tag{5.6}$$

where $\text{plan}_t^{(k, h)}$ is the tth action of the plan formed by looking h steps ahead in the meta-level state m_k. The meta-decision-maker receives this reward after the plan has been executed in its entirety. If the meta-decision-maker selects short planning horizons there can be multiple plan-act-reward-learn cycles within a single trial. The cost of planning $\text{cost}(h_k)$ is determined by the branching factor b of the decision tree according to

$$\text{cost}(h_k) = \lambda \cdot b^{h_k} \cdot h_k, \tag{5.7}$$

where b^{h_k} is the number of plans, h_k is the number of steps per plan, and λ is the cost per planning step.[††]

5.2.3   Metacognitive reinforcement learning

Solving the problem of deciding how to decide optimally is computationally intractable but the optimal solution can be approximated through learning (Russell & Wefald, 1991b). We propose that people use reinforcement learning (Sutton & Barto, 1998) to approximate the optimal solution to the meta-decision problem formulated in Equation 5.5.

Model

Our model of metacognitive reinforcement learning builds on the semi-gradient SARSA algorithm (Sutton & Barto, 1998) that was develop to approximately solve MDPs with large or continuous state spaces. Specifically, we assume that people learn a linear approximation to the meta-level Q-function

$$Q_{\text{meta}}(m_k, h_k) \approx \sum_{j=1}^{7} w_j \cdot f_j(m_k, h_k), \tag{5.8}$$

[††]This equation assumes a constant branching factor and an upper bound on the complexity of planning. People's planning time likely increases less than exponentially fast with the planning horizon but our approximation may be sufficient for small problems.

whose features $\mathbf{f}$ comprise one indicator variable for each possible planning horizon h ($f_1 = \mathbb{1}(h = 1), \cdots, f_4 = \mathbb{1}(h = 4)$), one indicator variable for whether or not the agent planned all l steps until the end of the task ($f_5 = \mathbb{1}(h = l)$), the number of steps that were left unplanned ($f_6 = \max\{0, l - h\}$), and the number of steps the agent planned too far ($f_7 = \max\{0, h - l\}$). The semi-gradient SARSA algorithm learns the weights of these features by gradient descent. To bring it closer to human performance, our model replaces its gradient descent updates by Bayesian learning. Concretely, the weights $\mathbf{w}$ are learned by Bayesian linear regression of the bootstrap estimate $\hat{Q}(m_k, h_k)$ of the meta-level value function onto the features $\mathbf{f}$. The bootstrap estimator

$$\hat{Q}(m_k, h_k) = r_{\text{meta}}(m_k, h_k) + \langle \mu_t, \mathbf{f}(m', h') \rangle \tag{5.9}$$

is the sum of the immediate meta-level reward and the predicted value of the next meta-level state m'. The predicted value of m' is the scalar product of the the posterior mean μ_t of the weights $\mathbf{w}$ given the observations from the first t actions (where $t = \sum_{n=1}^{k} h_n$) and the features $\mathbf{f}(m', c')$ of m' and the planning horizon h' that will be selected in that state.

We assume that the prior on the feature weights reflects that it is beneficial to plan until the end ($P(f_5) = \mathcal{N}(\mu = 1, \sigma = 0.1)$), although planning is costly ($P(f_1) = P(f_2) = P(f_3) = P(f_4) = \mathcal{N}(\mu = -1, \sigma = 0.1)$), and that planning too much is more costly than planning too little ($P(f_7) = \mathcal{N}(\mu = -1, \sigma = 0.1)$ and $P(f_6) = \mathcal{N}(\mu = 0, \sigma = 0.1)$).

Given the posterior on the feature weights $\mathbf{w}$, the planning horizon h is selected by Thompson sampling. Specifically, to make the kth meta-decision, a weight vector $\tilde{w}$ is sampled from the posterior distribution of the weights given the series of meta-level states, selected planning horizons, and resulting value estimates experienced so far. That is,

$$\tilde{w}_k \sim P(\mathbf{w}|\mathcal{E}_k), \tag{5.10}$$

where the set $\mathcal{E}_k = \{e_1, \cdots, e_k\}$ contains the meta-decision-maker's experience from the first k meta-decisions; to be precise, each meta-level experience $e_j \in \mathcal{E}_k$ is a tuple $\left(m_j, h_j, \hat{Q}(m_j, c_j; \mu_j)\right)$ containing a meta-level state, the computation selected in it, and the bootstrap estimates of its Q-value. The sampled weight vector $\tilde{w}$ is then used to predict the Q-values of each possible planning horizon $h \in \mathcal{H}$ according to Equation 5.8. Finally, the planning horizon with the highest predicted Q-value is used for decision-making.

By proposing metacognitive reinforcement learning as a mechanism of cognitive plasticity, our

model suggests that reward and feedback are critical for cognitive growth. Conceptualizing metacognitive reinforcement learning as a regression problem suggests that learning how to best think about a problem should require less practice the stronger the correlation between the features $\mathbf{f}(m, c)$ (i.e., the predictors) and the resulting reward net the cost of thinking (i.e., the criterion; Green, 1991). Here, we apply our model to predict how quickly people can learn that more planning leads to better results from the reward structure of the practice problems. According to the model, learning should be fastest when the reward increases deterministically with the planning horizon both within and across problems. By contrast, learning should be slower when this relationship is degraded by additional variability in the rewards that is unrelated to planning. The following experiments test this prediction and illustrate the model's utility for designing feedback structures that promote metacognitive learning.

5.2.4   Experiment 4: Reward structures can help or hinder learning to plan

Methods

We recruited 304 adult participants from Amazon Mechanical Turk. The task took about 25 minutes, and participants were paid $2.50 plus a performance-dependent bonus of up to $2.00. Participants played a series of *flight planning* games. The environment consisted of six different cities, each connected to two other cities (Figure 5.10). Participants began each trial at a given city, and were tasked with planning a specified number of flights. Each flight was associated with a known gain or loss of money, displayed onscreen. Thus, the participants' task was to plan a route that would maximize their earnings or minimize their losses, based on the number of planning steps required for that game.

The experiment comprised thirteen trials total: a sequence of three practice problems which required planning 2, 3, and 3 steps ahead, respectively, followed by ten 4-step problems, with a break after trial eight. The order of the two 3-step problems was randomized, and the order of the ten 4-step problems was randomized across the last ten trials of the experiment. Participants were assigned randomly to one of two conditions: environments with reward structures designed to promote learning ("diagnostic rewards"), or environments with reward structures designed to hinder learning ("non-diagnostic rewards").

The problems of the diagnostic rewards condition were automatically generated to exhibit four characteristics:

1. For each l-step problem, planning $h < l$ steps ahead generates $l - h$ suboptimal moves. In other words, each myopic planner makes the maximum possible number of mistakes.

2. When the number of moves is l, then planning l steps ahead yields a positive return, but planning $h < l$ steps ahead yields a negative return.

3. The return increases monotonically with the planning horizon from 1 to the total number of moves.

4. Each starting position occurs at least once.

Round 1 of 13

Location: Smithsville   **Flight: 1 of 2**   **Earnings: $0**   **Bonus: $0**
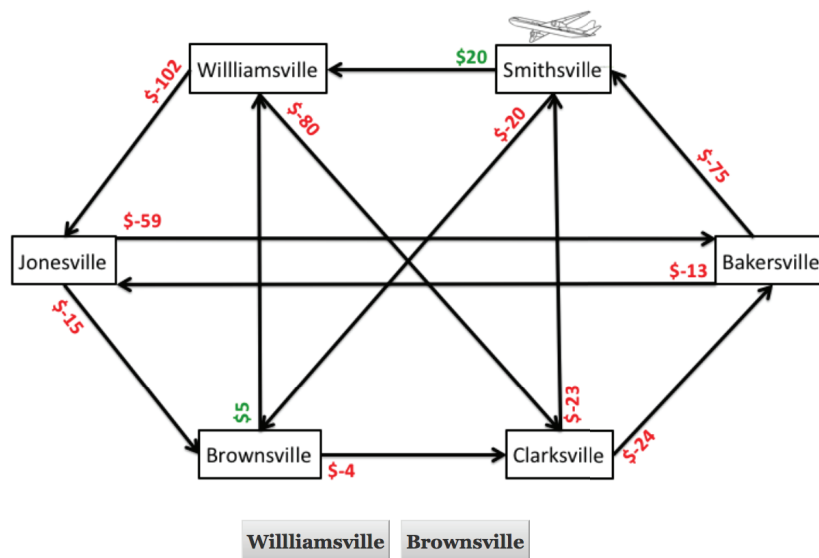

Figure 5.10: Screenshot of a problem from Experiment 4.

The reward structures used for the non-diagnostic rewards condition were created by shifting the diagnostic reward structures so as to degrade the correlation between planning horizon and reward. Concretely, for half of the problems all rewards were shifted down such that no amount of planning could achieve a return better than $-\$10$. Since the original problems were such that the 1-step planner always performed worst, the shift was $\frac{-r_1 + X}{l}$ where r_1 is the return of the 1-step planner,

l is the number of steps in the planning problem, and X is a random number between 10 and 20 that differed across problems ($X \sim \text{Uniform}([10, 20])$). For the other half of the problems, all rewards were shifted up by $-\frac{r_1 + X}{l}$ such that all planners achieve a return of at least $+\$10$. These reward structures make it extremely difficult for metacognitive reinforcement learning to discover that planning is valuable, because the random shifts greatly diminish the correlation between planning horizon and reward.

Results

Both model simulations and human behavior demonstrated enhanced learning in environments with diagnostic rewards. Figure 5.11 shows the mean performance of the metacognitive reinforcement learning model, and the mean performance of human participants. Here, performance is measured as relative reward

$$R_{rel} = (R - R_{min})/(R_{max} - R_{min}), \tag{5.11}$$

where R is the total reward received during the trial, and R_{min} and R_{max} are the highest and lowest possible total reward on that trial, respectively.

To measure the effects of condition and trial number on performance in human participants, we ran a repeated-measures ANOVA. This revealed a significant effect of both trial number ($F(9, 2989) = 3.44, p < 0.001$) and condition ($F(9, 3029) = 15.26, p < 0.0001$), such that participants improved over time, and participants with diagnostic feedback performed better than those without. To measure learning in each group, we ran a simple linear regression of the relative reward on the trial number. This revealed a significant regression equation for participants who received diagnostic rewards ($F(2, 302) = 11.28, p < 0.01$), with an R^2 of 0.59, but not for participants who received non-diagnostic rewards ($F(2, 302) = 3.51, p > 0.05$), with an R^2 of 0.31, suggesting that improvement in performance occurred with diagnostic rewards, but not without.

To analyze the frequency with which participants chose the optimal route, we performed a multinomial logistic regression of whether or not each participant chose the optimal route on trial number and group. This revealed significant effects of trial number ($p < 10^{-6}$) and group ($p < 0.0001$).

In addition, we found that participants interacting with a diagnostic reward structure learned to plan significantly further ahead than participants interacting with the non-diagnostic reward structure. When there were four steps left, the average planning horizon was 2.96 with diagnostic
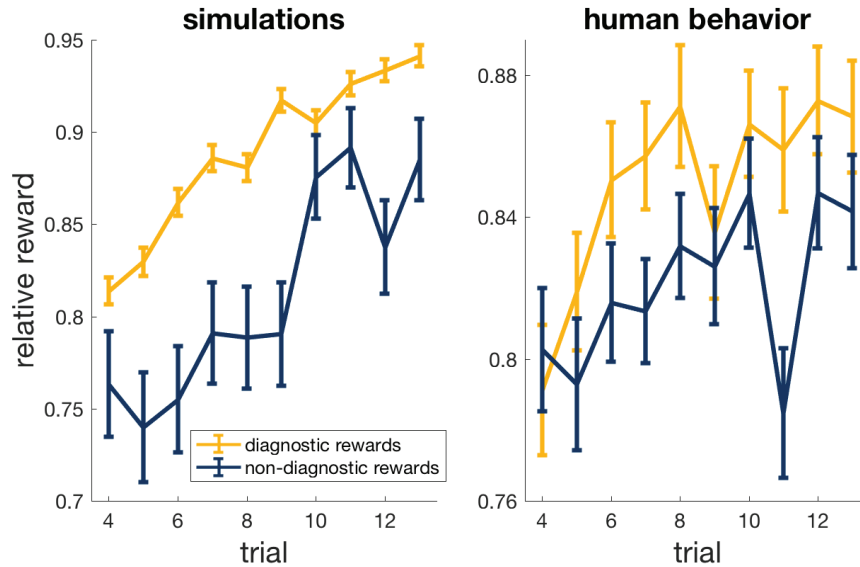
Figure 5.11: Model predictions and human performance in Experiment 4. Error bars indicate the standard error of the mean. Model predictions were averaged over 500 simulations.

rewards compared to 2.65 with non-diagnostic rewards ($t(596) = 2.94, p < 0.01$). When the rewards were diagnostic of good planning, participants' choices in the first step of the 4-step problems accorded 10.3% more frequently with 4-step planning ($t(302) = 3.57, p < 0.001$). For 3 remaining steps there was a significant increase in choices according with optimal 1-step ($p < 0.01$), 2-step ($p < 0.01$) and 4-step planning ($p < 0.01$). For 2 remaining steps, there was a significant increase in choices according with optimal 1-step planning ($p < 0.0001$) without a decrease in agreement with other planning horizons. Finally, on the last move participants' choices in the environment with diagnostic rewards corresponded 5.8% more frequently with optimal 1-step planning ($t(302) = 3.71, p < 0.001$), and significantly less frequently with 2-step and 3-step planning ($p < 0.01$ and $p < 0.001$). In summary, diagnostic rewards led to better agreement between the planning horizon and the number of remaining steps.

5.2.5 EXPERIMENT 5: USING FEEDBACK TO PROMOTE LEARNING TO PLAN

When one has control over the reward structure of an environment, creating rewards tailored to faster learning may be feasible. However, often environmental rewards are fixed. In Experiment 5, we tested whether providing feedback may be an effective alternative approach to accelerating

learning. When participants do not plan enough to find the optimal route, this could be because the time cost of planning an optimal route outweighs its benefits. To change that, we provided feedback in the form of timeout penalties for short-sighted decisions.

Methods

We recruited 324 adult participants on Amazon Mechanical Turk. The task took about 30 minutes, and participants were paid \$3.00 plus a performance-dependent bonus of up to \$2.00. Participants played twenty trials of the flight planning game described above. These trials were divided into a training block and a testing block. The training block consisted of six trials requiring 2-step planning, followed by ten trials requiring 3-step planning. The testing block consisted of four additional 3-step trials. The order of the 2-step trials and the order of the 3-step trials were randomized across subjects. Participants were randomly assigned to either the feedback condition or the control condition.

In the training block, participants in the feedback condition were told their apparent planning horizon at the end of every trial and penalized with a timeout that reflected the amount of planning they had eschewed. Concretely, we set the durations of the timeouts such that the cost of short-sighted decisions was proportional to the amount of necessary planning the participant had eschewed. Specifically, the forgone cost of planning was estimated by cost $= 2^{l-\hat{h}}$, where l is the number of moves for that trial, $\hat{h}$ is the participant's apparent planning horizon, and 2 is the branching factor since each step entailed a binary decision. The participant's planning horizon was estimated by the number of consecutive moves consistent with the optimal policy, beginning with the last move, followed by the second-to-last, etc. At the end of each trial of the first block, participants in the feedback group were penalized with a timeout delay for sub-optimal routes. The delay was calculated as $7 \cdot (\text{cost} - 1)$ seconds. During this period, participants were unable to proceed to the next trial. If participants performed the optimal route, they were able to proceed immediately to the next trial.

The control group received no feedback and had to wait a fixed amount of time at the end of every trial in block 1, regardless of their performance. This fixed period was set to 8 seconds, to match the mean timeout period for participants in the feedback group (7.9 seconds). Neither group received feedback or delays in the test block.

The planning problems presented in this experiment were created in two steps. In the first step,

we created 2- and 3-step problems with maximally diagnostic reward structures (according to the criteria used in Experiment 4) subject to the constraint that the first move with the highest immediate reward was optimal for exactly half of those problems. In the second step, we modified these problems so as to deteriorate the correlation between planning horizon and reward using the same method we employed to create the non-diagnostic reward structures used in Experiment 4.

Model Predictions

We applied the metacognitive reinforcement learning model described above to the problem of learning how many steps one should plan ahead. We simulated a run of the experiment described above with 1000 participants in each condition. The simulations predicted a gradual increase in the relative return from the first 3-step problem to the last one (see Figure 5.12). With feedback, the relative return increased faster and reached a higher level than without feedback.

Results

To quantify the effects of condition and trial number on performance (measured as relative reward), we ran a mixed-design repeated-measures ANOVA on participant performance during the 3-step trials. This revealed a significant effect of feedback ($F(9, 4521) = 8.54, p < 0.01$) and trial number ($F(9, 4521) = 1.85, p < 0.05$) on relative reward. To measure learning in each group, we performed a simple linear regression of relative reward on trial number for the 3-step trials in the training block (i.e., when participants in the feedback group received feedback). This revealed a significant regression equation for the feedback group ($F(2, 322) = 5.28, p = 0.05$), with an R^2 of 0.40 but not for the control group ($F(2, 322) = 1.57, p > 0.05$), with an R^2 of 0.16. This suggests that participants who received feedback improved during the training block but the control group did not.

Feedback increased the model's average performance in both the training block and the transfer block. We next tested whether the enhanced learning of the feedback group during training resulted in better performance in the transfer block (trials 17-20) where they no longer received any feedback. A two-sample t-test revealed that the feedback group's advantage in the testing block was nearly significant ($t(1294) = 1.53, p = 0.063$). Figure 5.12 compares our participants' performance to the model predictions.
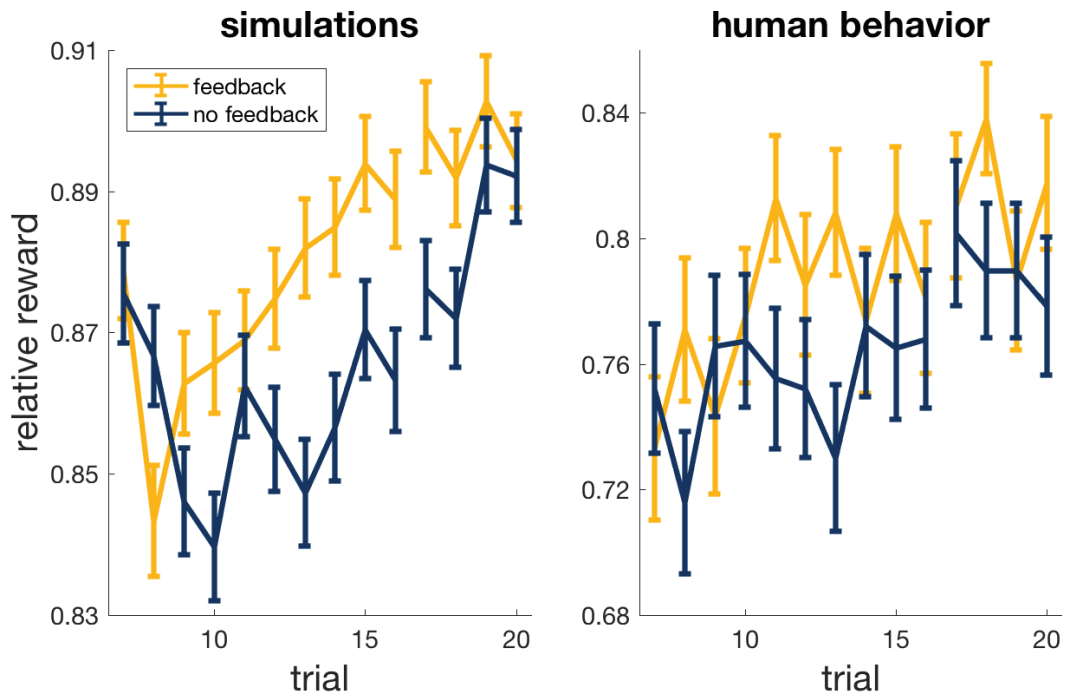
Figure 5.12: Results of Experiment 5. The metacognitive RL model predicts that feedback accelerate learning to plan. Human behavior shows a similar pattern of results.

As predicted by our model, a multinomial logistic regression of whether or not each participant chose the optimal route on trial number and feedback, revealed significant effects of trial number ($p < 0.0001$) and feedback ($p < 0.01$).

Feedback appeared to increase people's planning horizons: when there were two remaining moves, the choices of the feedback group accorded 4% less often with myopic choice ($t(1398) = -2.17, p < 0.05$), 7% more often with optimal 2-step planning ($t(1398) = 3.44, p < 0.001$), and 4% more often with optimal 3-step planning ($t(1398) = 2.43, p < 0.05$).

5.2.6   Discussion

This section introduced a computational model of how people learn to decide better. Its central idea is that learning how to think can be understood as metacognitive reinforcement learning. The metacognitive reinforcement learning model presented in this section extends the strategy selection learning model introduced in Chapter 4 by capturing that choosing cognitive operations is a sequential decision problem with potentially delayed rewards rather than a one-shot decision. The new model correctly predicted the effects of reward structure and feedback on learning to plan: Experiment 4 suggested that our model captures the effect of reward structures on the speed of metacognitive learning. We then applied our theory to design feedback for people's performance in environments whose reward structure is not diagnostic of good planning. Experiment 5 confirmed the model's prediction that this intervention would be effective.

The results suggest two pragmatic approaches to promoting cognitive growth: the first approach is to design reward structures that are diagnostic of the quality of reasoning, planning, and decision-making; the second approach is to provide feedback on the process by which a decision was made. In Experiment 5 we followed the latter approach by designing feedback based on the cost of planning; but other types of feedback may also be useful. If cognitive plasticity is based on model-free reinforcement learning as assumed by our theory, then its speed should critically depend on how well the feedback people receive upon performing cognitive operations reflects their value. Therefore, feedback structures that align immediate feedback with long-term value should be maximally effective at promoting cognitive plasticity and learning to make better decisions. This idea for designing feedback structures can be implemented using the optimal gamification method introduced by Lieder and Griffiths (2016). Feedback designed using optimal gamification could be especially beneficial because the underlying method of reward shaping is designed to accelerate model-free reinforcement learning (Ng et al., 1999). Critically, to promote learning how to decide, people should

decide without any assistance and only receive feedback *after* their choice.

The theory of metacognitive reinforcement learning presented in this section is a step towards establishing a scientific foundation for designing feedback for cognitive training and other interventions for promoting cognitive growth. Future work will evaluate alternative forms of feedback, address the problem of transfer and retention, and design more effective training paradigms where the feedback people receive is maximally informative about how people think and decide. As a first step in this direction, Chapter 7 applies the findings of this section to develop a cognitive tutor that gives people metacognitive feedback to teach them optimal planning strategies.

6

An automatic method for strategy discovery*

6.1 Introduction

The idea that people use simple heuristics is central to a substantial body of work on bounded rationality, and the school of ecological rationality assumes that people's simple heuristics are rational Gigerenzer (2008a); Todd and Gigerenzer (2012). To the extent that this is true, discovering rational heuristics will give us insights into how people think and decide. Conversely, to the extent that

people's cognitive strategies deviate from those rational heuristics it should be possible to improve human judgment and decision-making by teaching people to use such heuristics. Unfortunately, there is no principled way to discover rational heuristics and most known heuristics lack any normative justification. To overcome these problems, this chapter develops an automatic method for discovering rational heuristics.

The results presented in Chapter 5 suggest that people might be discovering cognitive strategies via metacognitive reinforcement learning. Inspired by this finding, this chapter translates the idea of metacognitive reinforcement learning into a computational method for discovering rational heuristics.

The first two sections present and evaluate a computational method for deriving near-optimal cognitive strategies from first principles. The third section applies this method to discover rational heuristics for multi-alternative risky-choice. We find that our method automatically rediscovers known heuristics and also uncovers a novel heuristic. An experiment confirmed that people do indeed use each of the discovered heuristics under the conditions for which our method reveals it to be resource-rational. The chapter concludes with a discussion of these findings and directions for future research.

6.2 Defining optimal cognitive strategies

The key idea of this chapter is that optimal cognitive strategies can be defined and computed as the solution to a meta-level MDP (see Figure 6.1). To recap from Chapter 1, a metalevel MDP

$$M_{\mathrm{meta}} = (\mathcal{B}, \mathcal{A}, T_{\mathrm{meta}}, r_{\mathrm{meta}}) \tag{6.1}$$

is a Markov decision process (Puterman, 2014) where the actions $\mathcal{A}$ are cognitive operations, the states $\mathcal{B}$ encode the agent's beliefs, and the transition function T_{meta} describes how cognitive operations change the beliefs. $\mathcal{A}$ includes computations $\mathcal{C}$ that update the belief, as well as a special metalevel action $\perp$ that terminates deliberation and initiates acting on the current belief. A belief state b encodes a probability distribution over parameters θ of a model of the domain. The parameters θ determine the utility of acting according to a policy π, that is $U_\pi^{(\theta)}$. For one-shot decisions, $U_\pi^{(\theta)}$ is the expected reward of a taking a single action. In sequential decision-problems, $U_\pi^{(\theta)} = V_\pi^{(\theta)}(s)$ is the expected sum of rewards the agent will obtain by acting according to policy π if the environment has the characteristics encoded by θ. Since b encodes the agent's belief about θ, its subjective utility $\hat{U}_\pi^{(b)}$ of acting according to π is $\mathbb{E}_{\theta \sim b}[U_\pi^{(\theta)}]$.
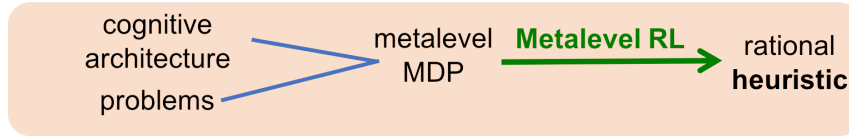
Figure 6.1: Illustration of the strategy discovery method developed in this chapter.

The metalevel reward function r_{meta} captures the cost of thinking (Shugan, 1980) and the external reward r the agent expects to receive from the environment. The computations $\mathcal{C}$ have no external effects, thus they always incur a negative reward $r_{\text{meta}}(b, c) = -\text{cost}(c)$. In the problems studied below, all computations that deliberate have the same cost, that is $\text{cost}(c) = \lambda$ for all $c \in \mathcal{C}$ whereas $\text{cost}(\perp) = 0$. An external reward is received only when the agent terminates deliberation and makes a decision based on the current belief state b. To reduce the variance of this reward signal, the metalevel reward of terminating deliberation is defined as the expectation of the external reward, that is

$$r_{\text{meta}}(b, \perp) = \max_\pi \hat{U}_\pi^{(b)} = \max_\pi \mathbb{E}_{\theta \sim b}\left[U_\pi^{(\theta)} \right]. \tag{6.2}$$

Early work on rational metareasoning (Russell & Wefald, 1991b) defined the optimal way to select computations as maximizing the value of computation (VOC), that is

$$\arg\max_c \text{VOC}(c, b), \tag{6.3}$$

where $\text{VOC}(c, b)$ is the expected improvement in decision quality that can achieved by performing computation c in belief state b and continuing optimally minus the cost of the optimal sequence of computations (Russell & Wefald, 1991b). When no computation has positive value, the policy terminates computation and executes the best object-level action, thus $\text{VOC}(\perp, b) = 0$. Using the formalism of metalevel MDPs (Hay et al., 2012), this definition can be rewritten as

$$\text{VOC}(c, b) = Q_{\text{meta}}^\star(b, c) - r_{\text{meta}}(b, \perp), \tag{6.4}$$

and the optimal selection of computations can be expressed as the optimal metalevel policy $\pi_{\text{meta}}^\star(b) = \arg\max_c Q_{\text{meta}}^\star(b, c)$.

6.3 Computing optimal cognitive strategies through meta-level reinforcement learning

6.3.1 Approximations to rational metareasoning

Previous work (C. H. Lin, Kolobov, Kamar, & Horvitz, 2015; Russell & Wefald, 1991b) has approximated rational metareasoning by the meta-greedy policy

$$\pi_{\text{greedy}}(b) = \arg \max_c \text{VOC}_1(b, c), \tag{6.5}$$

where

$$\text{VOC}_1(c, b_t) = \mathbb{E}\left[r_{\text{meta}}(B_{t+1}, \bot)|b_t, c_t\right] + r_{\text{meta}}(b_t, c) - r_{\text{meta}}(b_t, \bot), \tag{6.6}$$

is the myopic value of computation (Russell & Wefald, 1991b). This is optimal when the improvement from each additional computation is less than that from the previous one but deliberates too little when this assumption is violated.

Hay et al. (2012) approximated rational metareasoning by combining the solutions to smaller metalevel MDPs that formalize the problem of deciding how to decide between one object-level action and the expected return of its best alternative. While this *blinkered* approximation is more accurate than the meta-greedy policy, it is also significantly less scalable and not directly applicable to metareasoning about planning.

It has been proposed that people approximate optimally selecting individual computations by metareasoning over a small subset of all possible sequences of computations (Milli et al., 2017). The solution to this simplified problem can be approximated efficiently (Lieder & Griffiths, 2017), but this approximation neglects the sequential nature of selecting individual computations.

To our knowledge these are the main approximations to rational metareasoning. Hence, to date, there appears to be no accurate and scalable method for solving general metalevel MDPs.

6.3.2 Metalevel reinforcement learning

It has been proposed that metareasoning can be made tractable by learning an approximation to the value of computation (Russell & Wefald, 1991b). However, despite some preliminary steps in this direction (Harada & Russell, 1998; Lieder, Krueger, & Griffiths, 2017; Lieder, Plunkett, et al., 2014) and related work on meta-learning (Schaul & Schmidhuber, 2010; Smith-Miles, 2009; Thornton,

Hutter, Hoos, & Leyton-Brown, 2013; J. X. Wang et al., 2017), learning to approximate bounded optimal information processing remains an unsolved problem in artificial intelligence.

   Previous research in cognitive science suggests that people circumvent the intractability of metareasoning by learning a metalevel policy from experience (Cushman & Morris, 2015; Krueger et al., 2017; Lieder & Griffiths, 2017). At least in some cases, the underlying mechanism appears to be model-free reinforcement learning (RL) (Cushman & Morris, 2015; Krueger et al., 2017; J. X. Wang et al., 2017).This suggests that model-free reinforcement learning might be a promising approach to solving metalevel MDPs. To our knowledge, this approach is yet to be explored in artificial intelligence. Here, we present a proof-of-concept that near-optimal metalevel policies can be learned through metalevel RL.
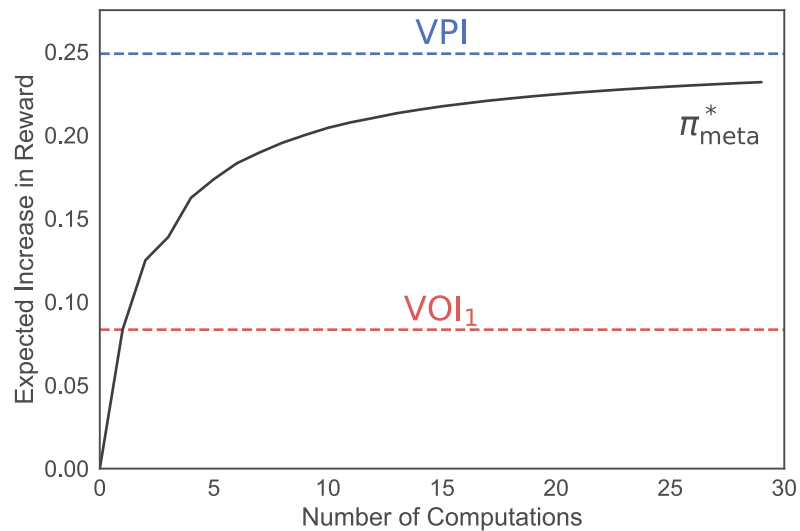
Figure 6.2: Expected performance in metareasoning about how to choose between three actions increases monotonically with the number of computations, asymptoting at the value of perfect information (VPI). Consequently, the value of executing a single computation must lie between the myopic value of information (VOI$_1$) and the VPI.

6.3.3   A METALEVEL RL ALGORITHM FOR SELECTING COMPUTATIONS

According to rational metareasoning, one should continue to reason until none of the available computations has a positive VOC. Until then, one should always choose the computation that confers

the highest improvement in decision-quality net its cost. While the improvement in decision quality contributed by a computation c under the optimal continuation is generally intractable to compute, it can be bounded. Figure 6.2 illustrates that if the expected decision quality improves monotonically with the number of computations, then the improvement achieved by the optimal sequence of computations should lie between the advantage of deciding immediately after the first computation over making a decision without it (Russell & Wefald, 1991) and the benefit of obtaining perfect information about all actions (Howard, 1966). The former is given by the myopic value of information[†], that is

$$\text{VOI}_1(c, b_t) = \mathbb{E}_{B_{t+1}|b_t, c} \left[\max_\pi \hat{U}_\pi^{(B_{t+1})} \right] - \max_\pi \hat{U}_\pi^{(b)}. \tag{6.7}$$

The latter is given by the value of perfect information about all actions, that is

$$\text{VPI}_{\text{all}}(b) = \mathbb{E}_{\theta \sim b} \left[\max_\pi U_\pi^{(\theta)} \right] - \max_\pi \hat{U}_\pi^{(b)}. \tag{6.8}$$

In problems with many possible actions, this upper bound can be very loose, and the VOC may be closer to the value of knowing the value functions of the policies Π_c about whose returns the computation c is informative, that is

$$\text{VPI}_A(b, c) = \mathbb{E}_{\theta \sim b} \left[\max \left(\mathcal{U}_c^{(\theta)} \cup \hat{\mathcal{U}}_{\neg c}^{(b)} \right) \right] - r_{\text{meta}}(b, \perp), \tag{6.9}$$

where $\mathcal{U}_c^{(\theta)} = \{U_\pi^{(\theta)} : \pi \in \Pi_c\}$ are the unknown utilities of the policies that computation c is informative about, and $\hat{\mathcal{U}}_{\neg c}^{(b)} = \{\hat{U}_\pi^{(b)} : \pi \notin \Pi_c\}$ is the set of the expected utilities of all policies that c is not informative about. This definition generalizes the value of perfect information about a single action (Dearden, Friedman, & Russell, 1998) to policies.

Critically, the myopic value of information (VOI_1), the VPI about all actions, and the VPI_A can all be computed efficiently or efficiently approximated by Monte-Carlo integration (Hammersley & Handscomb, 1964). Our method thus approximates the expected improvement in decision quality gained by a computation by linearly interpolating between its myopic VOI and the value of perfect information, that is

$$\text{VOC}(c, b) \approx w_1 \cdot \text{VOI}_1(c, b) + w_2 \cdot \text{VPI}_{\text{all}}(b) + w_3 \cdot \text{VPI}_A(b, c) - w_4 \cdot \text{cost}(c), \tag{6.10}$$

with the constraints that $w_1, w_2, w_3 \in [0, 1]$, $w_1 + w_2 + w_3 = 1$, and $w_4 \in [1, h]$ where h is an upper bound on how many computations can be performed. Below we propose an algorithm through which the agent can learn these weights from experience.

Since the VOC defines the optimal metalevel policy (Equation 6.3), we can approximate the opti-

---

[†]The VOI_1 defined here is equal to the myopic VOC defined by Russell and Wefald (1991) plus the cost of the computation.

mal policy by plugging in our VOC approximation (Eq. 6.10) into Equation 6.3. This yields

$$\pi_{\text{meta}}(b; \mathbf{w}) = \arg\max_c w_1 \cdot \text{VOI}_1(c, b) + w_2 \cdot \text{VPI}_{\text{all}}(b) + w_3 \cdot \text{VPI}_A(b, c) - w_4 \cdot \text{cost}(c). \quad (6.11)$$

The parameters $\mathbf{w}$ of this policy are estimated by maximizing the expected return

$$\mathbb{E}\left[\sum_t r_{\text{meta}}(b_t, \pi_{\text{meta}}(b_t; \mathbf{w}))\right].$$

Together with the constraints on the weights stated above, this effectively reduces the intractable problem of solving metalevel MDPs to a simple 3-dimensional optimization problem. There are many ways this optimization problem could be solved. Since estimating the expected return for a given weight vector can be expensive, we use Bayesian optimization (BO) (Mockus, 2012) to optimize the weights in a sample efficient manner.

The novelty of our approach lies in leveraging rational metareasoning and machine learning to discover optimal cognitive strategies. In the following sections, we validate the assumptions of our approach, evaluate its performance on increasingly complex metareasoning problems, compare it to existing methods, and apply it to discover rational heuristics for risky choice.

6.3.4   Evaluation of the method in simulations

We evaluate how accurately our method can approximate rational metareasoning against two state-of-the-art approximations–the meta-greedy policy and the blinkered approximation–on three increasingly difficult metareasoning problems: deciding when to stop thinking, deciding how to decide, and deciding how to plan.

1. Metareasoning about when to stop deliberating

How long should an agent deliberate before answering a question? Our evaluation mimics this problem for a binary prediction task (e.g., "Will the price of the stock go up or down?"). Every deliberation incurs a cost and provides probabilistic evidence $X_t \sim \text{Bernoulli}(p)$ in favor of one outcome or the other. At any point the agent can stop deliberating and predict the outcome supported by the majority of its deliberations so far. The agent receives a reward of $+1$ if its prediction is correct, or incurs a loss of -1 if it is incorrect. The goal is to maximize the expected reward of this one prediction minus the cost of computation.
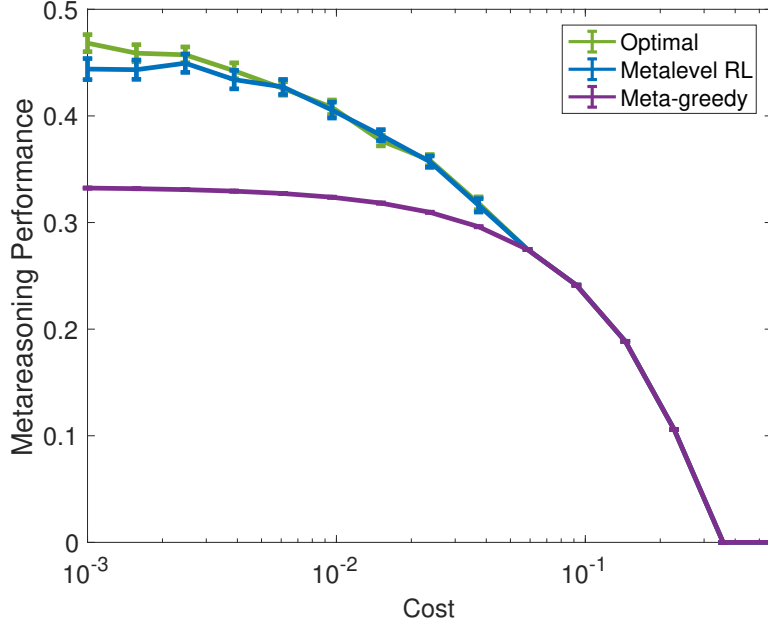
Figure 6.3: Results of performance evaluation on the problem of metareasoning about when to terminate deliberation.

METALEVEL MDP: We formalize the problem of deciding when to stop thinking as a metalevel MDP $M_{\text{meta}} = (\mathcal{B}, \mathcal{A}, T_{\text{meta}}, r_{\text{meta}})$ where each belief state $(\alpha, \beta) \in \mathcal{B}$ defines a beta distribution over the probability p of the first outcome. The metalevel actions $\mathcal{A}$ are $\{c_1, \perp\}$ where c_1 refines the belief by sampling, and $\perp$ terminates deliberation and predicts the outcome that is most likely according to the current belief. The transition probabilities for sampling are defined by the agent's belief state, that is $T_{\text{meta}}((\alpha, \beta), c_1, (\alpha + 1, \beta)) = \frac{\alpha}{\alpha + \beta}$ and $T_{\text{meta}}((\alpha, \beta), c_1, (\alpha, \beta + 1)) = \frac{\beta}{\alpha + \beta}$. Predicting (executing $\perp$) always transitions to a terminal state. The reward function r_{meta} reflects the cost of computation ($r_{\text{meta}}(b, c_1) = -\lambda$) and the probability of making the correct prediction $(r_{\text{meta}}(b, \perp) = +1 \cdot p_{\text{correct}}(\alpha, \beta) - 1 \cdot (1 - p_{\text{correct}}(\alpha, \beta)))$ where $p_{\text{correct}}(\alpha, \beta) = \max\{\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta}\})$. We set the horizon to $h = 30$, meaning that the agent can perform at most 30 computations before making a prediction.

Since there is only one object-level action (i.e., to predict the outcome that appears most likely) the VPI about all actions is identical to the VPI for a single action. When reporting on this problem, we will thus not distinguish between them and use the term VPI instead. For the same reason, the blinkered approximation is equivalent to solving the problem exactly.

EVALUATION PROCEDURE: We evaluated the potential of our method in two steps: First, we performed a regression analysis to evaluate whether the proposed features are sufficient to capture the value of computation. Second, we tested whether the proposed features are sufficient to learn a near-optimal metalevel policy. The metalevel RL agent learns the weights $\mathbf{w}$ of the policy defined in Equation 6.11 that maximize expected return through Gaussian process Bayesian optimization. We ran 500 iterations of optimization, estimating the expected return of the policy entailed by the probed weight vector by its average return across 2500 episodes. The performance of learned policy was evaluated on an independent test set of 3000 episodes.

To perform these evaluations, we first established the ground truth by solving the metalevel MDP with backward induction (Puterman, 2014).

RESULTS: First, linear regression analyses confirmed that three simple features ($VOI_1(c, b)$, $VPI(c, b)$, and $cost(c)$) are sufficient to capture between 90.8% and 100.0% of the variance in the value of computation for performing a simulation ($VOC(b, c_1)$) across different states b depending on the cost of computation. Concretely, as the cost of computation increased from 0.001 to 0.1 the regression weights shifted from $0.76 \cdot VPI + 0.46 \cdot VOI_1 - 4.5 \cdot cost$ to $0.00 \cdot VPI + 1.00 \cdot VOI_1 - 1.00 \cdot cost$ and the explained variance increased from 90.8% to 100.0%. The explained variance and the weights remained the same for costs greater than 0.1. Figure 6.4a) illustrates this fit for $\lambda = 0.02$.

Second, we found that the VOI_1 and the VPI features are sufficient to learn a near-optimal metalevel policy. As shown in Figure 6.3, the performance of metalevel RL policy was at most 5.19% lower than the performance of the optimal metalevel policy across all costs. The difference in performance was largest for the lowest cost $\lambda = 0.001$ ($t(2999) = 3.75, p = 0.0002$) and decreased with increasing cost so that there was no statistically significant performance difference between our method and the optimal metalevel policy for costs greater than $\lambda = 0.0025$ (all $p > 0.15$). The policy learned with BO performed between 6.78% and 35.8% better than the meta-greedy policy across all costs where the optimal policy made more than one observation (all $p < 0.0001$) and 20.3% better on average ($t(44999) = 42.4, p < 10^{-15}$).

a)   b)

$$\widehat{\mathrm{VOC}} = 0.0977 * \mathrm{VPI} + 0.926 * \mathrm{VOI_1} - 1.05 * \mathrm{cost}$$
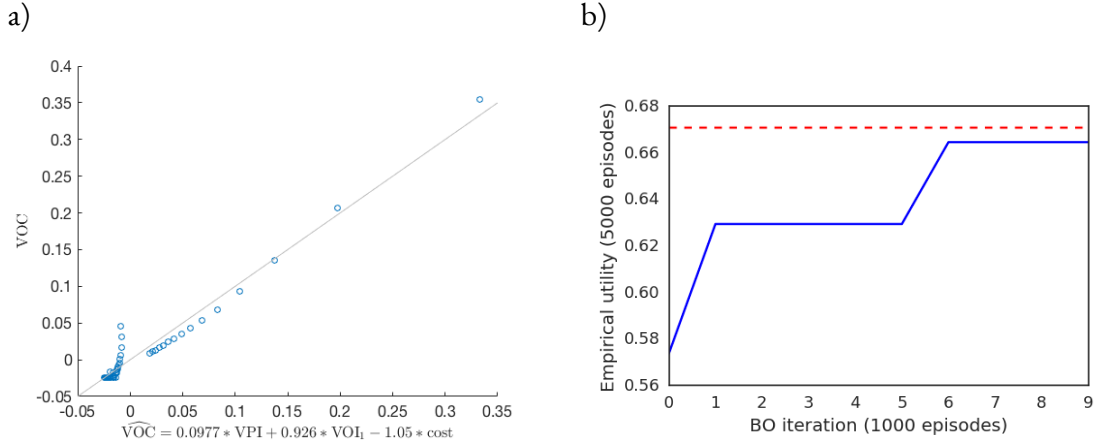
Figure 6.4: a) Linear fit of the true value of computation in terms of the the myopic VOI, the value of perfect information, and the cost of computation. b) Example of a convergence plot of meta-level reinforcement learning method.

2. METAREASONING ABOUT DECISION-MAKING

How should an agent allocate its limited decision-time across estimating the expected utilities of multiple alternatives? To evaluate how well our method can solve this kind of problem, we evaluate it on the *Bernoulli metalevel probability model* introduced by Hay et al. (2012). This problem differs from the previous one in two ways. First, instead of having only a single object-level action (i.e., make a prediction), there are now $k \geq 2$ object-level actions. Second, instead of making a prediction and being rewarded for its accuracy, the agent chooses an action a_i and receives a payoff that is sampled from its outcome distribution, that is $r(s, a_i) \sim \mathrm{Bernoulli}(\theta_i)$ where θ_i is the action's unknown reward probability. This problem differs from the standard multi-armed bandit problem in two ways: First, the agent takes only a single object-level action and thus receives only one external reward. Second, the agent is equipped with a simulator that it can use to estimate the reward probabilities $\theta_1, \cdots, \theta_k$ via sampling; simulated outcomes do not count towards the agent's reward, but each simulation has a cost.

METALEVEL MDP: The Bernoulli metalevel probability model is a metalevel MDP $M_{\mathrm{meta}} = (\mathcal{B}, \mathcal{A}, T_{\mathrm{meta}}, r_{\mathrm{meta}}, h)$ where each belief state b defines k Beta distributions over the reward probabilities $\theta_1, \cdots, \theta_k$ of the k possible actions. Thus b can be represented by $((\alpha_1, \beta_1), \ldots, (\alpha_k, \beta_k))$ where $b(\theta_i) = \mathrm{Beta}(\theta_i; \alpha_i, \beta_i)$ for all $1 \leq i \leq K$. For the initial belief state b_0 these parame-

ters are $\alpha_i = \beta_i = 1$ for all $1 \leq i \leq k$. The metalevel actions $\mathcal{A}$ are $\{c_1, \ldots, c_k, \perp\}$ where c_i simulates action a_i and $\perp$ terminates deliberation and executes the action with the highest expected return, that is action $\arg\max_i \frac{\alpha_i}{\alpha_i + \beta_i}$. The metalevel transition probabilities $(T_{\mathrm{meta}}(b_t, c_i, b_{t+1}))$ encode that performing computation c_i increments α_i with probability $\frac{\alpha_i}{\alpha_i + \beta_i}$ and increments β_i with probability $\frac{\beta_i}{\alpha_i + \beta_i}$. The metalevel reward function $r_{\mathrm{meta}}(b, c)$ is $-\lambda$ for $c \in \{c_1, \cdots, c_k\}$ and $r_{\mathrm{meta}}(b, \perp) = \max_i \frac{\alpha_i}{\alpha_i + \beta_i}$. Finally, the horizon h is the maximum number of metalevel actions that can be performed and the last metalevel action has to be to terminate deliberation and take action ($\perp$).

EVALUATION PROCEDURE: We evaluated our method on Bernoulli metalevel probability problems with $k \in \{2, \cdots, 5\}$ object-level actions, a horizon of $h = 25$, and computational costs ranging from 10^{-4} to 10^{-1}. We evaluated the performance of metalevel RL against the optimal metalevel policy and three alternative approximations: the meta-greedy heuristic (Russell & Wefald, 1991b), the blinkered approximation (Hay et al., 2012), and the metalevel policy that always deliberates as much as possible. We trained the metalevel RL policy with Bayesian optimization as described above, but with 100 iterations of 1000 episodes each. To combat the possibility of overfitting, we evaluated the average returns of the five best weight vectors over 5000 more episodes and selected the one that performed best. The optimal metalevel policy and the blinkered policy were computed using backward induction (Puterman, 2014). We evaluated the performance of each policy by its average return across 2000 episodes for each combination of computational cost and number of object-level actions.

RESULTS: An analysis of variance confirmed that these four methods differed significantly in their performance ($F(4, 279932) = 123078.5, p < 10^{-15}$). We found that the policy obtained by metalevel RL attained 99.2% of optimal performance (0.6540 vs. 0.6596, $t(1998) = -6.98, p < 0.0001$) and significantly outperformed the meta-greedy heuristic (0.60, $t(1998) = 86.9, p < 10^{-15}$) and the full-deliberation policy (0.20, $t(1998) = 475.2, p < 10^{-15}$). The performance of our method (0.6540) and the blinkered approximation (0.6559) differed by only 0.29%.

Figure 6.5b shows the metareasoning performance of each method as a function of the number of options. We found that our method's performance scaled well with the size of the decision problem. For each number of options, the relative performance of the different methods was consistent with the results reported above.
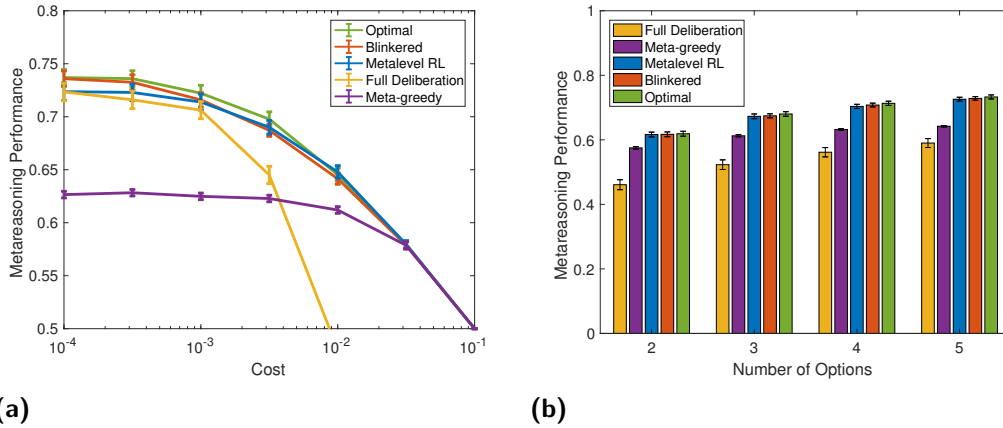
(a)   **(b)**

Figure 6.5: (a) Metareasoning performance as a function of the cost of computation. Error bars enclose 95% confidence intervals. (b) Metareasoning performance (i.e. expected reward of chosen option minus cost of the decision process) of alternative methods on the Bernoulli metalevel probability model as a function of the number of actions. Error bars enclose 95% confidence intervals.

Figure 6.5a shows the methods' average performance as a function of the cost of computation. An ANOVA confirmed that the effect of the metareasoning method differed across different costs of computation ($F(24, 279932) = 64401.4, p < 10^{-15}$). Our method outperformed the meta-greedy heuristic for costs smaller than 0.03 (all $p < 10^{-15}$), and the full-deliberation policy for costs greater than 0.0003 (all $p < 0.005$). For costs below 0.0003, the blinkered policy performed slightly better than our method (all $p < 0.0005$). For all other costs both methods performed at the same level (all $p > 0.1$), with the exception of the cost $\lambda = 0.01$, for which our method outperformed the blinkered approximation ($t(1998) = 3.2, p = 0.001$). Additionally, for costs larger than 0.01, our method's performance becomes indistinguishable from the optimal policy's performance (all $p > 0.24$).

Finally, as illustrated in Figure 6.4b), we found that our metalevel RL algorithm learned surprisingly quickly, usually discovering near-optimal policies in less than 10 iterations.

3. Metareasoning about planning

Having evaluated our method on problems of metareasoning about how to make a one-shot decision, we now evaluate its performance at deciding how to plan. To do so, we define the *Bernoulli metalevel tree*, which generalizes the Bernoulli metalevel probability model by replacing the one-shot

decision between k options by a tree-structured sequential decision problem that we will refer to as the *object-level MDP*. The transitions of the object-level MDP are deterministic and known to the agent. The reward associated with each of $K = 2^{h+1} - 1$ states in the tree is deterministic, but initially unknown; $r(s, a, s_k) = \theta_k \in \{-1, 1\}$. The agent can uncover these rewards through reasoning at a cost of $-\lambda$ per reward. When the agent terminates deliberation, it executes a policy with maximal expected utilty. Unlike in the previous domains, this policy entails a sequence of actions rather than a single action.

METALEVEL MDP: The Bernoulli metalevel tree is a metalevel MDP $M_{\mathrm{meta}} = (\mathcal{B}, \mathcal{A}, T_{\mathrm{meta}}, r_{\mathrm{meta}})$ where each belief state b encodes one Bernoulli distribution for each transition's reward. Thus, b can be represented as $(p_1, \cdots, p_K)$ such that $b(\theta_k = 1) = p_k$ and $b(\theta_k = -1) = 1 - p_k$. The initial belief b_0 has $p_k = 0.5$ for all k.u The metalevel actions are defined $\mathcal{A} = \{c_1, \cdots, c_K, \perp\}$ where c_k reveals the reward at state k and $\perp$ selects the path with highest expected sum of rewards according to the current belief state. The transition probabilities $T_{\mathrm{meta}}(b_t, c_k, b_{t+1})$ encode that performing computation c_k sets p_k to 1 or 0 with equal probability (unless p_k has already been updated, in which case c_k has no effect). The metalevel reward function is defined $r_{\mathrm{meta}}(b, c) = -\lambda$ for $c \in \{c_1, \cdots, c_K\}$, and $r_{\mathrm{meta}}((p_1, \cdots, p_k), \perp) = \max_{\mathbf{t} \in \mathcal{T}} \sum_{k \in \mathbf{t}} \mathbb{E}[\theta_k \mid p_k]$ where $\mathcal{T}$ is the set of possible trajectories $\mathbf{t}$ through the environment, and $\mathbb{E}[\theta_k | p_k] = 2p_k - 1$ is the expected reward attained at state s_k.

THE RECURSIVELY BLINKERED POLICY   The blinkered policy of Hay et al. (2012) was defined for problems where each computation informs the value of only one action. This assumption of "independent actions" is crucial to the efficiency of the blinkered approximation because it allows the problem to be decomposed into one independent subproblem for each action. When there are few computations associated with each action, each subproblem can be efficiently solved.

Critically, the Bernoulli metalevel tree violates the assumption of independent actions. This is because here "actions" are policies, and the reward at each state affects the values of all policies visiting that state. One can still apply the blinkered policy in this case, approximating the value of a computation c_k by assuming that future computations will be limited to $\mathcal{E}_{c_k}$, the set of computations that are informative about *any* of the policies the initial computation is relevant to. However, for large trees, this only modestly reduces the size of the initial problem. This suggests a recursive generalization: Rather than applying the blinkered approximation once and solving the resulting subproblem exactly, we recursively apply the approximation to the resulting subproblems.

Finally, to ensure that the subproblems decrease in size monotonically, we remove from $\mathcal{E}_{c_k}$ the computations about rewards on the path from the agent's current state to the state s_k inspected by computation c_k and call the resulting set $\mathcal{E}'_{c_k}$. Thus, we define the *recursively blinkered policy* as $\pi^{\text{RB}}(b) = \arg\max_c Q^{\text{RB}}(b, c)$ with $Q^{\text{RB}}(b_t, \perp) = r_{\text{meta}}(b_t, \perp)$ and

$$Q^{\text{RB}}(b_t, c_t) = \mathbb{E}\left[r_{\text{meta}}(b_t, c_t) + \max_{c_{t+1} \in \mathcal{E}'_{c_t}} Q^{\text{RB}}(B_{t+1}, c_{t+1})\right]. \tag{6.12}$$

EVALUATION PROCEDURE: We evaluated each method's performance by its average return over 5000 episodes for each combination of tree-height $h \in \{2, \cdots, 6\}$ and computational cost $\lambda \in \{2^{-7}, \cdots, 2^0\}$. To facilitate comparisons across planning problems with different numbers of steps, we measured the performance of meta-level policies by their expected return divided by the tree-height.

We trained the metalevel RL policy with Bayesian optimization as described above, but with 100 iterations of 1000 episodes each. To combat the possibility of overfitting, we evaluated the average returns of the three best weight vectors over 2000 more episodes and selected the one that performed best.

For metareasoning about how to plan in trees of height 2 and 3, we were able to compute the optimal metalevel policy using dynamic programming. But for larger trees, computing the optimal metalevel policy would have taken significantly longer than 6 hours and was therefore not undertaken.

RESULTS: We first compared our method with the optimal policy for $h \in \{2, 3\}$, finding that it attained 98.4% of optimal performance (0.367 vs. $0.373, t(159998) = -2.87 p < 10^{-15}$). An ANOVA of the performance of the approximate policies confirmed that the metareasoning performance differed significantly across the four methods we evaluated ($F(3, 799840) = 4625010; p < 10^{-15}$), and that the magnitude of this effect depends on the height of the tree ($F(12, 799840) = 1110179, p < 10^{-15}$) and the cost of computation ($F(21, 799840) = 1266582, p < 10^{-15}$).

Across all heights and costs, our method achieved a metareasoning performance of 0.392 units of reward per object-level action, thereby outperforming the meta-greedy heuristic ($0.307, t(399998) = 72.84, p < 10^{-15}$), the recursively blinkered policy ($0.368, t(399998) = 20.77, p < 10^{-15}$), and the full-deliberation policy ($-1.740, t(399998) = 231.18, p < 10^{-15}$).
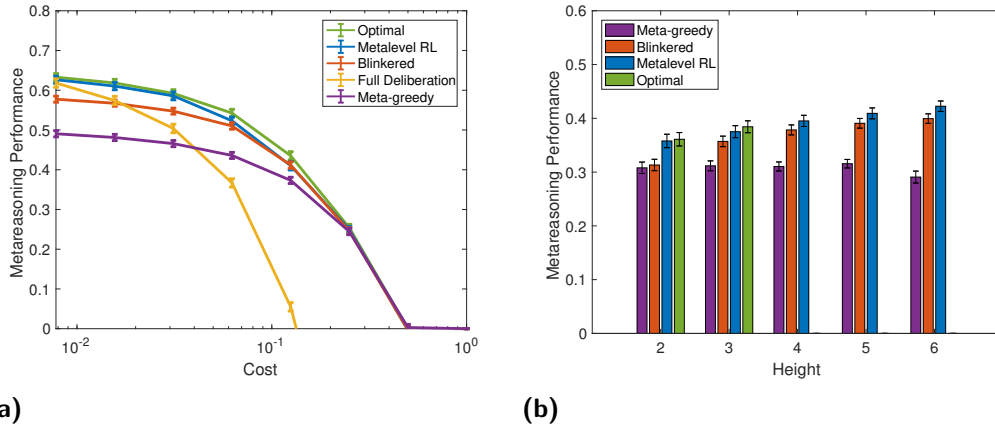
Figure 6.6: (a) Metareasoning performance as a function of computation cost on a Bernoulli tree of height three. Metareasoning performance is the average reward earned per object-level state visited. (b) Metareasoning performance as a function of tree height. The optimal policy is only shown for heights at which it can be computed in under six hours. The full observation policy is not shown because its performance is negative for all heights.

As shown in Figure 6.6a, our method performed near-optimally across all computational costs, and its advantage over the meta-greedy heuristic and the tree-blinkered approximation was largest when the cost of computation was low, whereas its benefit over the full-deliberation policy increased with increasing cost of computation. Finally, Figure 6.6b shows that the performance of our method scaled very well with the size of the planning problem, and that its advantage over the meta-greedy heuristic increased with the height of the tree.

6.3.5 Discussion

This section has introduced the first computational method for discovering resource-rational cognitive strategies. Its basic idea is to learn a near-optimal mapping from belief states to cognitive operations. We have validated this approach by showing that it learns near-optimal meta-level policies and outperforms the state-of-the-art methods for approximate metareasoning that could have been used instead. Since our method approximates the value of computation as a linear combination of the myopic VOI and the value of perfect information, it can be seen as a generalization of the meta-greedy approximation (C. H. Lin et al., 2015; Russell & Wefald, 1991a). It is the combination of multiple tractable features that capture different aspects of the value of computation with RL that makes our method tractable and powerful. Metalevel RL works well across a wider range of

267

problems than previous approximations because it reduces arbitrarily complex metalevel MDPs to low-dimensional optimization problems.

While we illustrated this approach using a policy search algorithm based on Bayesian optimization, there are many other RL algorithms that could be used instead, including policy gradient algorithms, actor-critic methods, and temporal difference learning with function approximation (e.g., Lieder, Krueger, & Griffiths, 2017).

Critically, our method can be used to derive the optimal cognitive strategy that people should use in a particular situation from assumptions about the mind's cognitive architecture and the nature of the problems to be solved (see Figure 6.1). To demonstrate this, the following section shows that it can be used to discover rational heuristics for multi-alternative risky choice.

6.4   Discovering rational heuristics for risky choice

The human mind appears to be equipped with multiple different decision strategies (Gigerenzer & Selten, 2002; Payne et al., 1988). This toolbox is assumed to include fast-and-frugal heuristics (Gigerenzer & Goldstein, 1996) as well as slower and more effortful strategies. Examples of fast-and-frugal heuristics are Take-The-Best (TTB), which chooses the alternative that is favored by the most predictive attribute and ignores all other attributes, satisficing (SAT) (Simon, 1956), which chooses the first alternative whose expected value exceeds some threshold, and random choice. In addition to fast-and-frugal heuristics people also also appear to use strategies that trade more mental effort for higher accuracy across a wider range of problems, such as the Weighted-Additive Strategy (WADD), which computes all gambles' expected values based on all possible payoffs. Work on risky choice suggests that people adaptively switch between multiple different strategies depending on how much time is available and whether one of the outcomes is much more likely than the others (Payne et al., 1988). Uncovering the strategies that people use to make decisions in everyday life is the subject of ongoing research, and it remains unclear whether and under which conditions it is rational for people to use them. There is currently no systematic to discover which strategies people should use in a given environment or to prove that a discovered strategy is indeed optimal. The closest the field has come to a model of optimal decision strategies in the Mouselab paradigm is the Directed Cognition model (Gabaix et al., 2006). This model assumes that people select decision operations according to a myopic cost-benefit analysis that approximates the VOC of a decision operation by how much

better the myopic value of computation and introduces a family of macro-operators

$$(O_{g,n})_{1 \leq g \leq \text{nr. gambles}, 1 \leq n \leq \text{nr. outcomes}}, \tag{6.13}$$

that are defined such that $O_{g,n}$ inspect the n most informative payoffs of gamble g. The directed cognition model captures some adaptive aspects of human decision-making. But its predictions are neither normative nor do they perfectly capture people's strategies.

To address these problems, we apply the automatic strategy discovery method described above to derive resource-rational heuristics for multi-alternative risky choice, measure how people's decision mechanisms compare to those rational strategies, and formally test our resource-rational theory against the Directed Cognition model. To reveal people's decision strategies in a way that makes them comparable to the rational strategies discovered by our method, we employ the Mouselab paradigm that is widely used to study multi-alternative risky choice (Johnson, Payne, Bettman, & Schkade, 1989). In this paradigm the alternatives are gambles and the attributes of each gamble are its payoffs in the event of different outcomes. As illustrated in Figure 6.7, the Mouselab paradigm traces people's decision process by recording the order in which they inspect different pieces of information. Concretely, participants are presented with a payoff matrix where the columns correspond to the alternatives they are choosing between and the rows corresponding to different outcomes. Each cell in the payoff matrix specifies how much the alternative corresponding to its column would pay if the event of corresponding to its row was to occur. Critically, all of the payoffs are initially occluded and the participant has to click on a cell to reveal its entry. The probabilities of the different outcomes are known to the participant. Each click comes at a cost, and participants are free to inspect as many or as few cells as they would like.

Our method rediscovered two known heuristics, TTB and random choice, as resource-rational strategies. TTB emerged as the resource-rational strategy for the majority of high-stakes decisions where one outcome is much more probable than any other outcome, whereas random choice emerged as the resource-rational strategy for the majority of low-stakes decisions in an environment where almost all outcomes are equally probable. In addition, our computational method discovered a novel heuristic that combines TTB with satisficing. Our experiment demonstrated that people do indeed use the newly discovered heuristic and confirmed our rational model's predictions of when people use which strategy: people used simple heuristics more frequently when the stakes were low, employed fast-and-frugal heuristics more often when one outcome was much more likely than any other outcome, and invested more time and effort when the stakes were high. This is the first demonstration that the principle of resource-rationality can be leveraged to discover people's cognitive strategies automatically.

6.4.1 Optimal decision-making

To model the meta-decision problem posed by the Mouselab task, we characterize the decision-maker's belief state b_t by probability distributions on the expected values $e_1 = \mathbb{E}[v_{O,g_1}], \cdots e_n = \mathbb{E}[v_{O,g_n}]$ of the n available gambles $g_1, \cdots, g_n$. Furthermore, we assume that for each element $v_{o,g}$ of the payoff matrix V there is one computation $c_{o,g}$ that inspects the payoff $v_{o,g}$ and updates the agent's belief about the expected value of the inspected gamble according to Bayesian inference. Since the entries of the payoff matrix are drawn from the normal distribution $\mathcal{N}(\bar{v}, \sigma_v^2)$, the resulting posterior distributions are also Gaussian. Hence, the decision-maker's belief state b_t can be represented by $b_t = (b_{t,1}, \cdots, b_{t,n})$ with

$$b_{t,g} = \left(b_{t,g}^{(\mu)}, b_{t,g}^{(\sigma^2)} \right), \tag{6.14}$$

where $b_{t,g}^{(\mu)}$ and $b_{t,g}^{(\sigma^2)}$ are the mean and the variance of the probability distribution on the expected value of gamble g of the belief state b_t.

Given the set $\mathcal{O}_t$ of the indices $(k_o^{(1)}, k_g^{(1)}), \cdots, (k_o^{(t)}, k_g^{(t)})$ of the t observations made so far, the means and variances characterizing the decision-maker's beliefs are given by

$$b_{t,g}^{(\mu)} = \sum_{(o,g)\in\mathcal{O}} p(o) \cdot v_{o,g} + \sum_{(o,g)\notin\mathcal{O}} p(o) \cdot \bar{v} \tag{6.15}$$

$$b_{t,g}^{(\sigma^2)} = \sum_{(o,g)\notin\mathcal{O}} p(o)^2 \cdot \sigma_v^2. \tag{6.16}$$

The meta-level transition function $T(b_t, c_{o,g}, b_{t+1})$ encodes the probability distribution on what the updated means and variances will be given the observation of a payoff value $V_{o,g}$ sampled from $\mathcal{N}(\bar{v}, \sigma_v^2)$. The meta-level reward for performing the computation $c_{o,g} \in C$ encodes that acquiring and processing an additional piece of information is costly. We assume that the cost of all such computations is an unknown constant λ. The meta-level reward for terminating deliberation and taking action is $r_{\text{meta}}(b_t, \bot) = \max_g b_t^{(\mu)}(g)$.

6.4.2 Decision environments

To investigate how the structure of the environment affects optimal and human decision-making, we looked at how optimal and human decision strategies depend on the stakes of the decision, and we also looked at how they differ between scenarios where one outcome is much more likely than all other outcomes (high dispersion of outcome probabilities) versus scenarios where all outcomes are almost equally likely (low dispersion of outcome probabilities). Concretely, we applied our method

separately to a low-stakes environment where the payoffs range from \$0.01 to \$0.25 and a high-stakes environment where the payoffs range from \$0.01 to \$9.99. In each of these environments 50% of the problems have outcome probabilities with high-dispersion and the other 50% have outcome probabilities with low dispersion. For simplicity, the number of alternatives is always 7, the number of outcomes is always 4, and the cost of computation corresponds to \$0.01 per processed payoff. In the low-stakes environment each payoff is independently drawn from a truncated normal distribution over the range [\$0.01, \$0.25] with mean 13 cents and standard deviation of 0.7 cents ($\mathcal{N}$(\$0.13, \$0.07)). In the high-stakes environment the payoffs are independently drawn from a truncated normal distribution over the range [\$0.01, \$9.99] with mean \$5 and standard deviation of \$3 ($\mathcal{N}$(\$5, \$3)). For low-dispersion problems each outcome probability lies between 0.1 and 0.4. For the high-dispersion problems one of the outcome probabilities is at least 0.85.

6.4.3 Model predictions

Computing the optimal policy for the meta-level MDP defined above is intractable, but it can be approximated with the method introduced above. To do so, we applied our meta-level reinforcement learning method separately to each of the four types of environments: high stakes and high dispersion, high stakes and low dispersion, low stakes and high dispersion, and low stakes and low dispersion. For each environment our metalevel RL algorithm was run for 30 iterations with 1000 episodes per iterations.

The meta-level MDP described above formalizes the costs and benefits of acquiring and processing information: acquiring additional information can improve the decision that will be taken later on but also incurs an immediate cost. The optimal decision strategy thus has to tradeoff decision quality versus decision time and mental effort. This tradeoff depends on the stakes of the decision such that higher stakes usually warrant more deliberation. Likewise, since processing probable outcomes is more likely to improve the quality of the resulting decision than processing improbable outcomes, we expect our model to prioritize probable outcomes over less probable outcomes—especially when one outcome is much more likely than all other outcomes.

Our computational method automatically discovered two strategies that people are known to use in the Mouselab paradigm: TTB and random choice. It revealed TTB to be resource-rational primarily for high-stakes decisions where one outcome is substantially more probable than all other outcomes; and it revealed random choice to be resource-rational primarily for low-stakes decisions where all outcomes are almost equally likely. Most importantly, it also discovered a novel hybrid

Figure 6.7: The Mouselab paradigm, showing an example sequence of clicks generated by the SAT-TTB strategy, which was discovered through approximate rational metareasoning.

strategy that combines TTB with satisficing (SAT-TTB). Like TTB, SAT-TTB inspects only the payoffs for the most probable outcome. But unlike TTB and like SAT, SAT-TTB terminates as soon as it finds a gamble whose payoff for the most probable outcome is high enough. We estimated the satisficing level of SAT-TTB by the lowest subjective expected value at a time when SAT-TTB stopped prior to having inspected all alternatives on a high-dispersion trial. For the low-stakes condition with high dispersion, where SAT-TTB was resource-rational, this value was \$0.16 (i.e., 0.40 standard deviations above the average payoff) . Our resource-rational analysis revealed SAT-TTB to be resource-rational for all decisions with low-stakes and high dispersion and for 27% of the low-stakes decisions with low-dispersion. Note that SAT-TTB generates the click sequence of TTB when it does not encounter any payoff that exceeds its aspiration level prior to inspecting the last alternative. This suggests that when people's click sequence accords with TTB they might actually be using SAT-TTB. Alternatively, we can interpret TTB as an extreme version of SAT-TTB that has a high aspiration level. Figure 6.7 illustrates this strategy.

Furthermore, our model makes intuitive predictions about the contingency of people's choice processes on the stakes and outcome probabilities. First, our model predicts that people should use the one-reason decision rules TTB and SAT-TTB more frequently on high-dispersion trials (i.e., for 90.0% of their decisions) than on low-dispersion trials (i.e., 0% of their decisions). This is intuitively rational because high dispersion means that one outcome is much more likely than all others and one-reason decision-making ignores all outcomes except for the most probable one (assuming that there are no ties). Within the high-dispersion environments, TTB is resource-rational more frequently than SAT-TTB when the stakes are high (80% vs. 0%), but when the stakes are low then

SAT-TTB is resource-rational more frequently than TTB (87% vs. 13%).

Second, our resource-rational analysis predicts that unless one outcome is much more likely than all other outcomes, people should use the simple heuristics (i.e., TTB, SAT-TTB, and random choice) primarily when the stakes are low. Random choice was resource-rational for 100% of the low-stakes problems with low dispersion, but for none of the high-stakes problems with low-dispersion. This too is intuitively rational because fast-and-frugal heuristics tend to be faster but less accurate than more effortful strategies. Thus, when the stakes are high and the dispersion is low, then it becomes resource-rational to inspect multiple outcomes (4/4 outcomes in 58% of the decisions and 3/4 outcomes in 34% of the decisions) and more alternative-outcome pairs (14.0/28 compared to 0/28 for low-stakes problems with low-dispersion). In this environment, the resource-rational strategies TTB+, which starts like TTB but then continues to inspect additional payoffs, SAT-TTB3, which only inspects some payoffs of the three most probable outcomes and ignores the least probable outcome, and SAT-TTB2, which inspects some payoffs of the two most probable outcomes but no payoffs of the two less probable outcomes; see Figure 6.8.

Third, our resource-rational analysis predicts that when the stakes are high, people should invest more time and effort (8.8 clicks per trial vs. 3.7 clicks per trial) to reap a higher fraction of the highest possible expected payoff (96.9% vs. 76.1%). This is also consistent with the rational speed-accuracy tradeoff inherent in the theory of resource-rationality Note that even when the resource-rational strategy achieved near-optimal performance in the high-stakes conditions it was always substantially more efficient than the WADD strategy (8.8 clicks per trial vs. 28 clicks per trial).

Figure 6.8 shows a detailed breakdown of the how frequently each strategy is resource-rational for each of the four types of decision-problems. As this figure shows, our resource-rational analysis almost perfectly predicted which strategy people use most frequently in each of the four types of environments. However, this figure also shows that people's strategy choices were less concentrated on the modal strategy as they should be. Instead, people's strategy choices were more spread out and differed less sharply between the four types of environments than would be resource-rational according to our analysis. This suggests that while people's strategy choices reflect the structure of the environment they do not always fully exploit the structure of each individual decision.
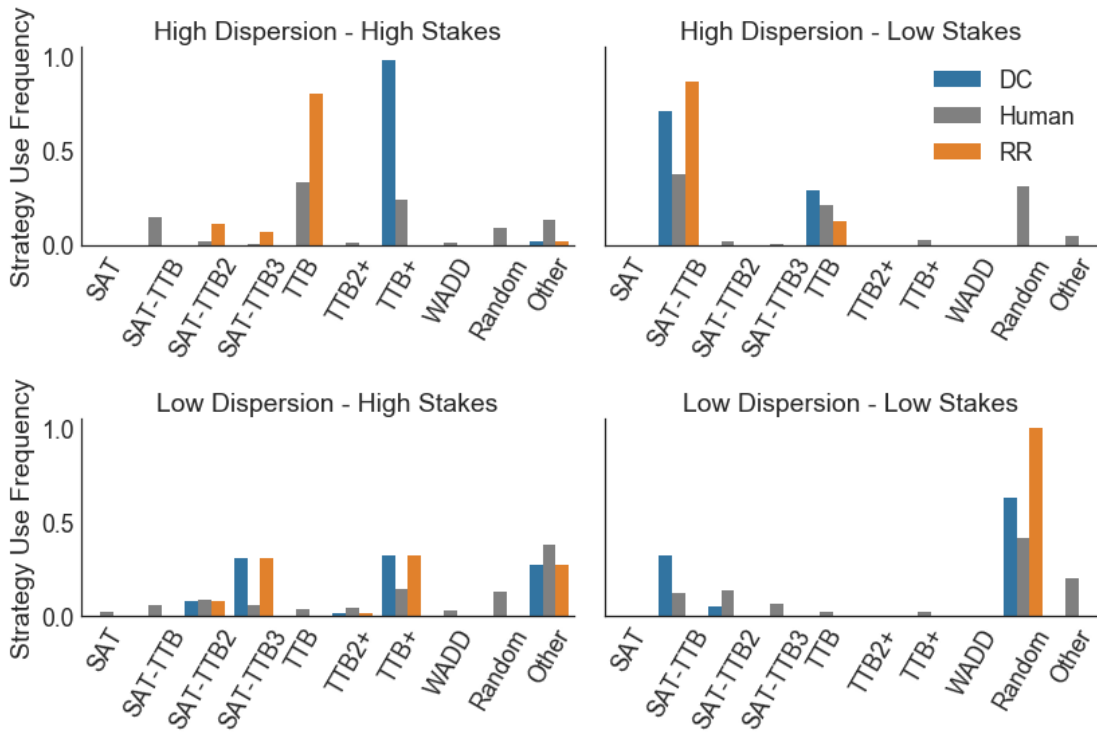
Figure 6.8: Strategy use frequencies of the resource-rational model versus people for different stakes and the outcome probabilities.

6.4.4 Experimental test of novel predictions

To test the predictions of our model, we conducted a new Mouselab experiment that manipulated the stakes of the decision and the dispersion of the outcome probabilities within subjects. Concretely, this experiment put people in the exact same environments whose optimal decision strategies are characterized above.

Methods

Participants We recruited 100 participants on Amazon Mechanical Turk. Participants received a base pay of $0.25 for about 15 minutes of work (average duration 13.5 min) plus a performance-dependent bonus of up to $5.12 (average bonus $3.39).

Experimental design The experiment used a 2×2 within-subjects design, and was structured into two blocks of 10 trials each. Within each block, all decisions had either low-stakes or high-stakes, and the order of the high-stakes block and the low-stakes block was randomly counterbalanced across participants. In low-stakes decisions, the possible payoffs ranged from $0.01 to $0.25, whereas in the high-stakes decisions the payoffs ranged from $0.01 to $9.99. All payoffs were drawn from a truncated normal distribution with mean $\frac{r_{\max}+r_{\min}}{2}$ and standard deviation $0.3 \cdot (r_{\max} - r_{\min})$. Within each block, there were five low-dispersion trials and five high-dispersion trials, ordered randomly. In low-dispersion trials, the probability of each of the four outcomes ranged from 0.1 to 0.4, whereas in high-dispersion trials, the probability of the most likely outcome ranged from 0.85 to 0.97.

Procedure Following the instructions and a comprehension check, participants performed a variation of the Mouselab task (Payne et al., 1988). Participants played a series of 20 games divided into two blocks. Figure 6.7 shows a screenshot of one game. Every game began with a 4×7 grid of occluded payoffs: there were seven gambles to choose from (columns) and four possible outcomes (rows). The occluded value in each cell specified how much the gamble indicated by its column would pay if the outcome indicated by its row occurred. The outcome probabilities were described by the number of balls of a given color in a bin of 100 balls, from which the outcome would be drawn. For each trial, participants were free to inspect any number of cells before selecting a gamble. Clicking on a cell revealed its payoff and participants were charged $0.01 per click. The value

of each inspected cell remained visible onscreen for the duration of the trial. There was no upper limit on how much time participants could spend on a trial but they were required to spend at least 30 seconds collecting information and/or waiting before they could choose a gamble. This restriction served to eliminate the opportunity cost of the participant's time so that the cost of deliberation was virtually reduced to the price they were charged were acquiring a piece of information. When a gamble was chosen the sum of the click costs was subtracted from its payoff, and participants were informed about which outcome had occurred, the resulting payoff of their chosen gamble, and their net earnings (payoff minus click costs). After the last trial, 1 of the 10 high-stakes trials and 1 of 10 low-stakes trials were selected at random and each participant received the average of their net earnings on those two trials as a bonus.

The instructions explained the task by walking the participant through the demonstration of a trial with step-by-step explanations. These explanations covered the cost of clicking and the way that their payoff was determined. The instructions also conveyed the range of payoffs in the high-stakes block and in the low-stakes block. Participants were alerted that on some trials there would be many more balls of a certain color than of other colors. After these instructions, participants were given a quiz that assessed their understanding of all critical information conveyed in the instructions including that they could inspect as many or as few payoffs as they want, the meaning of the number of balls of a given color, the fact that payoffs change from trial to trial, the fact that on some trials one outcome would be much more likely than all other outcomes, the range of payoffs for the high-stakes problems and low-stakes problems, and the fact that the outcome probabilities would be different on every trial. If a participant answered one or more questions incorrectly they were required to re-read the instructions and retake the quiz until they answered all questions correctly.

Strategy identification   We interpreted the click sequence of people and the optimal strategy in terms of six different decision strategies. A click sequence was identified as TTB when it inspected all cells in the row corresponding to the most probable outcome and nothing else. A click sequence was interpreted as SAT when one gamble's payoffs were inspected for all four outcomes consecutively, potentially followed by the inspection of all outcomes of another gamble, and so on, but leaving at least one gamble unexamined. The hybrid strategy SAT-TTB was defined as inspecting the payoffs of 1 to 6 gambles for the most probable outcome and not inspecting payoffs for any other outcome. Any click sequence that started like TTB but continued with at least one additional click was classified as TTB+. Any click sequence that a) inspected at least one payoff of each of the two most probable outcomes, b) inspected no payoffs of any other outcome, and c) did not inspect

all payoffs of the two most probable outcomes was classified as SAT-TTB2. Likewise, any strategy that a) inspected at least one payoff of each of the three most probable outcomes, b) inspected no payoffs for the least probable outcome, and c) did not inspect all payoffs of the three most probable outcomes was classified as SAT-TTB3. WADD was defined as inspecting all 28 cells column by column.The random strategy was defined as choosing an alternative without collecting any information. Click sequences that did not match any of these definitions were classified as "Other".

Results

Decision strategies   Figure 6.8 compares how often each participants used each of the six strategies introduced above for each of the four types of decision problems to the predictions of our resource-rational analysis. Our process-tracing confirmed the existence of the previously unnoticed SAT-TTB heuristic discovered by our method. Overall, people used SAT-TTB more frequently than any other heuristic; just as our resource-rational analysis had predicted. Concretely, participants used SAT-TTB for 41.5% of all decision-problems, chose randomly on 23.8% of the problems, used TTB 15.2% of the time, and rarely used WADD (1%), or SAT (0.7%). Together, the strategies reported in Figure 6.8 account for people's decision strategies on about 81.1% of all trials and for about 70.1% of all resource-rational decision mechanisms observed in this paradigm.

Our resource-rational analysis accurately predicted the relative frequency with which people used each heuristic and how it depends on the stakes of the decision and the dispersion of its outcome probabilities (see Figure 6.8). This is remarkable given that our automatic strategy discovery method did not even know that any of these strategies existed. People's strategy preferences generally agreed with the predictions of our resource-rational analysis but tended to be less extreme. This let to some discrepancies. The most pronounced discrepancy occurred for low-stakes decisions with high dispersion. Here, resource-rational analysis predicted that people should always use SAT-TTB. While people did use SAT-TTB for the majority of these decisions (69%), they also used random choice and TTB for a minority of those decisions. The second most noticeable discrepancy is that while both people and our resource-rational model used unidentified decision strategies in the high-stakes decisions with low dispersion, people also used SAT-TTB or random choice for about one fifth of them.

Consistent with our model's first prediction, people used the fast-and-frugal heuristics TTB and SAT-TTB more frequently when one outcome was much more probable than all other outcomes compared to when all outcomes were almost equally probable (39% vs. 74%; $\chi^2(1) = 21.86$,

$p =< 0.0001$). Concretely, high-dispersion increased people's reliance on TTB from 3% to 27.3% ($\chi^2(1) = 229.7, p =< 0.0001$), and increased their reliance on SAT-TTB from 36.3% to 46.6% ($\chi^2(1) = 21.9, p =< 0.0001$).

Consistent with the second prediction, participants switched to more effortful and more accurate strategies as the stakes increased. When the stakes were high, people considered a larger number of possible outcomes (9.8 ± 7.9 vs. $2.9 \pm 3.1, t(1998) = 25.4, p < 0.0001$). This increase in deliberation was reflected by a decreased reliance on simple heuristics: Overall, the frequency with which people relied on fast-and-frugal heuristics (TTB, SAT-TTB, SAT, or random choice) decreased significantly from 73.1% on low-stakes problems to 40.1% on high-stakes problems ($\chi^2(1) = 116.9, p < 0.0001$). Concretely, the frequency of random choice—the simplest heuristic—decreased significantly from 36.4% on low-stakes problems to 11.2% on high-stakes problems ($\chi^2(1) = 175.1, p < 0.0001$), and so did the frequency of the second simplest heuristic, SAT-TTB (61.4% vs. 21.5%, $\chi^2(1) = 328.0, p < 0.0001$). These substantial decreases in the frequency of the most frugal heuristics were not nearly offset by slight increases in the frequencies of TTB (11.7% vs. 18.6%, $\chi^2(1) = 18.5, p < 0.0001$) and SAT (0.2% vs. 1.2%, $\chi^2(1) = 7.2, p = 0.0073$). Finally, the data also confirmed the predicted switch from SAT-TTB to TTB: Increasing the stakes in the high-dispersion environment increased the proportion of SAT-TTB click sequences that were TTB sequences by 34% (95% CI: [29%, 40%]) from 24% to 58%.

DECISION STYLES To investigate the effects of stakes and dispersion on people's decision-processes without having to restrict the analysis to trials were people's strategy fell into one of the predefined categories, we quantified people's decision style by four metrics introduced by Payne et al. (1988): the number of inspected cells (*acquisitions*), the proportion of those inspections that pertained to the most probable outcome (*prioritization*), the degree to which subsequent acquisitions inspected the payoffs of different gambles (alternatives) for the same outcome (attribute) versus the payoffs of the same alternative (gamble) for different attributes (outcomes) (*attribute-based processing*: $\frac{n_{\text{same attribute}} - n_{\text{same alternative}}}{n_{\text{same attribute}} + n_{\text{same alternative}}}$), and the average ratio of the expected value of the chosen gamble over the expected value of the optimal choice (*relative performance*). To further test our model's predictions, we ran a 2-way mixed-effects ANOVA for each of these four metrics.

As shown in Figure 6.9, the effects of the stakes and outcome probabilities on the four metrics confirmed the predictions of our resource-rational analysis. Our model's first prediction that high dispersion promotes the use of fast-and-frugal heuristics was confirmed by a decrease in the number of acquisitions ($F(1, 1898) = 21.97, p < 0.0001$) in conjunction with an increases in attribute-

	Low Dispersion	High Dispersion
Low Stakes	People: 2.7 RRA: 0.0	People: 3.1 RRA: 3.6
High Stakes	People: 11.0 RRA: 14.0	People: 8.6 RRA: 7.4

Table 6.1: Number of information acquisitions (clicks) by people compared to the predictions of resource-rational analysis (RRA) by experimental condition.

based processing ($F(1, 1252) = 403.5, p < 0.0001$) and prioritization ($F(1, 1425) = 812.2, p < 0.0001$). The increase in prioritization was especially striking: while only 49.4% of participants' clicks inspected the most probable outcome when dispersion was low, 84.8% of them focused on the most probable outcome when the dispersion was high.

The second prediction that higher stakes should decrease people's reliance on fast-and-frugal heuristics was confirmed by a significant increases in the number of acquisitions ($F(1, 1898) = 974.6, p < 0.0001$) which was accompanied by a slight decrease in prioritization (67.0% vs. 63.7%, $F(1, 1425) = 73.0, p < 0.0001$) and an increase in relative performance ($F(1, 1898) = 127.2, p < 0.0001$).

The third prediction that higher stakes make people think harder and perform better was confirmed by the finding that higher stakes significantly increased the number of acquisitions ($F(1, 1898) = 974.6, p < 0.0001$) and relative performance ($F(1, 1898) = 127.2, p < 0.0001$) while reducing attribute-based processing ($F(1, 1252) = 60.4, p < 0.0001$). In quantitative terms, We found that the empirical effects of raising the stakes on the number of information acquisitions were similar to the predictions of our resource-rational analysis: 6.9 clicks (2.9 for low stakes vs. 9.8 for high stakes) for people versus 8.9 clicks for the resource-rational strategy (1.8 for low-stakes and 10.7 for high stakes); see Table 6.1. The main discrepancy was that people appeared to collected too much information in the low-stakes condition with low-dispersion.

RESOURCE-RATIONALITY   To assess the degree to which people's decision-strategies are resource-rational, we evaluated their net-performance (payoff of the chosen alternative minus decision cost) against the net-performance of resource-rational decision-making. This analysis revealed that, on average, the net-performance of our participants' decision-strategies was about 88.8% of the net-performance of resource-rational decision-making. This metric can be interpreted as a rationality quotient. Since our method only provides a lower bound on the performance of the resource-
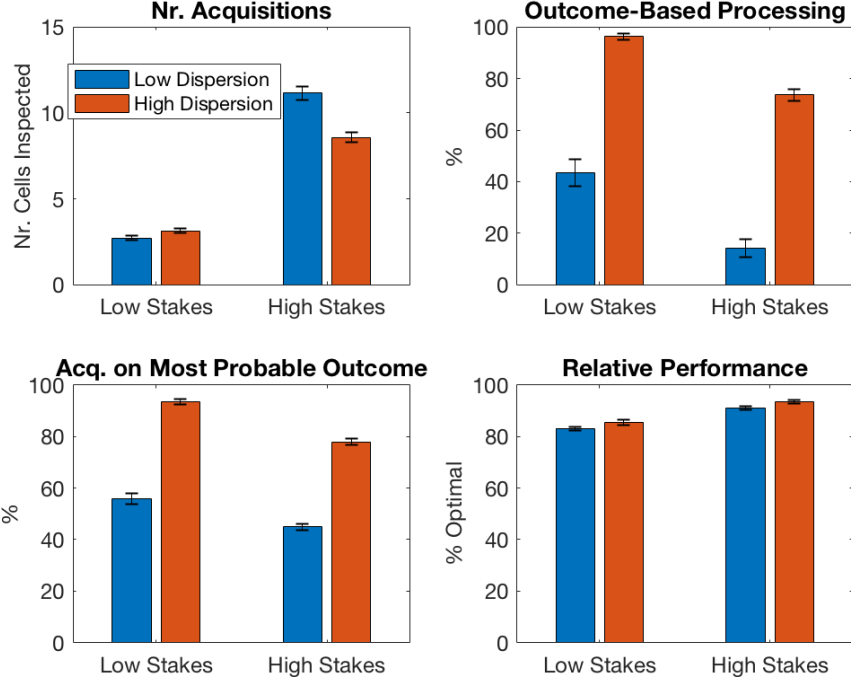
Figure 6.9: People's decision style as a function of the stakes of the decision and the dispersion of the outcome probabilities.

	Low dispersion	High dispersion
Low stakes	$\frac{\$0.11}{\$0.13} = 84.6\%$	$\frac{\$0.14}{\$0.16} = 87.5\%$
High stakes	$\frac{\$5.78}{\$6.42} = 90.0\%$	$\frac{\$7.22}{\$7.77} = 92.9\%$

Table 6.2: Relative net-performance of people (numerator), resource-rational decision-making (denominator), and their ratio (right hand side) by stakes (rows) and dispersion (columns).

rational strategy, the observed rationality quotient means that, on average, our participants were at most 88% as resource-rational as they could be. Overall, the degree to which participants' decision strategies are resource-rational was similar across all four experimental conditions – ranging from 84.6% for low-stakes trials with low-dispersion to 92.9% for high-stakes trials with high dispersion. Interestingly, their relative net-performance tended to be slightly higher for high-dispersion trials compared to low-dispersion trials (90.2% vs. 86.5%) and for high-stakes trials relative to low-stakes trials (90.7% vs. 86.1%) but these differences were not statistically significant (all $p \geq 0.0861$). For a more detailed breakdown of people's relative net-performance by condition see Table 6.2.

MODEL COMPARISONS   To formally test our model against the Directed Cognition model, we translated each of them into a likelihood model of our participants' click sequences. The likelihood models' basic assumption is that people select a click that is compatible with their strategy with probability $1 - \varepsilon$ or randomly deviate from their strategy and select one of the available decision operations uniformly at random with probability $1 - \varepsilon$, where $0 \leq \varepsilon \leq 0.5$ is a free parameter. In addition, we consider a null model according to which all decision operations are selected uniformly at random.

A formal model comparison using the Bayesian Information Criterion (G. Schwarz, 1978, BIC,) and the Akaike Information Criterion (Akaike, 1974, AIC,) showed that, overall, the resource-rational model explained participants' click sequences significantly better than the Directed Cognition model and a null model that selects computations at random ($\text{BIC}_{\text{RR}} = 79360.88 < \text{BIC}_{\text{DC}} = 81606.35 < \text{BIC}_{\text{random}} = 89172.66$; $\text{AIC}_{\text{RR}} = 78785.03 < \text{AIC}_{\text{DC}} = 81030.51 < \text{AIC}_{\text{random}} = 89172.66$). Furthermore, the data provided strong evidence for the resource-rational model over its alternatives for a majority of our participants: according to both the BIC and the AIC the click sequences of 64/100 participants were best explained by the resource-rational model, compared to only 22/100 for the Directed Cognition model and 14/100 for the null model.

Discussion

In summary, our resource-rational analysis of multi-alternative risky choice predicted some of the main strategies people use in the Mouselab paradigm and the conditions under which they are used. In addition to automatically discovering known strategies and contingencies, our computational approach also discovered a novel, previously unknown heuristic that integrates TTB with satisficing (SAT-TTB), and our experiment confirmed that people do indeed use SAT-TTB on the majority of the risky choice problems we examined — especially when the stakes are low. Comparing people's decision processes against the near-optimal decision mechanisms discovered by our method suggested that people's heuristics for multi-alternative risky choice are about 85% as resource-rational as they could be. Building on this finding, Chapter 9 explores whether it is possible to bring people even closer to resource-rationality by teaching them bounded-optimal decision strategies. Future work will provide a more precise characterization of the resource-rational strategy for high-stakes environment with low dispersion and perform an in-depth analysis of how people's decision strategies deviate from resource-rational decision-making at the level of individual decision-operations (clicks).

The application of our meta-level reinforcement learning method to discovering resource-rational decision mechanisms extends the previous approaches reviewed in Chapter 1, including the approach to solve a meta-level MDP to derive optimal stopping rules for drift-diffusion models (Tajima et al., 2016), to substantially larger spaces of more sophisticated sequential decision strategies. Our approach is can be seen as an extension of the myopic cost-benefit analysis of the directed cognition model (Gabaix et al., 2006) to looking multiple cognitive operations ahead. Concretely, while the directed cognition model approximated the value of computation by the VOI_1 minus the cost of computation, our method additionally incorporates two variants of the value of perfect information.

6.5 General Discussion

An important step in resource-rational analysis is to derive the bounded-optimal cognitive strategy from assumptions about the problem to be solved and the cognitive resources available to solve it (see Figure 1.3). Deriving the optimal strategy analytically is challenging and often requires restricting the space of possible strategies so that the optimal strategy can be found by optimizing a performance metric with respect to a few parameters (Lewis et al., 2014; Lieder, Griffiths, et al., 2018a) or a distribution (Lieder, Griffiths, & Hsu, 2017). This chapter presented an automatic computa-

tional method that can be used to optimize over a much richer class of cognitive strategies involving sequences of cognitive operations. We have validated this method by showing that it can compute near-optimal policies for simple metareasoning problems and outperforms previously proposed methods for approximate metareasoning in meta-level MDPs that are too complex to be solved exactly. We have then applied this method to discover resource-rational decision strategies for different multi-alternative risky choice environments. Our method rediscovered previously proposed heuristics and discovered a previously unknown heuristic (SAT-TTB). An experiment confirmed that people do indeed use the discovered heuristics in the environments for which they are resource-rational.

These findings support the conclusion that people's decision strategies are qualitatively consistent with the rational use of finite time and limited cognitive resources. Quantiatively speaking, we found that people's heuristics for multi-alternative risky-choice are about 85% as resource-rational as they could be. This supports a more nuanced perspective on human rationality according to which people are neither hopelessly irrational as previous work on heuristics and biases might suggest (Ariely, 2009; Marcus, 2009; Sutherland, 1992) nor optimal as a superficial reading of the literature on rational models of perception, cognition, and motor control might suggest (Griffiths & Tenenbaum, 2006; Knill & Pouget, 2004; Knill & Richards, 1996; Körding & Wolpert, 2004; Todorov, 2004; Wolpert & Ghahramani, 2000). Instead, it appears that while people generally make good use of their limited time and bounded cognitive resources, there is still room for improvement. Therefore, it might be possible to leverage our automatic strategy discovery method to enhance human judgment and decision-making by teaching people resource-rational cognitive strategies as illustrated in Chapter 9.

6.5.1 Limitations and future directions

The main limitation of the method introduced here is that it approximates bounded optimality by rational metareasoning. It optimizes a tradeoff between cognitive performance and computational cost over all mappings from belief state to computations, but some of these mappings might be infeasible for the human mind to implement. Additionally, our method ignores the computational cost of selecting computations. Thus, mappings that select inferior computations could be preferable to the optimal mapping approximated by our method if they can be implemented with less expensive metareasoning (Milli et al., 2017, 2018). These limitations may be less severe than they sound because the human brain might select computations through an associative mapping from features of belief states to computations similar to the one in the meta-level reinforcement learning

method proposed here (see also Lieder, Shenhav, et al., 2018, and Section 5.2). Concretely, if the mapping from belief states to computations was implemented in a feed-forward neural network, then all mappings require the same amount of computation. Furthermore, a sufficiently expressive neural network would be able to approximate most mappings from belief states to computations fairly accurately. While this proposal addresses both concerns in principle, whether the brain does in fact select computations in this way remains to be demonstrated. Future work will also develop efficient methods for directly computing the optimal heuristic defined in Equation 1.8. This will require explicitly postulating which cognitive strategies the mind can and cannot implement whereas the current approach was able to abstract away from the fine details of how the mind would select the optimal sequence of computations prescribed by the method presented above.

Future work might continue the resource-rational analysis cycle (see Figure 1.5) begun in the previous section by refining its assumptions about the problem to be solved and the cognitive architecture available to solve them. The problem formulation could be improved by capturing that in everyday life people are not given the outcome probabilities but have to estimate them through reasoning and information gathering. This could be addressed by augmenting the meta-level MDP with computations that inspect or estimate the outcome probabilities. Furthermore, a more realistic model of people's cognitive architecture should not assume that each piece of acquired information is always integrated Bayes optimally. Instead, alternative ways of integrating information, such as pairwise comparisons and counting should be included in the set of computations. Furthermore, a recent study found that people's strategies for multi-alternative risky choice are also shaped by memory constraints (Sanjurjo, 2017). Memory constraints could be incorporated into our meta-level MDP using a model similar to the one proposed by Yang et al. (2015).

Future work will also test an improved version of our resource-rational process model of multi-alternative risky choice against alternative models, including established heuristics (Payne et al., 1993), stochastic cumulative prospect theory (Erev et al., 2010), the directed cognition model (Gabaix et al., 2006), and the strategy selection model developed in Chapter 4. Future work will also provide a more fine-grained evaluation of the extent to which people's multi-alternative risky choice strategies are resource-rational.

Furthermore, the automatic strategy discovery method introduced in this chapter can also be used to derive rational heuristics for other cognitive domains, such as reasoning, memory, and problem solving. Our recent work on planning (Callaway et al., 2018) is just one example, and there are many more to come. Its generality and versatility make resource-rational analysis a promising modeling paradigm for all of cognitive psychology. I am therefore optimistic that resource-rational analysis

with automatic strategy discovery will enable significant advances in our understanding of the cognitive mechanisms that give rise to human intelligence and cognitive biases.

6.5.2 Conclusion

Overall, the findings presented in this chapter suggest that formulating the problem of making optimal use of finite time and limited cognitive resources as a meta-level MDP is a promising approach to discovering cognitive strategies. Automatic strategy discovery expands the scope of resource-rational analysis by enabling it to optimize over a large space of cognitive strategies, such as Take-The-Best, that proceed in a step-by-step manner. This makes it possible to discover both simple and complex rational strategies automatically; which makes it a promising starting point for uncovering people's cognitive strategies. Comparing people's heuristics against the realistic normative standard of resource-rationality will not only shed new light on the debate about human rationality, but it can also help us reverse-engineer the computational principles of human intelligence, and discover new ways to improve human judgment and decision-making. To the extent that people's heuristics are resource-rational strategies, resource-rational analysis will allow us to uncover them and reverse-engineer the cognitive and functional constraints that make them rational. And in cases where people's heuristics are not resource-rational, resource-rational analysis can help us identify genuine sub-optimalities in human reasoning and generate simple heuristics that people can use to perform better.

In the long term, our approach could be used to generate the curriculum for a course on how to make good decisions and reason effectively in the real-world. This approach would refine the idea to translate ideas from computer science into decision strategies that people can use in everyday life (Christian & Griffiths, 2016) by incorporating additional insights about the nature of the human mind. Resource-rational analysis might enable us to find out which strategies people should use in situations where their performance is genuinely poor. This might give us a much better handle on improving human reasoning and decision-making than the traditional approach of debiasing. This is because resource-rational strategies are designed to be tractable solutions to complex problems whereas the prescription to be logical, Bayesian, and maximizes expected utility is intractable given the limited resources that people have to work with. I therefore believe that leveraging the method presented in this chapter to discover optimal cognitive strategies and teach them to people is a promising new avenue for improving human judgment and decision-making, and the findings presented in Chapter 9 support this conclusion.

7

Conclusion of Part I

7.1   Resource-rational analysis of heuristics and biases

Despite their seemingly irrational cognitive biases, people have the ability to efficiently solve problems that defy artificial intelligence. Resolving this paradox is a fundamental open problem. Ideally, we would discover unifying theoretical principles that account for both people's strengths and their weaknesses. However, it is unclear whether a unifying explanation of our numerous heuristics and disparate cognitive biases is even possible. In order to tackle these problems, I have developed the theoretical framework of resource-rationality. I employ it to derive heuristics that make optimal use of people's finite cognitive resources, and to establish a rational mechanism for choosing when to use which heuristic.

The findings presented in Chapters 1-3 and 6 show that the principle of resource-rationality can explain classic biases in judgment and decision-making, while revealing the heuristics that generate them. For instance, the resource-rational process models derived in Chapters 2 and 3 provided a unifying explanation for a wide range of anchoring biases in numerical estimation, and accounted for an even wider range of availability biases in memory recall, frequency estimation, decisions from experience, and decisions from description. These rational models do more than simply list deviations from the standard picture of rationality: they explain many seemingly disparate and irrational biases

using a single rational principle. The principle of resource-rationality provides valuable constraints on the otherwise ill-posed problem of inferring people's heuristics from a limited number of decisions and judgments. This makes it possible to predict people's judgments and decisions in novel circumstances from their performance on related tasks in the laboratory. Looking forward, resource-rational analysis could be used to provide mechanistic explanations of the many seemingly unrelated and irrational cognitive biases documented in the literatures on judgment and decision-making (Gilovich et al., 2002) and behavioral economics (Ariely, 2009). Taking this approach a step further, Chapter 6 shows that we can employ the principle of resource-rationality to identify rational heuristics automatically. This is a significant advance over the haphazard theorizing predominant in research on heuristics and biases. The introduced method can be applied across all domains of human cognition including perception, problem solving, social cognition, decision-making, learning, and reasoning. I am therefore optimistic that it will allow us to uncover the principles and mechanisms responsible for both our most impressive cognitive achievements, and our most embarrassing errors.

Chapters 1–3 and 6 showed that resource-rationality can help us uncover and understand people's many heuristics, while Chapters 4 and 5 provided a resource-rational account of how people learn when to use which heuristic. Concretely, the rational metareasoning model of strategy selection offers a principled explanation for the variability, contingency, and change of people's strategy choices across multiple domains, ranging from sorting to decision-making, mental arithmetic, and problem solving. By addressing the problems of both strategy discovery and strategy selection, the resource-rational framework developed in this dissertation adds two critical missing pieces to our understanding of bounded rationality.

7.2 Redefining rationality

The anchoring bias can cause people's judgments to violate the rules of logic and probability theory. Similarly, the availability bias often leads to choices which violate the maxims of expected utility theory. These and many other cognitive biases could be interpreted as signs of human irrationality. However, as I have argued in Chapter 1, adherence to the rules of logic, probability theory, and expected utility theory is a flawed notion of rationality, because it ignores people's computational limitations. What then is to blame for people's violations of those normative principles? Is it the limitations of the human mind, or rather, the limitations of our theories of rationality? To answer this question, I have developed a realistic standard of human rationality which grants that people

have only limited cognitive resources and finite time to tackle the many millions of big and small decisions they have to make throughout life. Redefining rationality as the optimal use of finite time and limited cognitive resources has allowed me to revisit past interpretations of cognitive biases as signatures of human irrationality. My findings suggest that far from being irrational, the anchoring bias and the numerous availability biases in memory recall, judgment, and decision-making could reflect the rational use of limited resources.

Heuristics and rational models are often seen as opposites, but once the cost of computation is taken into account, heuristics can be resource-rational. This shows that resource-rational analysis has the potential to reconcile cognitive biases with the fascinating capacities of human intelligence. In addition, resource-rational analysis can be used to build bridges between rational theories, such as Bayesian models of cognition, and heuristics and other psychological process models (Griffiths et al., 2015).

Redefining rationality has profound implications not only for the interpretation of classic heuristics and biases which have shaped the debate about human rationality, but also for our approaches to cognitive modeling and our efforts to improve human judgment and decision-making. I discuss these implications in the remainder of this chapter. Then, the research presented in Part 2 will illustrate how the theory of resource-rationality can be leveraged to develop more effective tools and interventions for improving the human mind.

7.3   The debate about human rationality needs to be revisited

The standard picture of rationality posits that people should reason according to the rules of logic and probability theory and act so as to maximize their expected utility. This account was the foundation of Kahneman and Tversky's highly influential research program on heuristics and biases. Their discovery that human judgment and decision-making violate the traditional account of human rationality sparked an ongoing debate. The standard picture of human rationality was a fire waiting to happen; and the idea that people reason by applying the rules of logic and probability theory, and make calculated decisions that maximize expected utility, has since been reduced to ashes.

In Part I of my dissertation, I have identified and addressed several key limitations of the standard picture of rationality. In its place, I have proposed a qualitatively different view of what it means to be rational. This new perspective challenges the pervasive interpretation that violating the rules of logic, probability theory, and expected utility theory is a sign of irrationality. My proposal thereby

insulates the question of human rationality from the empirical demonstrations that first set the standard picture of rationality on fire. This means that there is still hope for human rationality, since the real question of the ways in which, and the extents to which people are (ir)rational remains unanswered. From this, I conclude that the debate about human rationality should be revisited, with resource-rationality as the gold standard for human judgment and decision-making. Redefining human rationality in this way calls for the reinterpretation of "cognitive bias" as the systematic suboptimal use of finite time and limited cognitive resources; this could manifest as thinking too much, thinking too little, or thinking ineffectively. Many violations of the standard picture of rationality will likely turn out not to to be cognitive biases when analysed from the resource-rational perspective. This means that the implications of virtually all previous findings on heuristics and biases will have to be re-evaluated.

One can revisit the rationality debate by going through the list of demonstrated violations of logic, probability theory, and expected utility and determine which of them are (in)compatible with resource-rationality. I have begun this endeavour in Chapters 1-3. Another approach, which I have taken in Chapters 5-6, is be to derive and test qualitative predictions of resource-rationality, and then evaluate them in human experiments. This could be seen as rebooting the research program of Tversky and Kahneman using a more accurate standard of human rationality. I believe that there is great value in identifying systematic violations of resource-rationality, because they would be genuine irrationalities that could promising targets for interventions aimed at improving human judgment and decision-making.

My findings suggest that violations of logic, probability, and expected utility theory may not be signs of irrationality but a window on resource-rational information processing and the computational constraints faced by the human mind. Overall, the findings of my dissertation paint a brighter and more nuanced picture of human rationality. According to my resource-rational perspective, people learn to make more rational use of their finite time and limited cognitive resources throughout life. My initial findings are merely the starting point of what I hope will be a widely adopted, fruitful, and long-lived research program..

7.4 IMPLICATIONS FOR COGNITIVE MODELING

Normative principles are commonly used to constrain models of human behavior and cognition. But many researchers have criticized the methodological assumption of rationality as being unreal-

istic. These criticisms arise from the observation that people often appear to violate the normative principles that we use to model their behaviors. This problem has become most evident for the rational actor models used in economics and the social sciences (Kahneman & Tversky, 1979) though Bayesian models of cognition have not gone unchallenged either (Jones & Love, 2011). The findings presented in Chapters 2-3 and the studies reviewed in Chapter 1 suggest that resource-rationality might enable us to overcome these challenges, since it reconciles the methodological assumption that people are optimal with the empirical observations that their judgments are often biased, and that their decisions often fail to maximize expected utility. It thereby enables us to develop computational models of human cognition that combine the generalizability and predictive power of rational principles with the accuracy of descriptive theories derived from empirical observations. This leaves me optimistic that resource-rationality may fill a theoretical void left by the demolition of expected utility theory, and other classical theories based on the standard picture of rationality.

While the standard picture of rationality provided constraints modeling *behavior*, resource-rationality is a methodology for modeling *cognitive mechanisms* themselves. This makes valuable headway in the transition from using rational principles to formalize the functions of cognitive systems, to using resource-rational principles to reveal the underlying cognitive mechanisms (Griffiths et al., 2015). Early rational process models were constrained by the relatively weak requirement that their outputs should lead to an optimal solution in the limit of infinite computation. Resource-rationality provides a much stronger constraint that can uniquely identify optimal, yet realistic, cognitive mechanisms. This makes it possible to automatically derive rational process models of cognitive mechanisms from a mathematical specification of their function and the cognitive architecture that executes them (see Figure 1.4).

7.5   Implications for improving human judgment and decision-making

Redefining rationality as the optimal use of limited cognitive resources has important implications for improving judgment and decision-making. One such implication is that the classic approach of debiasing might be misguided, specifically when it aims to remove 'cognitive biases' that arise from people's use of resource-rational heuristics. Another implication is that teaching people the rules of logic, probability theory, and expected utility theory might be ineffective or even harmful, since those strategies are *not* optimal for people because they neither have infinite time nor unlimited computational resources. In fact, it would seem irrational for people to apply the rules of logic, probability theory and expected utility theory to solve complex problems in limited time, because

this would likely lead to decision paralysis. Even in benign cases, using these strategies would likely cause people to invest too much time into a decision that should be made more quickly, or to make errors because the strategy requires more working memory or time than people can afford it.

Resource-rationality offers four alternative approaches to improving judgment and decision-making. Its first two proposals are well aligned with the philosophy of boosting (Hertwig & Grüne-Yanoff, 2017). First, it suggests that in order to make the best decisions possible, people should rely on efficient heuristics that are well adapted to both their cognitive capacities and the problem they are trying to solve. Resource-rationality enhances this idea with concrete mathematical and computational tools for discovering the rational heuristics that people should use (Chapter 1 and Chapter 6). Second, resource-rationality allows for the possibility that people's heuristics are already optimal even when the resulting judgments and decisions are sub-optimal. In these cases, resource-rationality suggests that we simplify or reformulate the problem at hand, so as to reduce its computational complexity. Reducing the number of options that people have to choose between is a simple example of this approach. Reformulating Bayesian reasoning problems in terms of natural frequencies is another example (Sedlmeier & Gigerenzer, 2001). Second, resource-rationality seconds the proposal of ecological rationality that we should restructure our environment so as to meet the implicit assumptions of people's heuristics. The two other implications of resource-rationality are better aligned with the philosophy of cognitive training. First, by identifying how limited cognitive resources constrain rational performance, resource-rationality can be used to identify which basic cognitive capacity, such as working memory or processing speed, should be trained to remedy a particular cognitive bias. Second, the view that people learn to make increasingly more rational use of their limited cognitive resources over time implies that reasoning and decision-making can be improved through practice. The observation that this metacognitive learning is at least partly driven by reinforcement suggests that giving people feedback on the quality of their cognitive strategies might be a promising way to help them discover resource-rational heuristics. In Part II, I build on these ideas to establish resource-rational approaches to improving human decision-making.

Part II

Expanding the bounds on human rationality

Introduction to Part II

I hope that the chapters of Part I have convinced you that resource-rationality is a promising methodological framework for modeling the mechanisms of human cognition and revisiting the debate about human rationality.

The thesis behind Part II of my dissertation is that we can leverage both the theory of resource-rationality developed in Part I and empirical and insights into bounded rationality to improve and augment the human mind using technology.

Chapter 8 argues that insights into people's cognitive limitations can guide the design of intelligent systems that enable people to overcome their cognitive biases. As a proof-of-concept, we translate insights about the bounded rationality of human decision-making into a cognitive prosthesis that restructures decision problems so that people's heuristics lead to better decisions. We find that this cognitive prosthesis helped people avoid making short-sighted decisions, overcome procrastination, and achieve their goals on time.

Chapter 9 illustrates the utility of the resource-rational framework for improving how people think and decide. In brief, the approach is to leverage the automatic strategy discovery method presented in Chapter 6 and to develop a cognitive tutor that teaches them to people. This intelligent tutoring system gives people feedback on how they plan so as to maximally accelerate the metacognitive learning mechanism identified in Chapter 5. This illustrates that the resource-rational framework can inform both the curriculum and the pedagogy of interventions for improving human judgment and decision-making. An empirical evaluation suggests that the cognitive tutor is highly effective at improving people's planning strategies, and follow-up experiments show that this improvement transfers to more difficult planning problems in more complex environments and these benefits persist for at least 24 hours.

8

Cognitive Prostheses for Goal Achievement*

WHILE ARTIFICIAL INTELLIGENCE (AI) IS PROGRESSING STEADILY and the computing power of our electronic devices continues to grow, the computing power of the human brain does not. Our bounded cognitive resources continue to constrain our decision-making and often lead to simple heuristics. Previous research has shown that these heuristics can fail miserably in certain scenarios (Ariely, 2009; Gilovich et al., 2002; Tversky & Kahneman, 1974) but perform very well in the environments they evolved for (Chater & Oaksford, 1999; Griffiths et al., 2015; Oaksford & Chater, 1994; Todd & Gigerenzer, 2007, 2012). These two observations suggest that in the future the human mind could be augmented with cognitive prostheses that use AI to automatically restructure problems on which people's heuristics perform poorly into problems on which those heuristics perform very well.

In line with this vision, previous work has found that human judgment and decision-making

can be significantly improved by restructuring how information is presented to people (Gigerenzer & Edwards, 2003; Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Johnson et al., 2012; Thaler & Sunstein, 2008), and parallel work in operations research and computer science has developed decision-support systems (Aronson, Liang, & Turban, 2004; Power et al., 2015) that use planning algorithms to solve complex, sequential decision-problems for people (Aviv & Pazgal, 2005; Bhatnagar, Fernández-Gaucherand, Fu, He, & Marcus, 1999; Gadomski, Bologna, Costanzo, Perini, & Schaerf, 2001; Nunes, de Carvalho, & Rodrigues, 2009; Song, Liu, Lawarrée, & Dahlgren, 2000). However, to the best of our knowledge, these two approaches have rarely been combined to help people overcome motivational obstacles and achieve their personal long-term goals.

One class of decision problems in which people underperform systematically involves choices whose proximal rewards are misaligned with their long-term value (e.g., persevering on a frustrating challenge versus getting drunk and watching TV). In situations like these, people's heuristics tend to reach short-sighted decisions (Ainslie & Haslam, 1992; Huys et al., 2012; Myerson & Green, 1995) that can manifest in procrastination (Steel, 2007) and impulsivity (Mischel, Shoda, & Rodriguez, 1989). This apparently myopic nature of human decision-making suggests that decision environments can be repaired by aligning each action's immediate reward with the value of its long-term consequences.

While it is generally difficult to change how people experience the actions necessary to achieve their goals (e.g., dieting, debugging, or filing taxes) relative to actions that do not (e.g., eating chocolate or watching TV), it is possible to incentivize those actions with game elements such as points, levels, and badges. This approach is known as *gamification* (Deterding, Dixon, Khaled, & Nacke, 2011). Previous research has found that gamification can have positive effects on motivation, engagement, behavior, and learning outcomes (Hamari, Koivisto, & Sarsa, 2014). Yet, determining which actions should be incentivized and by how much is still an art rather than a science and misspecified incentives can have devastating consequences (Callan, Bauer, & Landers, 2015; Devers & Gurung, 2015).

Here, we leverage ideas from artificial intelligence to develop a principled theory defining optimal incentives for helping people make better decisions. The resulting system can be interpreted as a cognitive prosthesis that uses artificial intelligence to solve people's complex sequential decision problems and uses gamification to restructure them in such a way that people can easily identify the course of action that is best for them in the long-run. This approach offloads the computational challenges of long-term planning into the reward structure of the environment, and the underlying

theory ensures that the added game elements will never incentivize counter-productive behavior.

8.1   An optimal gamification method for decision-support

The first step of our approach to optimal gamification is to model the decision environment as a Markov Decision Process (MDP; see Section 5.2.1). As a refresher, please recall that the expected sum of discounted rewards that a policy π will generate in the MDP M starting from a state s is known as its value function

$$V_M^\pi(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r\left(S_t, \pi(S_t), S_{t+1}\right)\right]. \tag{8.1}$$

The optimal policy $\pi_M^\star$ maximizes the expected sum of discounted rewards, that is

$$\pi_M^\star = \arg\max_\pi \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r\left(S_t, \pi(S_t), S_{t+1}\right)\right], \tag{8.2}$$

and its value function satisfies the Bellman equation

$$V_M^\star(s_t) = \max_a \mathbb{E}\left[r\left(s_t, a, S_{t+1}\right) + \gamma \cdot V_M^\star(S_{t+1})\right]. \tag{8.3}$$

We can therefore rewrite the optimal policy as

$$\pi_M^\star(s) = \arg\max_a \mathbb{E}\left[r(s_t, a, S_{t+1}) + \gamma \cdot V_M^\star(S_{t+1})\right], \tag{8.4}$$

which reveals that it is myopic with respect to the sum of the immediate reward and the discounted value of the next state.

Here, we leverage the MDP framework to model game elements such as points and badges as *pseudo-rewards* $f(s, a, s')$ that are added to the reward function $r(s, a, s')$ of a decision environment M to create a modified environment $M' = (\mathcal{S}, \mathcal{A}, T, \gamma, r', P_0)$ with a more benign reward function $r'(s, a, s') = r(s, a, s') + f(s, a, s')$. From this perspective, the problem with misspecified incentives is that they change the optimal policy $\pi_M^\star$ of the original decision problem M into a different policy $\pi_{M'}^\star$ that is optimal for the gamified environment M' but not for the original environment M. To avoid this problem we have to ensure that each optimal policy of M' is also an optimal policy of M.

Research on machine learning has identified which conditions pseudo-rewards must satisfy to achieve this: according to the *shaping theorem* (Ng et al., 1999) adding pseudo-rewards retains the optimal policies of any original MDP if and only if the pseudo-reward function f is *potential-based*, that is if there exists a *potential function* $\Phi : \mathcal{S} \mapsto \mathbb{R}$ such that

$$f(s, a, s') = \gamma \cdot \Phi(s') - \Phi(s), \tag{8.5}$$

for all states s, actions a, and successor states s'. Furthermore, the resulting pseudo-rewards f can be shifted and scaled without changing the optimal policy, because linear transformations of potential-based pseudo-rewards are also potential-based, that is

$$a \cdot f(s, a, s') + b = \gamma \cdot \Phi'(s') - \Phi'(s), \tag{8.6}$$

$$\text{for } \Phi'(s) = a \cdot \Phi(s) - \frac{b}{1 - \gamma}. \tag{8.7}$$

If gamification is to help people achieve their goals, then the pseudo-rewards added in the form of points or badges must *not* divert people from the best of course of action but make its path easier to follow. Otherwise, gamification would lead people astray instead of guiding them to their goals. Hence, the practical significance of the shaping theorem is that it gives the architects of incentive structures a method to rule out incentivizing counter-productive behaviors:

1. Model the decision environment as an MDP.

2. Define a potential function Φ that specifies the value of each state of the MDP.

3. Assign points according to Equation 8.5.

This method could thus be used to avoid some of the dark sides of gamification (Callan et al., 2015; Devers & Gurung, 2015).

While the shaping theorem constrains pseudo-rewards to be potential-based there are infinitely many potential functions one could choose. Given that people's cognitive limitations prevent them from fully incorporating distant rewards (Huys et al., 2012; Myerson & Green, 1995), the modified reward structure $r'(s, a, s')$ should be such that the best action yields the highest immediate reward, that is

$$\pi_M^\star(s) = \arg \max_a r'(s, a, s'). \tag{8.8}$$

Here, we show that this can be achieved with our method by setting the potential function Φ to the optimal value function $V_M^\star$ of the decision environment M, that is

$$\Phi^\star(s) = V_M^\star(s) = \max_\pi V_M^\pi(s). \tag{8.9}$$

First, note that the resulting pseudo-rewards are

$$f(s, a, s') = \gamma \cdot V_M^\star(s') - V_M^\star(s), \tag{8.10}$$

which leads to the modified reward function

$$r'(s, a, s') = r(s, a, s') + \gamma \cdot V_M^\star(s') - V_M^\star(s). \tag{8.11}$$

Hence, if the agent was myopic its policy would be

$$\pi(s) = \arg\max_a \mathbb{E}\left[r(s, a, s') + \gamma \cdot V_M^\star(s') - V_M^\star(s)\right] \qquad (8.12)$$
$$= \arg\max_a \mathbb{E}\left[r(s, a, s') + \gamma \cdot V_M^\star(s')\right].$$

According to Equation 8.4, this is the optimal policy $\pi_M^\star$ for the original decision environment M. Thus, people would act optimally even if they were completely myopic. And they should perform equally well if they do optimal long-term planning to fully exploit the gamified environment M' or learn its optimal policy $\pi_{M'}^\star$ through trial-and-error, because the shaping theorem (Eq. 8.5) guarantees that the gamified environment M' has the same optimal policy, that is $\pi_{M'}^\star = \pi_M^\star$. This suggests that potential-based pseudo-rewards derived from $V_M^\star$ should allow even the most myopic agent that only considers the immediate reward to perform optimally. In this sense, the pseudo-rewards defined in Equation 8.10 can be considered optimal. In addition, the optimal pseudo-rewards accelerate learning as long as the agent's initial estimate of the value function is close to 0 (Ng et al., 1999).

Computing the optimal pseudo-rewards requires perfect knowledge of the decision environment and the decision-maker's preferences. This information may be unavailable in practice. Yet, even when the optimal value function $V_M^\star$ cannot be computed, it is often possible to approximate it. If so, the approximate value function $\hat{V}_M$ can be used to approximate the optimal pseudo-rewards (Eq. 8.10) by

$$\hat{f}(s, a, s') = \gamma \cdot \hat{V}_M(s') - \hat{V}_M(s). \qquad (8.13)$$

For instance, one can estimate the value of a state s from its approximate distance to a goal (Ng et al., 1999).

Here, we develop and evaluate a novel approach to helping people make better decisions. The basic idea is to automatically restructure decision-environments in such a way that people can identify the optimal course of action even if they can look only a single step ahead. To achieve this, our method leverages artificial intelligence to compute optimal pseudo-rewards and delivers them through game elements. Based on previous simulations (Ng et al., 1999), we predict that adding approximate pseudo-rewards (Eq. 8.13) improves people's decisions and that adding optimal pseudo-rewards is even more beneficial. We test these predictions in three behavioral experiments.

8.2 Experiment 1: Optimal reward structures

To determine which incentive structures are most conducive to good decisions, Experiment 1 applied optimal gamification to a difficult sequential decision-making task (Figure 8.2A).

8.2.1 Methods

We recruited 250 adult participants on Amazon Mechanical Turk. Participants received $0.50 and a performance-dependent bonus of up to $2 for playing 24 rounds of the game shown in Figure 8.2. In this game, the player receives points for routing an airplane along profitable routes between six cities. In each round, the initial location of the airplane is chosen at random. Participants then choose which of two possible destinations to fly to, receive the profit or loss of that flight, and choose the next flight until the game ends. Concretely, after each flight there was a 1 in 6 chance that the trial would end. Participants were instructed to score as high as possible, and their financial bonus was proportional to the rank of their score among all participants in their condition. This game is based on the planning task developed by Huys et al. (2012) and is isomorphic to a MDP with six states, two actions, deterministic transitions, and a discount factor of $\gamma = 1 - 1/6$. The locations correspond to the states of the MDP, the two actions correspond to flying to the first or the second destination available from the current location, the routes correspond to state-transitions, and the points participants received for flying those routes are the rewards. The current state was indicated by the position of the aircraft and was updated according to the flight chosen by the participant.

Participants were randomly assigned to one of four conditions (Table D.1 and Figure D.1): In the control condition, participants were shown the true transition and reward structure of this task, and their incentives were identical to the task's reward function $r(s, a, s')$ (Figure 8.2). This task was such that finding the optimal path required planning 4 steps ahead. By contrast, in the experimental conditions, the incentives shown to the participants ($r'(s, a, s')$) differed from the task's true reward function by the addition of pseudo-rewards $f(s, a, s')$, that is

$$r'(s, a, s') = r(s, a, s') + f(s, a, s'). \tag{8.14}$$

We evaluated three different kinds of pseudo-rewards: In the first experimental condition, the pseudo-rewards were derived from the optimal value function according to the shaping theorem (Eq. 8.10), rewarding or punishing each move according to how much it increases or decreases the expected long-term reward respectively. In this condition, looking only 1 step ahead was sufficient to find the optimal path. The second experimental condition used the approximate potential-based

pseudo-rewards based on the distance-based heuristic value function

$$\hat{V}_M(s) = \hat{V}_M(s^\star) \cdot \left(1 - \frac{\text{distance}(s, s^\star)}{\max_s \text{distance}(s, s^\star)}\right), \tag{8.15}$$

where the goal state $s^\star$ was *Smithsville*, $\hat{V}_M(s^\star) = 140$ was the highest immediate reward that can be achieved from there, and distance(a, b) is the minimum number of moves required to get from state a to state b. The resulting pseudo-rewards simplified planning but not as much as the optimal pseudo-rewards. Finding the optimal path required planning 2-3 steps ahead and the immediate losses were smaller. In the third experimental condition, the pseudo-rewards violated the shaping theorem: the pseudo-reward was $+50$ for each transition that reduced the distance to the most valuable state (i.e. *Smithsville*) but there was no penalty for moving away from it. These incentives deviate from the optimal pseudo-rewards in two important ways: 1) they do not punish moves that lead away from the goal, and 2) they ignore that some paths are more costly than others. Since the experimental manipulation only affected the flights' payoffs, participants were unaware of the pseudo-rewards in Experiment 1.

In all three experimental conditions, the pseudo-rewards were mean-centered by subtracting their average to keep the average reward constant; since mean-centering is a linear transformation this retained the guarantees of the shaping theorem (see Eq. D.2). The mean-centered pseudo-rewards were added to the rewards of the control condition (see Figure D.1A) yielding the modified rewards shown in Figure D.1B-D and Table D.1, and the flight map was updated accordingly.

Regardless of the incentives shown to the participants, we measured their performance according to the reward function $r(s', a, s')$ of the original task. The complete experiment can be inspected at http://cocosci.dreamhosters.com/mturk/falk/FlightPlanning/.

Condition	Smiths-		Jones-		Williams-		Browns-		Clarks-		Bakers-	
No PR	140	30	−30	−70	−30	−70	−30	30	−30	−70	−30	−70
Optimal PR	2	−76	2	−5	−12	2	−4	2	2	0	2	−42
Approx. PR	8	−102	−22	−4	−22	−4	36	38	−34	−16	24	−32
Non-Potential-Based PR	119	9	−51	−41	−51	−41	−1	9	−51	41	−1	9

Table 8.1: Rewards in Experiment 1. The first entry of each cell is the (modified) reward of the counter-clockwise move and the second one is the (modified) reward of the other move.

INCLUSION CRITERIA.   The average completion time of the experiment was 13:37 min, and the median response time was 1.3 sec per choice. The relative score (i.e. $(R-r_{\min})/(r_{\max}-r_{\min})$ where R is the sum total of the player's points) was 79%. We excluded 3 participants who invested less than one
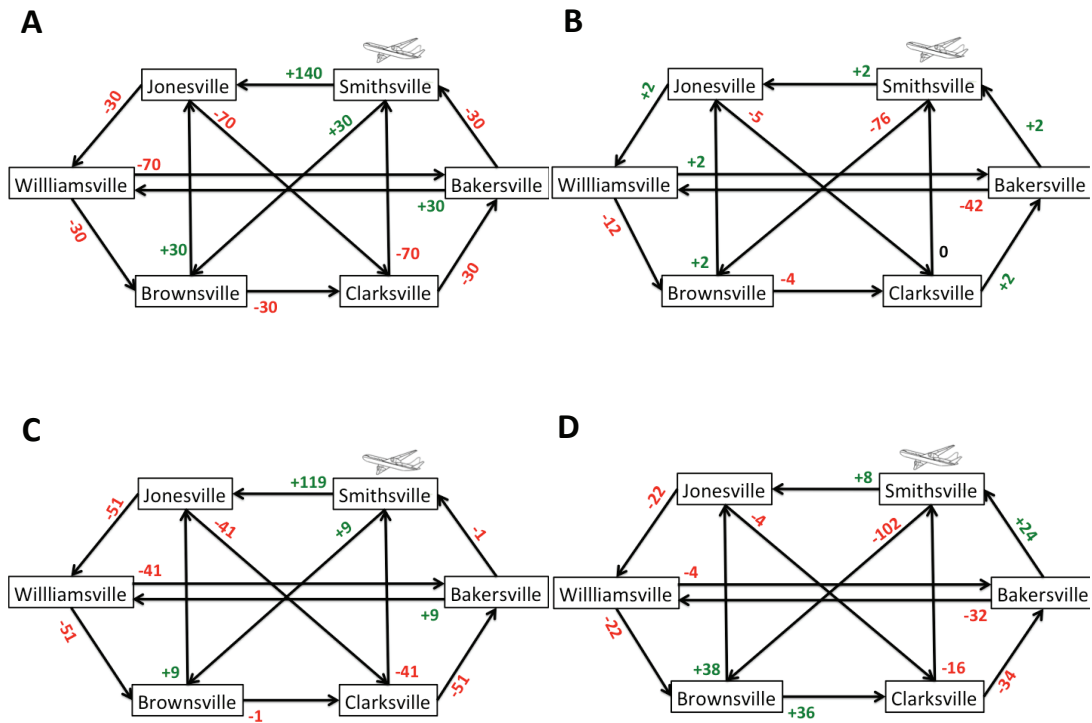
Figure 8.1: Conditions of Experiment 1. A: Control condition. B: Embedded pseudo-rewards. C: Separate pseudo-rewards. D: Integrated pseudo-rewards.

third of the median response time of their condition and 11 participants who scored lower than 95% of all participants in their condition. This led to the exclusion of 11 additional participants (5.5%), leading to a total exclusion rate of 7% ($14/200$).

8.2.2 RESULTS

A Kruskal-Wallis ANOVA revealed that the type of pseudo-rewards added to the reward function significantly affected people's performance in the original MDP ($H(3) = 40.35, p < 10^{-8}$; see Figure 8.2B). As expected, the unaided participants in the control condition performed very poorly attaining a median *loss* of 18.75 points per trial. Aiding participants with optimal pseudo-rewards led to significantly better performance ($Z = 4.76, p < 10^{-5}$) enabling them to achieve a median *gain* of $+5.00$ points/trial. Potential-based pseudo-rewards derived from an approximate value function also improved people's performance ($Z = 2.86, p = 0.0042$) but not as much as optimal pseudo-rewards ($Z = 2.68, p = 0.0074$). By contrast, the non-potential-based pseudo-rewards failed to improve people's performance ($Z = 0.72, p = 0.47$). In addition, optimal pseudo-rewards also accelerated the decision process (Figure D.3) supporting the conclusion that optimal gamification simplifies decision problems. Inspecting the four groups' choices frequencies revealed that the optimal pseudo-rewards significantly changed the choice frequencies in each of the six states and successfully nudged participants to follow the optimal cycle *Smithsville* $\rightarrow$ *Jonesville* $\rightarrow$ *Williamsville* $\rightarrow$ *Bakersville* $\rightarrow$ *Smithsville* (see Figure D.4).

REACTION TIMES.   A Kruskal-Wallis ANOVA revealed that the type of pseudo-rewards added to the reward function significantly affected people's reaction times ($H(3) = 29.96, p < 10^{-5}$). Given that the pseudo-reward type had a significant effect, we performed pairwise Wilcoxon rank sum tests to compare the medians of the four conditions (see Figure D.3). Optimal pseudo-rewards decreased the median response time from 1.72 to 1.14 sec. per decision ($Z = -4.19, p < 0.0001$), and non-potential-based pseudo-rewards decreased it to 1.12 sec. per decision ($Z = -3.38, p = 0.0007$). People in the condition with approximate potential-based pseudo-rewards took about the same amount of time as people in the control condition (1.65 sec.; $Z = -0.28, p = 0.78$).

EFFECT OF PSEUDO-REWARDS ON CHOICE FREQUENCIES.   The optimal strategy for this experiment was to take the counter-clockwise moves around the circle in all states except *Williamsville* and *Brownsville* (see Figure D.1A). Importantly, at *Williamsville* the optimal policy incurs a large

A Trial 3/24

Location: Jonesville   Flight 9   20 Points

Willliamsville: -30   Clarksville: -70
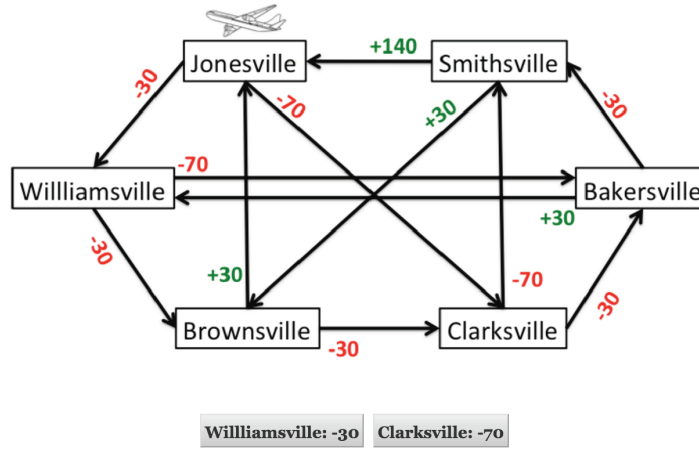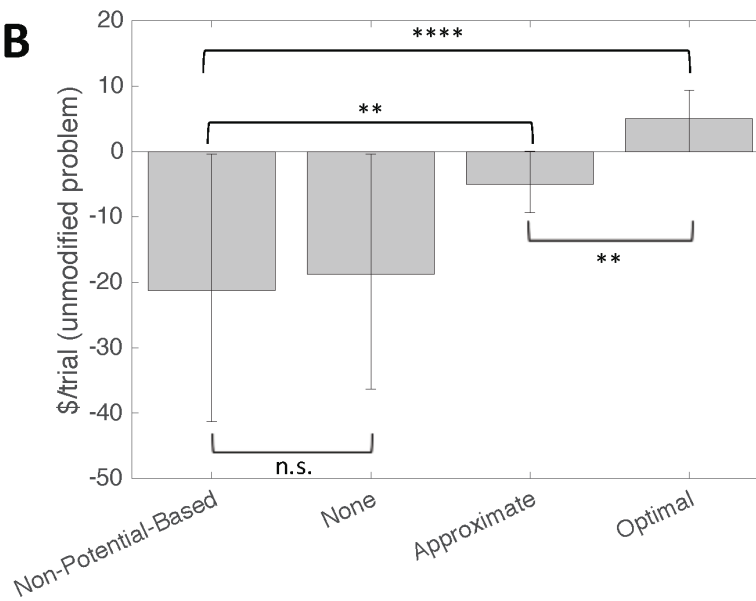
B

Figure 8.2: A: Task in Experiment 1: Control condition without pseudo-rewards. B: Median performance in Experiment 1 with 95% confidence intervals.
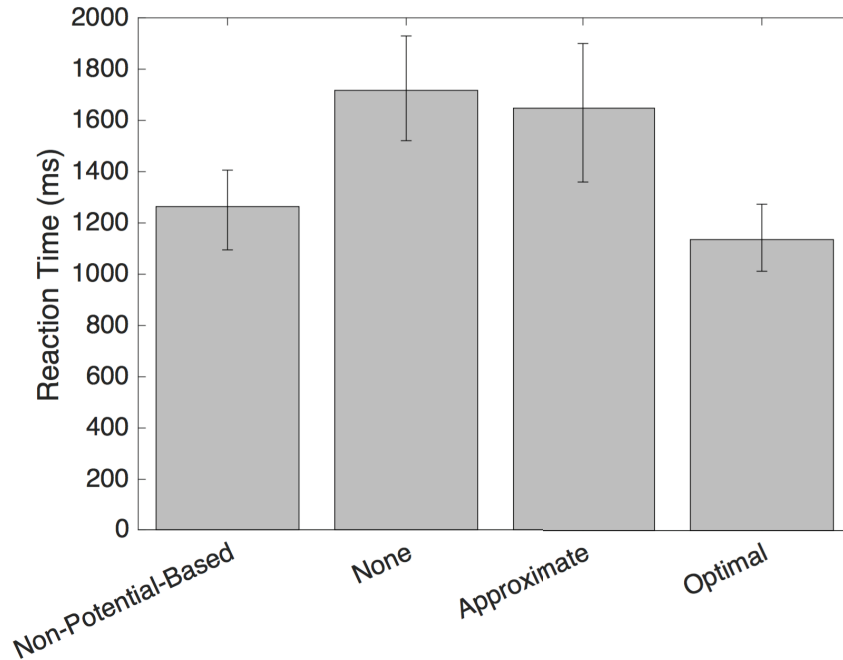
303

Figure 8.3: A: Median reaction times in Experiment 1 with 95% confidence intervals.

immediate loss, and no other policy achieves a positive reward rate. The optimal pseudo-rewards significantly changed the choice frequencies in each of the six states and successfully nudged participants to follow the optimal cycle *Smithsville* $\rightarrow$ *Jonesville* $\rightarrow$ *Williamsville* $\rightarrow$ *Bakersville* $\rightarrow$ *Smithsville* (see Figure D.1A). Their strongest effect was to eliminate the problem that most people would avoid the large loss associated with the correct move from *Williamsville* to *Bakersville* ($\chi^2(2) = 1393.8, p < 10^{-15}$). The optimal pseudo-rewards also increased the frequency of all other correct choices along the optimal cycle, that is the decisions to fly from *Bakersville* to *Smithsville* ($\chi^2(2) = 326.5, p < 10^{-15}$), from *Smithsville* to *Jonesville* ($\chi^2(2) = 7.9, p = 0.0191$), and from *Jonesville* to *Williamsville* ($\chi^2(2) = 299.8, p < 10^{-15}$). In addition, the optimal pseudo-rewards increased the frequency of the correct move from *Clarksville* to *Bakersville* ($\chi^2(2) = 92.0, p < 10^{-15}$). The only negative effect of the optimal pseudo-rewards was to slightly increase the frequency of the suboptimal move from *Brownsville* to *Clarksville* ($\chi^2(2) = 13.2, p = 0.0013$). By contrast, the non-potential-based pseudo-rewards misled our participants to follow the unprofitable cycle *Jonesville* $\rightarrow$ *Clarksville* $\rightarrow$ *Smithsville* $\rightarrow$ *Jonesville* by raising the frequency of the reckless moves from *Jonesville* to *Clarksville* ($\chi^2(2) = 1578.6, p < 10^{-15}$) and from *Clarksville* to *Smithsville* ($\chi^2(2) = 813.7, p < 10^{-15}$). The effect of the approximate pseudo-rewards was
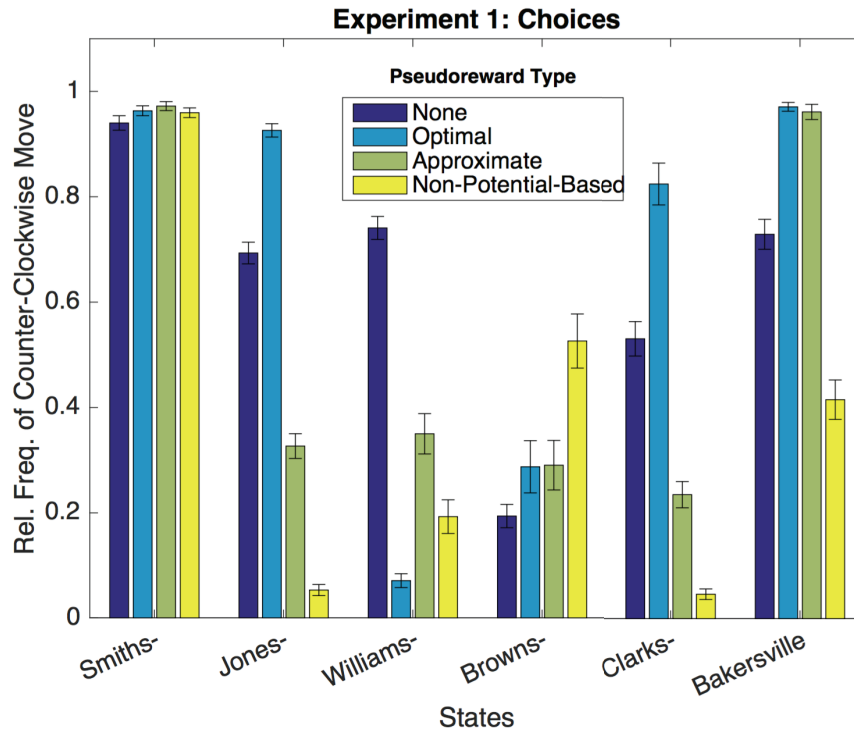
Experiment 1: Choices

Figure 8.4: Choice frequencies in each state of Experiment 1 by condition. Error bars enclose 95% confidence intervals.

beneficial in *Smithsville*, *Williamsville*, and *Bakersville*, but negative in *Jonesville*, *Brownsville*, and *Clarksville* (see Figure D.4). This explains why only potential-based pseudo-rewards had a positive net-effect on performance (Figure 1B in the Main Text).

8.3 Experiment 2: Conveying incentives with game elements

Given that optimal pseudo-rewards can significantly improve people's performance, we asked how they should be presented. In Experiment 1 pseudo-rewards were embedded directly into the reward structure of the decision environment, but this may be impossible to implement in the real world. Instead, a real-world application could convey pseudo-rewards through game elements. To evaluate the effectiveness of this presentation format, we augmented the task from Experiment 1 with game mechanics that conveyed the pseudo-rewards through stars and badges (see Figure 8.6A and Appendix D).
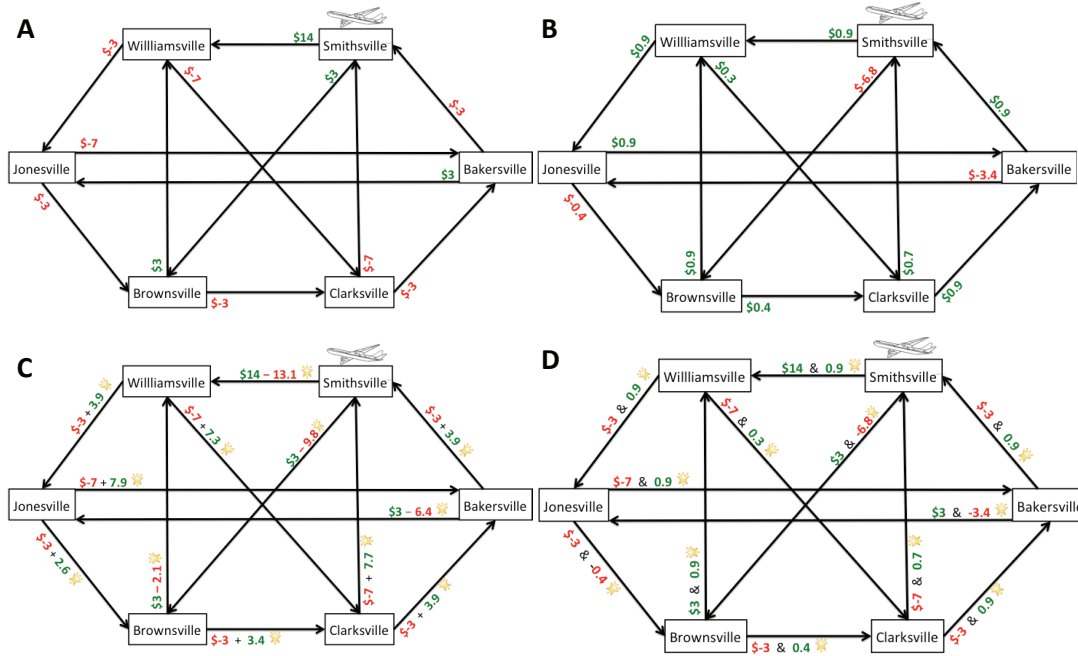
Figure 8.5: Conditions of Experiment 2. A: Control condition. B: Embedded pseudo-rewards. C: Separate pseudo-rewards. D: Integrated pseudo-rewards.

8.3.1 Methods

We recruited 400 participants on Amazon Mechanical Turk and paid them $2.50 for about 20-25 minutes of work plus a performance dependent bonus of up to $2. The average value of the bonus was $1. The median completion time of the experiment was 21.2 minutes.

The task was equivalent to the one used in Experiment 1 except that all rewards were scaled down by a factor of 10 to keep the arithmetic operations required to solve the task simple. Optimal pseudo-rewards were computed according to Equation 8.10 and shifted by the expected return of the optimal policy. This ensured that, on average, the sum of the immediate reward and pseudo-reward for the optimal action was equal to the expected long-run reward of the optimal strategy. This is appealing because it predicts exactly how much money the player will earn in the long-run if they act optimally. In the control condition pseudo-rewards were not presented at all (Figure D.2A). Three experimental conditions presented the optimal pseudo-rewards in three different formats: In the first experimental condition, the pseudo-rewards were embedded into the decision environment by adding them directly onto the flights profits and losses (Figure D.2B). In the second experimental

condition, the pseudo-rewards were presented separately from the monetary rewards in the form of stars (Figure D.2B). In the third experimental condition the number of stars communicated the sum of the shifted optimal pseudo-reward and the immediate reward. In all conditions, each flight's payoff and number of stars were rounded to one significant digit. In the conditions with stars participants were informed that the stars were designed to help the pilots make better, less short-sighted decisions. The instructions explained the meaning of the stars: In the second experimental condition, participants were told that the difference in the number of stars awarded for flying to destination A versus B predicts the difference in the amount of money that can be earned from there onward in the long run. In addition, these participants were given the tip that the flight with the highest sum of stars plus dollars is most profitable in the long run. In the third experimental condition participants were told that the difference between the number of stars awarded for flying to destination A versus B predicted the difference in how much profit they were going to make in the long run if they chose destination A over destination B. Participants in this condition were given the tip that they could earn the most by always flying the route with the larger number of stars. The stars had no monetary value, but they determined the player's level in the game.

Game mechanics. The character played by the participant could rise from *Trainee* to *ATP senior captain* via 15 intermediate levels. The number of points required to reach the next level increased according to the difficult curve proposed by Bostan and Öğüt (2009). Whenever the player reach the next level a congratulatory message was shown. In addition, participants were told how many stars and dollars were required to reach the next level in the game. To make the levels salient the pilot's shoulder badge was shown in the top right corner of the screen, and a feedback message was shown whenever the character was promoted and earned a badge or was demoted and lost a badge. The player started the game with +$50 so that their balance would remain positive as they learned to play the game.

The complete Experiment can be inspected at http://cocosci.berkeley.edu/mturk/falk/PNASExp2/index.html

Attention checks and inclusion criteria. To start the experiment participants had to pass a quiz comprising three questions on how their financial bonus would be determined and three questions testing their understanding of the mechanics of the task. Out of the 400 participants, 335 had not participated in any of our previous flight planning experiments and were included for in this study. Out of those 335 participants, we excluded subjects whose median response time was less

than one third of the median response time across all included subjects. In addition, we excluded the 5% of participants with the lowest scores of each group. This led to the exclusion of 19 out of the 335 included participants (5.7%), leaving 316 participants with about 80 participants per condition: 80 in the control condition, 81 in the condition with embedded pseudo-rewards, 81 in the condition with separate pseudo-rewards, and 74 in the condition with integrated pseudo-rewards.

8.3.2 RESULTS

The results of this experiment replicated the finding that pseudo-rewards significantly improve people's performance ($Z = 3.43, p = 0.0006$). Furthermore, the results suggested that presenting integrated pseudo-rewards in the form of stars could be just as effective as directly modifying the reward structure of the environment: Integrated pseudo-rewards significantly increased people's performance from -0.73 dollars/trial to $+0.17$ dollars/trial ($Z = 3.69, p = 0.0002$) which was not significantly lower than the performance of the group presented embedded pseudo-rewards (0.42 dollars/trial, $Z = 0.52, p = 0.62$). Presenting pseudo-rewards in this integrated format was critical to their effectiveness, since presenting them separately failed to significantly increase people's performance (median performance: -0.5 dollars/trial; $Z = 0.22, p = 0.83$). Inspecting participants' choice frequencies revealed that the three presentation formats had significantly different effects on people's decisions (see Appendix D and Figure D.5). In summary, incentivizing good decisions with game elements can be as effective as redesigning the decision environment, and this approach is most effective when the game elements make it very easy for people to identify the best course of action.
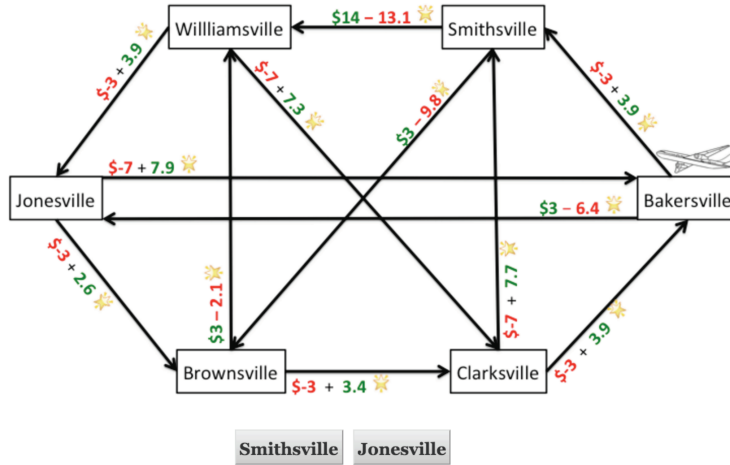
EFFECT OF PRESENTATION FORMAT ON RESPONSE TIMES AND CHOICE FREQUENCIES.   Participants were significantly faster when pseudo-rewards were embedded in the decision environment than when they were presented separately ($Z = -4.06, p < 0.0001$) or in the integrated format ($Z = -2.78, p = 0.0053$). Figure D.5 shows people's choice frequencies for each state depending on the experimental condition. Compared to separately presented pseudo-rewards, embedded pseudo-rewards were significantly more beneficial in all 6 states (all $p \leq 0.0218$) as were integrated pseudo-rewards (all $p \leq 0.0023$) but separately presented pseudo-rewards were never advantageous to either embedded or integrated pseudo-rewards. Embedded pseudo-rewards were more beneficial than integrated pseudo-rewards in 2 states (all $p \leq 0.0001$); conversely integrated pseudo-rewards were more beneficial than embedded pseudo-rewards in 1 state ($p < 10^{-9}$).
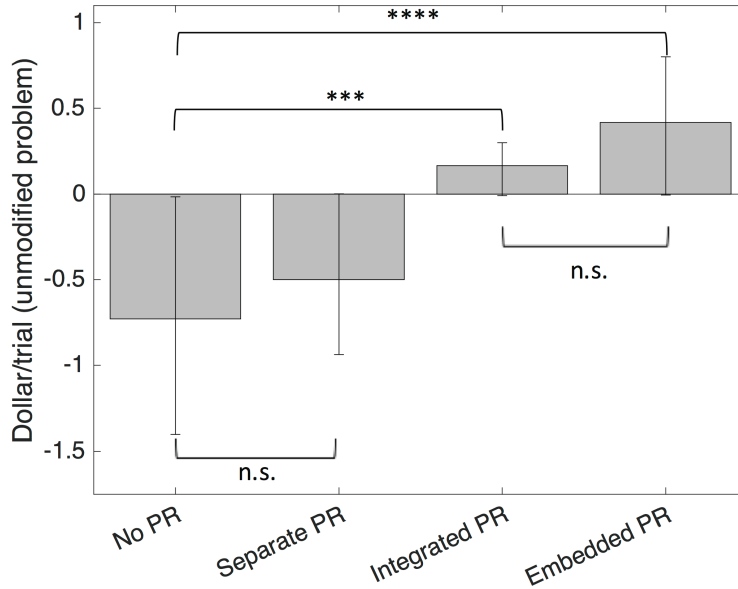
Figure 8.6: A: Pilot game with separately presented pseudo-rewards. B: Median performance in Experiment 2 by condition with 95% confidence intervals.
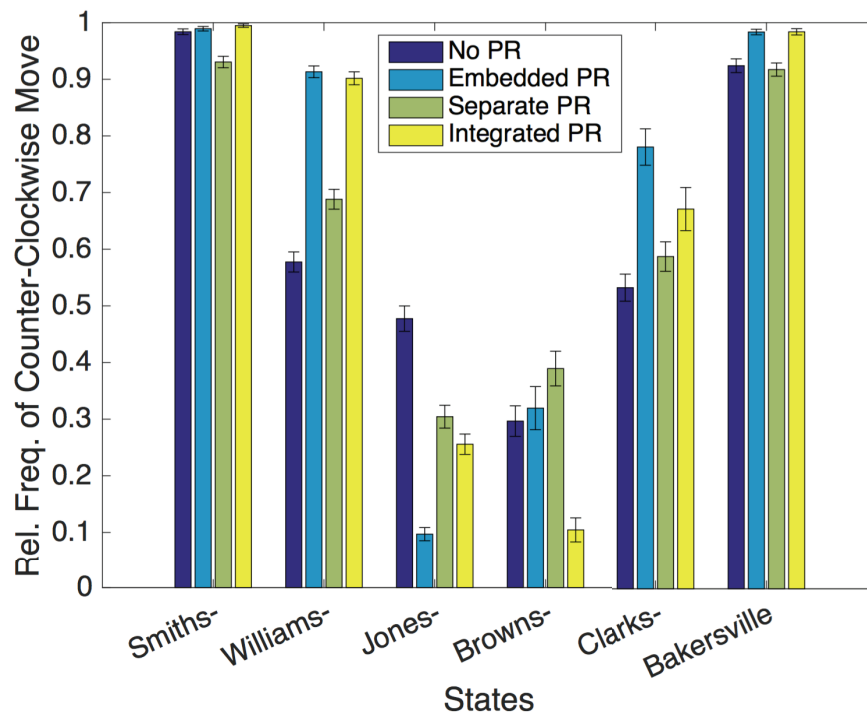
Figure 8.7: Choice frequencies in each state of Experiment 2 by condition. Error bars enclose 95% confidence intervals.
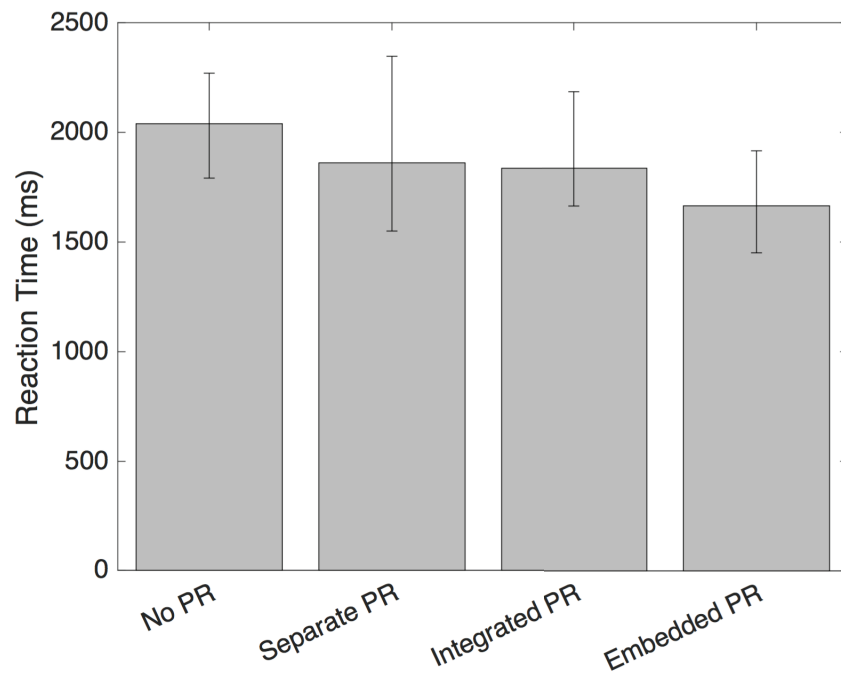
Figure 8.8: Median reaction times in Experiment 2 with 95% confidence intervals.

Follow-up experiment. The integrated pseudo-rewards differ from the separately presented pseudo-rewards in two respects: First, they simplify the decision process by allowing people to base their decision on a single signal. Second, they shift the pseudo-rewards such that the pseudo-reward for the optimal action is always positive. To tease apart the contributions of these two factors, we ran a follow-up experiment in which the separately presented pseudo-rewards were shifted such that the minimum pseudo-reward for an optimal action was the expected return of the optimal policy as it was for the integrated pseudo-rewards.

We recruited 339 participants on Amazon Mechanical Turk. Each participant was randomly assigned them to one of three conditions: no pseudo-rewards, shifted separately presented pseudo-rewards, and integrated pseudo-rewards. Condition 1 and 3 were identical to the equivalent conditions in Experiment 2. In the second condition, the optimal pseudo-rewards were shifted such that the minimum pseudo-reward for taking an optimal action was the expected reward rate of the optimal policy, that is 0.9. In all other regards, this follow-up experiment was identical to Experiment 2.

The median completion time was 23.35 minutes. We excluded 24 participants who had participated in previous flight planning experiments and 15 participants who performed worse or responded faster than 95% of the participants in their condition. Out of the remaining 290 participants 102 were in the condition without pseudo-rewards, 91 were in the condition with shifted separately presented pseudo-rewards, and 97 were in the condition with integrated pseudo-rewards.

We found that the shifted separately presented pseudo-rewards were significantly less effective than the integrated pseudo-rewards ($Z = -2.38, p = 0.0172$) and did not significantly improve people's performance relative to the control condition ($Z = 0.17, p = 0.8617$; median loss: \$11 vs. \$11 in the control condition; see Figure D.7). By contrast, participants in the condition with integrated pseudo-rewards performed significantly better than participants in the condition without pseudo-rewards ($Z = 2.46, p = 0.0140$; median loss: \$0 vs. \$14.50). Therefore, the primary benefit of the integrated pseudo-rewards appears to be that they simplify the decision process by offloading the computation of adding rewards and pseudo-rewards from the participants.

8.4   Experiment 3: To-do list gamification

Given that optimal gamification enabled people to act more farsightedly in the laboratory tasks of Experiments 1 and 2, we hypothesized that it could be used to help people overcome the myopic
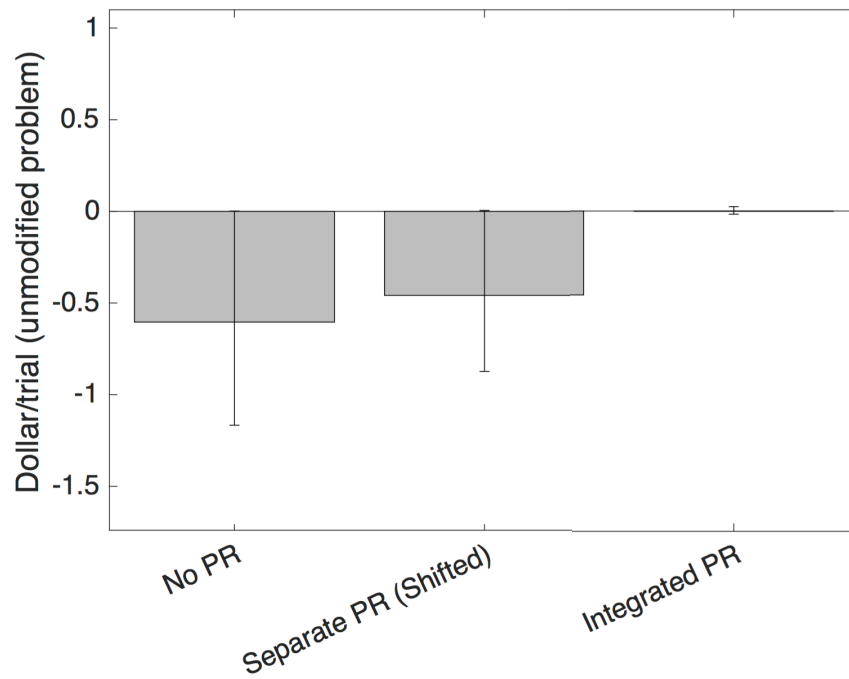
Figure 8.9: Performance in Experiment 2b by condition. Error bars enclose 95% confidence intervals.

biases that give rise to procrastination in everyday life. To test this hypothesis, we designed a third experiment where participants could earn a $20 bonus by completing six daunting writing assignments by a distant deadline.

8.4.1 Methods

PILOT STUDY AND TASK SELECTION. To select a suitable set of tasks for Experiment 3 we ran a pilot study that acquired subjective ratings of 21 candidate tasks. 100 participants recruited on Amazon Mechanical Turk evaluated 5 tasks each and were paid $0.50 in return. For each task, they estimated the fair price that should be paid for the task on Amazon Mechanical Turk and its duration. In addition, they rated the task's difficulty, their willingness to complete it for the price they had indicated, its enjoyableness compared to a typical MTurk HIT, its relative unpleasantness compared to a typical HIT an MTurk, and how likely they would be to postpone it on nine point Likert scales with appropriate anchors. We selected the 4 tasks that participants said they would be most likely to postpone and the task they said they were least likely to postpone. This procedure led to the selection of the following five writing assignments shown in Table D.2. Each assignment required that participants write at least 100 words (assignments 1-4) or at least 50 words (assignment 5).

PARTICIPANTS AND PROCEDURE For the main experiment, we recruited 120 participants by posting a sign-up form on Amazon Mechanical Turk. The sign-up form told potential participants that the study would comprise the five writing assignments shown in Table D.2. The sign-up form was posted on Monday, April 24 2017 and the deadline was at midnight on Wednesday of the following week (i.e., May 3rd 2017). The sign-up form can be inspected at cocosci.berkeley.edu/mturk/falk/ToDoListStudyPart1Writing. We informed potential participants that they could earn a bonus of $20 by completing all five assignments by a deadline 10 days later whereas missing the deadline would yield only $3 for each hour's worth of completed tasks. Potential participants could choose to either sign-up for the experiment and receive an immediate compensation of $0.05 or forego the opportunity to participate and receive $0.15. If they chose to sign-up for the second part of the study they were shown the link to the website hosting the to-do list experiment and asked to create an account or bookmark it.

The main experiment presented participants with a to-do list of five writing assignments (see Figure 8.10). Participants typed their answer into a text box on the to-do list website[†]. To tempt

---

[†]https://todo-list-study.herokuapp.com/

Writing Assignment	Fair Price	Duration	propensity to postpone	Minimum Length
How has North Korea's economic policy changed since the 1950s? What are the reasons and implications of these changes?	$3	15min	6.6/9	100 words
Please analyze the causes and implications of the British exit referendum in June 2016.	$3.25	25min	6.3/9	100 words
Describe with examples the importance of recognizing and responding to concerns about children and young people's development.	$2.25	20min	6.2/9	100 words
Write an essay about how society should assign value to human life.	$3	27.5min	6.1/9	100 words
What is your favorite TV show and why?	$1	7min	2.8/9	50 words

Table 8.2: Writing assignments and their ratings

Figure 8.10: Screenshot from Experiment 3.

participants to procrastinate, the to-do list website displayed a series of distracting links to Youtube videos, Reddit articles, news stories, or the game of Tetris.

Upon creating an account on the to-do list website, each participant was assigned to one of four conditions: In the first experimental condition, the incentives for completing each assignment were conveyed as points (see Figure 8.10), and the participant's total number of points determined their level in the game. The second experimental condition was like the first one except that the optimal pseudorewards were displayed as dollars rather than points. The first control group received no incentives, and in the second control condition the number of points was constant across all tasks. The incentives were saliently displayed next to each entry of the participant's to-do list, and the level and current number of points were saliently displayed above the current task (see Figure 8.10). The optimal pseudo-rewards were computed by applying the optimal gamification method described above to a finite-horizon MDP model of the experiment. This model comprised one action for each task and an additional action for taking a break. The reward function was set up such that each task-action incurred a cost that reflected the task's fair wage as determined in the pilot study described above. Finishing the experiment earns an additional reward of \$20. In the MDP model of the experiment, taking a break earns a reward equivalent to \$0.50 but also comes with a 2.5% chance of forgetting about the tasks. The benefit of finishing the experiment sooner rather than later was captured by a discount factor of $\gamma = 0.95$.

When a participant completed all tasks, they were shown a bonus code. After the deadline, we posted a reimbursement HIT on Amazon Mechanical Turk that included an exit survey. The exit survey asked participants to rate their motivation to complete the tasks and how rewarding it felt to complete the second task on 9-point Likert scales. In addition, the exit survey also recorded age and self-identified gender and inquired if the participant had used any strategies to stay engaged, and which components of the website they found helpful. We posted a separate reimbursement HIT for participants who decided to quit the experiment was posted on the first day of the experiment, and it also included an exit survey.

8.4.2 RESULTS

As shown in Figure 8.11, optimal pseudo-rewards significantly increased the completion rate from 56.1% in the control conditions to 85.2% in the experimental conditions with optimal pseudorewards ($\chi^2(1) = 11.20, p = 0.0008$). This benefit cannot be explained by the mere presence of incentives or game elements because adding constant point values failed to increase the completion

rate (53.6% with constant points vs. 58.6% without points, $\chi^2(1) = 0.15, p = .70$). Framing optimal PRs in terms of money led to a completion rate of 92.3% while presenting them as points led to a completion rate of 78.6% but this difference was not statistically significant ($\chi^2(1) = 2.01, p = 0.16$). Along with the increase in the completion rate, the average number of completed assignments increased from 2.61 out of 5 without optimal gamification to 4.37 out of 5 ($\chi^2(1) = 17.99, p < 0.0001$), and the average total number of words written by each participant increased from 408.25 ± 55.54 to 765.46 ± 71.45 ($\chi^2(1) = 16.19, p < 0.0001$).

Of the 40 participants who did not complete all tasks (33.9%), only 1 filled out the exit survey. We therefore cannot evaluate the effect of the pseudo-rewards on motivation and perceived reward per se. However, we can analyze its effect on participants who completed all tasks. The following analyses are therefore restricted to this biased subset of participants. Due to this selection bias the results have to be interpreted with caution. For the participants who completed all tasks neither motivation ($\chi^2(1) = 0.04, p = 0.84$) nor experienced reward ($\chi^2(1) = 0.14, p = 0.71$) were significantly affected by optimal gamification. Among these participants, optimal gamification also did not affect how long it took them to complete the tasks ($\chi^2(1) = 0.07, p = 0.79$) or the number of times they aborted a task ($F(76) = 0.27, p = 0.61$). While optimal gamification slightly increased the number of words written per assignment from 155 to 175, this difference was not statistically significant ($F(1, 81) = 1.33, p = 0.25$). Optimal gamification also had no statistically significant effect on the total length of the breaks that these participants took between tasks ($F(1) = 0.42, p = 0.52$) or the number of times that they played Tetris ($F(1, 76) = 0.16, p = 0.69$). Optimal gamification also did not affect how long it took them to submit their first assignment ($\chi^2(1) = 3.26, p = 0.07$), when they started working on it ($\chi^2(1) = 2.35, p = 0.13$), or the delay until the first time they opened an assignment ($\chi^2(1) = 2.18, p = 0.14$). These negative results suggest that the main effect of optimal gamification was to increase the probability that participants would start working on the first task from 59.65% to 87.04% ($\chi^2(1) = 11.01, p = 0.0009$), because regardless of gamification 95.1% of all participants who completed the first task went on to complete all of the tasks ($\chi^2(1) = 0.69, p = 0.41$) and their motivation and behavior appeared to be unaffected.

8.5 Discussion

The results of Experiments 1-3 suggest that optimal gamification can help people make better decisions, act more farsightedly, get started on daunting tasks, and become more productive. Our
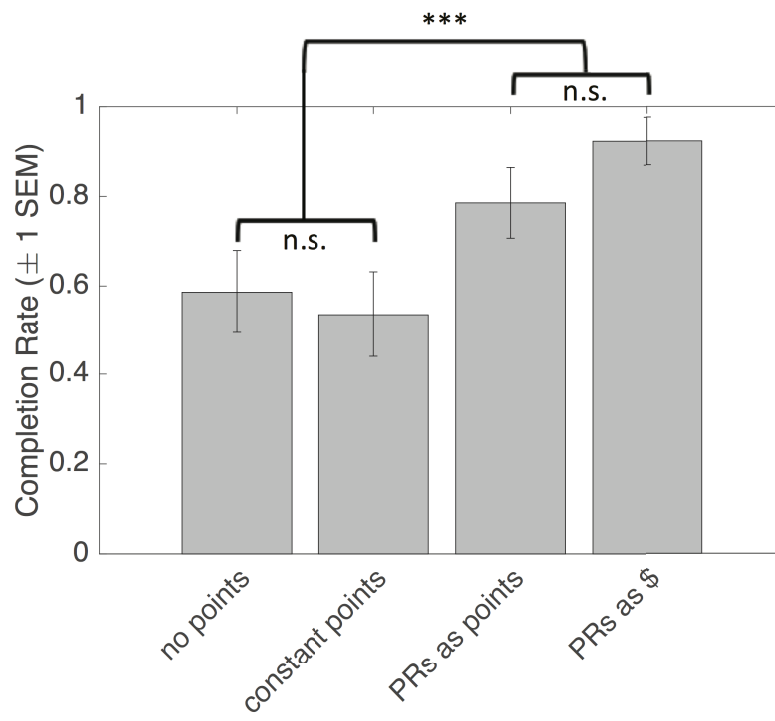
Figure 8.11: Results of Experiment 3: Proportion of participants who completed all assignments by the deadline with error bars showing ± 1 standard error of the mean.

method achieves this by leveraging artificial intelligence to solve sequential decision problems that are challenging for people and translating the solutions into incentives that align each action's immediate reward with its long-term value. The resulting incentive structures are implemented using game elements such as points and levels that motivate people to do what is best for them in the long run.

More generally, our results illustrate that AI can be used to automatically restructure decision problems in such a way that people's heuristics work well. This approach is in line with an extensive literature on bounded rationality that emphasizes that decision quality depends on the fit between people's heuristics and the structure of their environment (Chater & Oaksford, 1999; Griffiths et al., 2015; Oaksford & Chater, 1994; Todd & Gigerenzer, 2007, 2012).

There are already many decision support systems that solve Markov decision processes to compute optimal decisions and advise people to execute them (Aviv & Pazgal, 2005; Bhatnagar et al., 1999; Gadomski et al., 2001; Nunes et al., 2009; Song et al., 2000). However, research in psychology suggests that this approach to decision-support would likely undermine people's intrinsic motivation because it runs counter the fundamental human need for self-determination and autonomy (Gagné & Deci, 2005). Optimal gamification, by contrast, gives people complete freedom over what to do and can be applied to help people motivate themselves to take action towards their own goals. While using game elements to boost motivation is not a new idea, optimal gamification is unique in being based on a rigorous mathematical theory for determining which actions should be incentivized and by how much. This theory guarantees that optimal gamification will never incentivize counterproductive behavior (Ng et al., 1999). This avoids the perils of less principled approaches to motivating people with incentives and game elements (Callan et al., 2015; Devers & Gurung, 2015).

The results of Experiment 3 illustrate that optimal gamification can indeed help people to align their actions (e.g, whether or not to work on a writing assignment) with their long-term goal (e.g., to complete their tasks before the deadline to earn a financial bonus). The primary problem that optimal gamification solved in this setting was to help people overcome the motivational barriers of immediate effort that would only be rewarded much later. This suggests that optimal gamification might be useful for helping people overcome the myopic biases affecting their motivation (Steel & König, 2006), avoid self-control failure, and support the pursuit of long-term goals.

Beyond motivational issues, many decision problems that arise in the pursuit of long-term goals are simply too large and too complex for people to solve them optimally. Our approach could be used to overcome such challenges by augmenting people's bounded cognitive resources with the

power of computing and leveraging planning algorithms developed in artificial intelligence (Puterman, 2014) to build the solution of complex decision problems into the reward structure of the environment. Future work will investigate these hypotheses and explore optimal gamification as an interface between artificial and human intelligence. By integrating the power of computing with psychological insight into human motivation and decision-making, this line of research could lead to a new generation of cognitive prostheses that might significantly enhance human productivity and self-mastery. Our approach illustrates how advances in artificial intelligence can be leveraged to enhance human intelligence and overcome the cognitive limitations that hold people back from realizing their full potential. In this way, the continuing progress in artificial intelligence could enable a parallel growth in human intelligence.

9

Developing an intelligent system that teaches people optimal cognitive strategies[*]

The to-do list gamification app presented in Chapter 8 improved the people's decisions by adding incentives in such a way that their short-sighted decision mechanisms led to better choices. This approach combines the idea of nudging (Thaler & Sunstein, 2008), that is to restructure the environment in such a way that people's heuristics lead to better choices, with the idea of compensating for cognitive limitations by augmenting the human mind with external computational resources. In this chapter, we explore a third approach to expanding the bounds on human rationality: teaching people to think and decide as well as they possibly could. This is a form of boosting (Hertwig & Grüne-Yanoff, 2017).

Previous attempts to mitigate people's cognitive biases through education have been mostly unsuccessful (Larrick, 2002). I postulate that the main reason for these failures was that the curriculum was based on the unrealistically high standards of probability theory, logic, and expected utility

theory. As the research presented in Chapters 1-3 and Chapter 6 shows, adopting a more realistic normative standard that takes people's cognitive limitations into account leads to radically different prescriptions. In fact, as Chapters 2-3 show, rational heuristics often generate the very biases that previous approaches to improving judgment and decision-making sought to eliminate (Larrick, 2002). The strategy discovery method presented in Chapter 6 can be used to develop a more appropriate curriculum of strategies for good thinking and decision-making. I postulate that interventions based on this refined curriculum would be much more effective than interventions based on the standard picture of rationality. As a proof-of-concept, we derived a resource-rational planning strategy and evaluated the effectiveness of teaching it to people.

I believe that teaching people how to think can be accomplished without one-on-one instruction from a human teacher. Instead, we might be able to deliver high-quality training and instruction in rational thinking and decision-making via a fully-automated machine teaching system that is freely available via the internet. As a first step in this direction, we developed a web-based intelligent tutoring system for teaching people near-optimal planning strategies. The remainder of this chapter describes the underlying theoretical principles, the implementation of the system, and a series of experiments that assess the resulting learning gains as well as their transfer and retention.

9.1 THEORETICAL APPROACH

In Chapter 5, we found that people can learn to make more rational use of their limited cognitive resources via metacognitive reinforcement learning. Specifically, the findings presented in the second section of Chapter 5 suggested that people learn the value of alternative cognitive operations according to a model-free reinforcement learning mechanism. Viewing cognitive growth as a form of model-free reinforcement learning suggests that methods developed to accelerate model-free reinforcement learning in robots can be leveraged to accelerate metacognitive learning in humans. One such method is reward shaping (Ng et al., 1999). The basic idea of reward-shaping is to align each action's immediate reward more closely to its true long-term value. To accomplish this reward shaping adds additional rewards–called *pseudo-rewards*–on top of the rewards provided by the task environment. Critically, the pseudo-rewards should to be designed in such a way that the optimal policy does not change. Historically, designing good pseudorewards was a tricky problem, because intuitive incentive schemes can often be gamed by counterproductive behavior. For instance, rewarding a robot to touch the ball seems to be a sensible approach to help it learn to dribble. Unfortunately, it changes the optimal policy so that it becomes optimal for the robot to fall down on the ball and
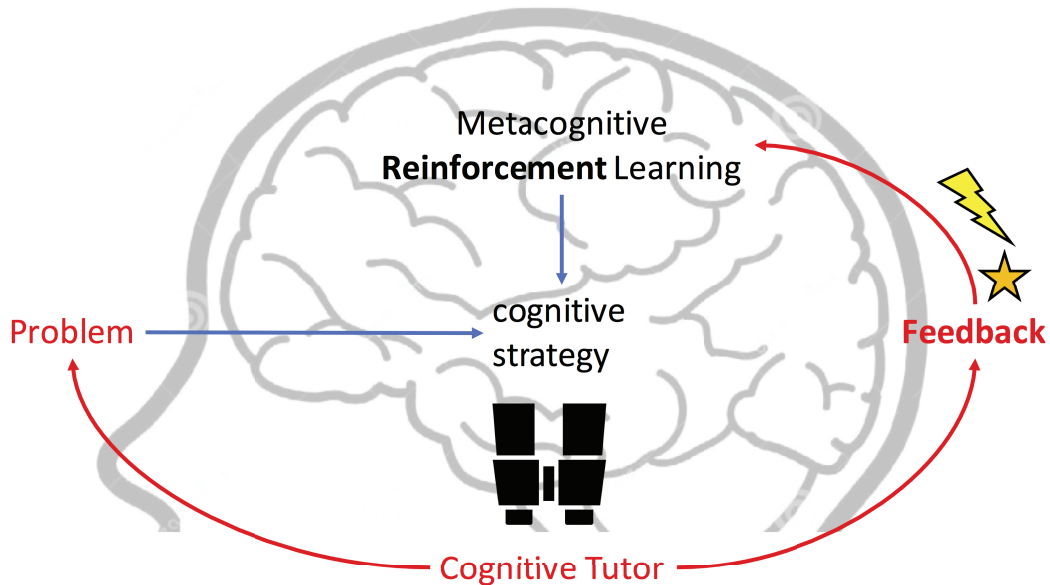
Figure 9.1: Illustration of the general idea behind the cognitive tutor presented in this chapter.

vibrate rapidly. Ng et al. (1999) proved that this problem can be avoided by first assigning a value $\Phi(s)$ to each state s and then computing each pseudo-reward as the difference between the potential $\Phi(s_{t+1})$ of the new state and the potential $\Phi(s_t)$ of the old state, that is

$$\text{PR}(s_t, a_t, s_{t+1}) = \gamma \cdot \Phi(s_{t+1}) - \Phi(s_t), \tag{9.1}$$

where γ is the discount factor of the MDP to be solved. For more details on the shaping theorem, see Chapter 8.

The cognitive training method developed in this chapter applies the shaping theorem (Ng et al., 1999) to design an optimal feedback mechanisms for cognitive training. The general idea is to provide immediate rewards that accurately communicate the long-term value of each cognitive operation. Such feedback would steer the metacognitive reinforcement learning mechanism identified in Chapter 5 towards the optimal cognitive strategy as quickly as possible. To assist people in how to change their strategy, the feedback should additionally include messages explaining what the optimal policy would have done instead. Figure 9.1 illustrates this pedagogical principle.

This line of thinking led me to develop the following approach to cognitive training:

1. Model the cognitive function to be improved (e.g., planning) and the available cognitive operations (e.g., simulating the outcome of taking a certain action in a certain state) and their

costs as a meta-level MDP M_{meta}.

2. Compute the values of the computations people might perform in different states (i.e., $Q_{\mathrm{meta}}(b, c)$) by solving the meta-level MDP either exactly or approximately.

3. Let people practice the cognitive function to be improved and infer their computations from process tracing data.

4. Score people's inferred computations by

$$\mathrm{score}(b, c) = \hat{Q}_{\mathrm{meta}}(b, c) - \max_c \hat{Q}_{\mathrm{meta}}(b, c). \tag{9.2}$$

5. Translate score into reinforcement and a feedback message.

The resulting reinforcement signal rewards people according to the expected sum of the immediate and long-term benefits of their actions minus the value of the previous belief state under the optimal strategy, that is

$$\mathrm{PR}_{\mathrm{opt}}(b_t, c_t) = \mathbb{E}\left[V^{\star}_{\mathrm{meta}}(B_{t+1}) + r_{\mathrm{meta}}(b_t, c_t, B_{t+1}) | B_t = b_t, C_t = c_t\right] - V^{\star}_{\mathrm{meta}}(b_t). \tag{9.3}$$

Adding these pseudo-rewards to the meta-level MPD retains the optimal meta-level policy according to the shaping theorem (Ng et al., 1999). This is because the expected difference in the state-value is the expected value of a potential-based pseudo-reward (Ng et al., 1999) and the addition of the expected immediate reward merely scales the expected rewards of all transitions by a factor of two.

In many cases solving the meta-level MDP exactly is intractable. In these cases, we can approximate the meta-level Q-function using the following four-step procedure:

1. Apply the strategy discovery method presented in Chapter 6 to compute a near-optimal meta-level policy π_{LC}.

2. Collect process tracing data from human participants solving the task without feedback.

3. Start the optimal strategy π_{LC} from all pairs (b, c) of a participant's information state b and the computation c they performed in that state.

4. Record the optimal strategy's meta-level return R from each of these starting points.

5. Use the resulting data set $\{(b, c)_t, R_t\}_{1 \leq t \leq T}$ to compute the maximum-likelihood estimate $\hat{\beta}_{\mathrm{ML}}$ of the parameters β of the regression model

$$Q_{\pi_{\mathrm{LC}}}(b, c) = \beta_0 \cdot r_{\mathrm{meta}}(b, \perp) + \beta_1 \cdot \mathrm{VOI}_1(b, c) + \beta_2 \tag{9.4}$$
$$\cdot \mathrm{VPI}_{\mathrm{all}}(b, c) + \beta_3 \cdot \mathrm{VPI}_{\mathrm{a}}(b, c) + \beta_4 \cdot \mathrm{cost}(c) + \varepsilon,$$

where the first regressor is the expected return of acting without further deliberation in the current state and the subsequent three features are defined in Chapter 6, and ε is a normally distributed error term. We can then use the maximum likelihood estimate $\hat{\beta}_{\mathrm{ML}}$ to approximate $Q_{\pi_{\mathrm{LC}}}$ by

$$\hat{Q}_{\pi_{\mathrm{LC}}} = \beta_0 \cdot r_{\mathrm{meta}}(b, \perp) + \hat{\beta}_1^{\mathrm{ML}} \cdot \mathrm{VOI}_1(b, c) + \beta_2^{\mathrm{ML}} \tag{9.5}$$
$$\cdot \mathrm{VPI}_{\mathrm{all}}(b, c) + \beta_3^{\mathrm{ML}} \cdot \mathrm{VPI}_{\mathrm{a}}(b, c) + \beta_4^{\mathrm{ML}} \cdot \mathrm{cost}(c).$$

9.2   A COGNITIVE TUTOR FOR PLANNING

The metacognitive feedback method presented above is very general and widely applicable. Here, we applied it to teaching people effective planning strategies via metacognitive feedback. Planning, like all of cognition, is a mental process that cannot be observed directly. Thus, to give people feedback on their planning strategy, we first have to make it observable. To do so, we developed a process tracing paradigm for the study of planning (Callaway, Lieder, Krueger, & Griffiths, 2017). Our Mouselab-MDP paradigm is inspired by the Mouselab paradigm (Payne et al., 1993) that traces how people choose between multiple risky gambles (see Chapter 6). The basic idea is to externalize people's mental simulations of alternative action sequences. To do so, the Mouselab-MDP paradigm presents participants with a route planning problem where each move earns or loses an initially unknown amount of money. The participant's goal is to choose a route in such as way that they earn as much money as possible. To find out how much money a move would yield, the participant has to click on it and pay a fee. Each click is recorded and the recorded sequence of clicks reveals which paths participants mentally simulated and in which order. As a proof-of-concept, this chapter develops a cognitive tutor that teaches people to plan backward from potential goals using the 3-step planning task shown in Figure 9.2.

To develop a cognitive tutor for planning, we applied Steps 1-5 of the methodology summarized above to the Mouselab-MDP paradigm.

In Step 1, we model the problem of deciding how to plan by the meta-level MDP

$$M_{\mathrm{meta}} = (\mathcal{B}, \mathcal{A}, \mathcal{T}, r_{\mathrm{meta}}), \tag{9.6}$$

where each belief state b encodes one Normal distribution for each transition's reward. Thus, the belief state $b^{(t)}$ at time t can be represented as $((\mu_1^{(t)}, \sigma_1^{(t)}), \cdots, (\mu_K^{(t)}, \sigma_K^{(t)}))$ such that $b^{(t)}(\theta_k = x) = \mathcal{N}(x; \mu_k^{(t)}, \sigma_k^{(t)2})$. The initial belief state $b^{(0)}$ encodes the joint distribution $\mathcal{N}\left(\mathbf{x}; (\mu_1^{(R)}, \cdots, \mu_K^{(R)}), \Sigma^{(R)}\right)$
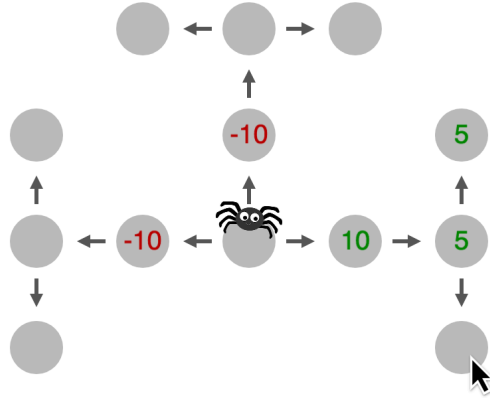
Figure 9.2: Illustration of the Mouselab-MDP paradigm.

that the rewards are sampled from. The metalevel actions are $\mathcal{A} = \{c_1, \cdots, c_K, \bot\}$ where c_k reveals the reward at location k and $\bot$ terminates planning and selects the path with highest expected sum of rewards according to the current belief state. The transition probabilities $T_{\mathrm{meta}}(b^{(t)}, c_k, b^{(t+1)})$ encode that performing computation c_k sets $(\mu_k^{(t+1)}, \sigma_k^{(t+1)})$ to $(x, 0)$ with probability density $\phi(x; \mu_k^{(t)}, \sigma_k^{(t)})$ where ϕ is the density function of the normal distribution. The metalevel reward function is $r_{\mathrm{meta}}(b, c) = -\lambda$ for $c \in \{c_1, \cdots, c_K\}$, and $r_{\mathrm{meta}}\left((\mu_1, \sigma_1), \cdots, (\mu_K, \sigma_K)\right), \bot) = \max_{\mathbf{t} \in \mathcal{T}} \sum_{k \in \mathbf{t}} \mu_k$ where $\mathcal{T}$ is the set of possible trajectories $\mathbf{t}$ through the environment.

For the second step, we solved this meta-level MDP exactly through backward induction (Puterman, 2014) with hashing. Using the Mouselab-MDP paradigm allows the cognitive tutor to automatically infer the participants' computations from their clicks by assuming that participants immediately update their belief state upon uncovering a new piece of information (Step 3). After each click a participant makes, the cognitive tutor scores the corresponding computation according to Equations 9.2 and 9.5 (Step 4).

In the final step, this score is translated into a penalty delay of round$(2 - \mathrm{score}(b_t, c_t))$ seconds, and the cognitive tutor displays a feedback message. If the participant made an error, then the feedback message informs them what they should have done differently (see Figure 9.3). Concretely, if the optimal action(s) involved clicking, then all optimal clicks are highlighted in blue. And when it is optimal to move without further deliberation, then the feedback message informs participants that the optimal strategy would not have performed any more clicks. When the participant's planning operation was optimal, then the feedback message says "Good job!" and they are allowed to proceed immediately.

The participant's score is updated after every click (by subtracting the cost per click) and move (by adding the collected reward). A timer enforces that each participant spends at least a required minimum amount of time on each trial. This serves to eliminate the opportunity cost of time, and as a result the cost of planning is entirely determined by the price of each click.

The following sections evaluate this cognitive tutor in a series of experiments: Experiment 1 assesses whether training with the cognitive tutor accelerates learning how to plan better compared to practicing without feedback. Experiment 2 tests whether the benefits of this training transfer to more difficult problems in more complex environments, and Experiment 3 tests whether those transferable benefits are retained over time.

9.3 Experiment 1: Metacognitive feedback accelerates learning to plan

To assess whether the cognitive tutor accelerates learning to plan, We employed a pre-post design where the intervening training block either gave participants the cognitive tutor's feedback or no feedback at all.

9.3.1 Methods

We recruited 119 participants on Amazon Mechanical Turk (average age 34.7 ± 9.8 years, range: 20–68 years). Participants received a base pay of $0.75 plus a performance dependent bonus of 1 cent for every $5 they earned in the test block (average bonus 1.34 ± 0.57). The average duration of the experiment was 13.4 ± 3.5 minutes. For the first six participants the condition was not recorded due to a technical error; these participants were therefore excluded from all subsequent analyses.

Each participant was assigned to be in either the experimental condition where participants practiced with feedback or the control condition where participants practiced without feedback. Counterbalancing the assignments to the two groups yielded 56 participants in the feedback group and 57 in the control group. The experiment was structured into instructions, a pretest, a training block, a post-test block, a quiz, and questions about the participant's strategies. The pretest block comprised 1 trial, the training block comprised 10 trials, and the post-test block comprised 20 trials.

Each trial presented participants with a 3-step planning problem with 3 choices in the first step, 1 choice in the second step, and 2 choices in the final step (see Figure 9.2). The participant's goal was to earn as much money as possible. Critically, the range of the attainable rewards increased

from the first step to the third step. Concretely, in the first condition the reward distributions were Uniform($\{-4, -2, +2, +4\}$), Uniform($\{-8, -4, +4, +8\}$), and Uniform($\{-48, -24, +24, +48\}$) for nodes reachable in one, two, and three steps, respectively. To operationalize the opportunity cost of planning, we charged participants \$1 per click. To eliminate the time cost of engaging in planning compared to speeding through the experiment, participants who spend less than 7 seconds on planning (e.g., only 3 seconds) were required to wait for the remaining time after executing their moves (e.g., for 4 seconds).

The instructions informed participants about how to move, how to collect information, the cost of information, the minimum time of 7 seconds per trial, the structure of the experiment, and the bonus. The quiz queried participants about the range of rewards in the first step, the range of rewards in the last step, the cost per click, and how the bonus would be calculated. Additional questions asked participants about how they decided where to click and where not to click, where they usually clicked first, their general strategy, what they had learned, and whether they found anything confusing.

In the post-test block participants were informed that they would receive a bonus of 1 cents for every \$5 they make in the game, and they received an endowment of 50 virtual dollars, which was worth 10 cents.

9.3.2 RESULTS

As shown in Figure 9.4, the cognitive tutor's feedback significantly accelerated people's learning. Concretely, a linear regression analysis confirmed that the feedback significantly increased the slope of participants' learning curve from 0.7710 \$/trial to 2.0994 \$/trial ($t(1305) = 2.461, p = 0.014$). Participants in the feedback condition improved significantly more from the pretest to the last trial of the training block than participants in the control condition (+22.6 \$/trial vs. +4.81 \$/trial, $t(111) = 2.22, p = 0.0281$). Consequently, participants who received feedback performed significantly better in the post-test than participants who practiced without feedback (36.2 \$/trial vs. 24.6 \$/trial, $t(2258) = 10.7, p < 0.0001$; see Figure 9.5).

To elucidate the source of these improvements, we compared participants' planning strategies in the post-test between the feedback condition and the control condition. As shown in Figure 9.6, we found that the proportion of trials on which participants started by inspecting a potential goal state was significantly higher in the feedback condition (98.6% of all first clicks) than in the control

Figure 9.3: Examples of feedback messages used by the cognitive tutor.
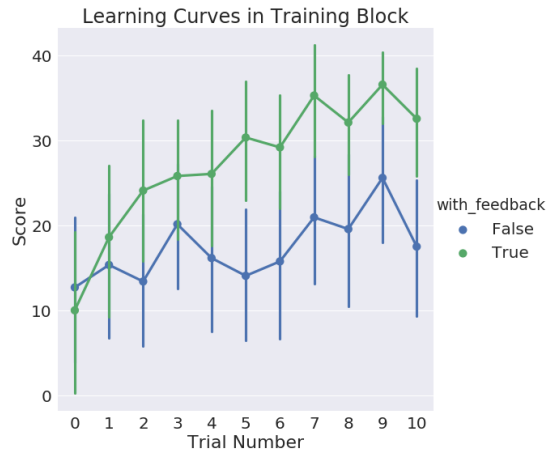
Figure 9.4: Optimal feedback accelerates learning. The error bars are 95% confidence intervals based on bootstrapping.
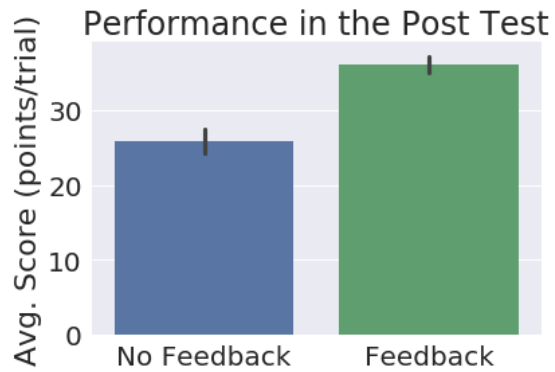
Figure 9.5: Optimal feedback increased performance in the test block. The error bars are 95% confidence intervals based on bootstrapping.
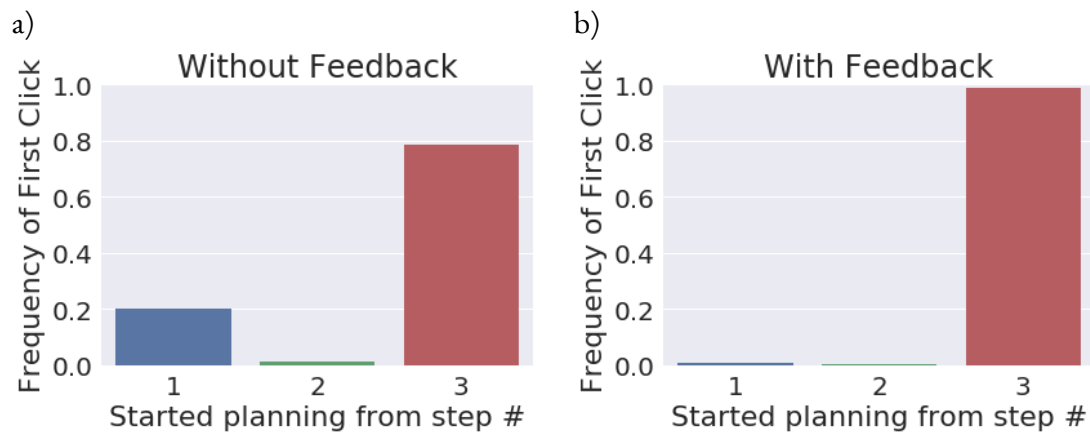
Figure 9.6: Effect of feedback on planning strategies. Compared to the control condition (a), participants who were trained with feedback engaged more often in backward planning and less often in forward planning.

condition (78.6% of all first clicks, $Z = 14.6, p < 0.0001$). Conversely, participants who had practiced with the cognitive tutor were less likely to start by inspecting immediate outcomes than the control conditoin (0.9% vs. 20.1% of all first clicks, $Z = -14.5, p < 0.0001$) or intermediate outcomes (0.46% vs. 1.25% of first clicks, $Z = -1.96, p = 0.0496$). Furthermore, 97% of the participants in the feedback group engaged in planning compared to only 76% of participants in the control group ($Z = 14.6, p < 0.0001$), and the feedback group also performed more planning overall (3.7 ± 3.0 vs. 2.4 ± 2.8 clicks on average, $t(2258) = 6.3, p < 0.0001$).

9.3.3 Discussion

The results of Experiment 1 suggested that training paradigm developed in this chapter could potentially to lead to effective cognitive training programs, because it accelerated learning to plan backwards. However, the critical question is whether training benefits achieved with the cognitive tutor would transfer to planning and decision-making in the real world. As a first step towards answering that question, we designed a follow-up experiment with a transfer task that assesses people's performance in a more complex environment. This mimics the scenario in which people engage in cognitive training with relatively simple tasks in hopes of improving their decision-making skills in everyday life.

9.4 Experiment 2: Do the training benefits transfer to other planning tasks?

To assess whether the training effects observed in Experiment 1 transfer to more complex environments, we modified Experiment 1 such that the test block uses a different and more complex task that requires planning five steps ahead.

9.4.1 Methods

We recruited 118 participants on Amazon Mechanical Turk. Each participant was paid $0.80 plus a performance-dependent bonus. The average bonus was 1.50 ± 0.59 and the average completion time was 16.3 ± 5.7 minutes. Exactly half of the participants were assigned to the experimental condition and the remaining half was assigned to the control condition.

The experiment was structured into instructions, a pretest block (1 trial), a training block (10 trials), a post-test block (20 trials), and an exit survey. The instructions explained how to move, how to collect information, the cost of information, and the minimum time of 7 seconds per trial.

The pre-test and the post-test presented participants with the complex 5-step planning task shown in Figure 9.7. In this task participants move a money-loving spider across a web of cash; they can choose between three directions in the first step, have no choice in the second and third step, have two choices in the fourth step, and three choices in the fifth step. The rewards of nodes at step $i \in \{1, 2, 3, 4, \}$ were drawn from a normal distributions with mean zero and standard deviation $\sigma_i = 2^{i-1}$, and the rewards at the last step ($i = 5$) were drawn from $\mathcal{N}(\mu = 0, \sigma_5 = 2^5)$. All sampled rewards were rounded to the nearest integer.

In the training block, participants solved a different planning problem that was simpler and had a different cover story (see Figure 9.8). This task was structurally identical to the training task used in Experiment 1. But it used a different cover story about navigating an airplane across a network of airports, and the icon of the spider was replaced by the icon of an airplane. To recap, this task required participants to solve a 3-step planning problem with three choices in the first step, no choice in the second step, and two choices in the third step. As in previous experiment, the reward distributions were Uniform($\{-4, -2, +2, +4\}$), Uniform($\{-8, -4, +4, +8\}$), and Uniform($\{-48, -24, +24, +48\}$) for nodes reachable in one, two, and three steps, respectively.

Each participant was randomly assigned to one of two conditions: Participants in the experimental group received optimal feedback in the training block. By contrast, participants in the control
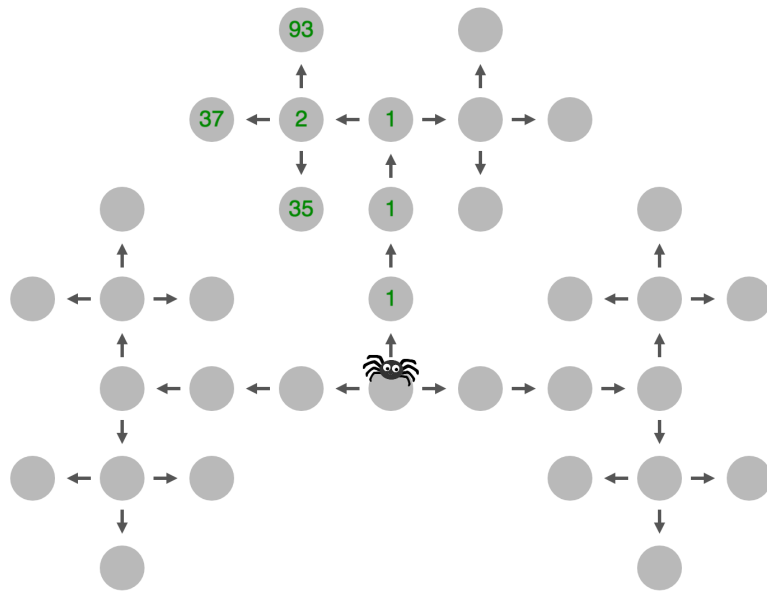
Clicking on a node reveals its value for a $1 fee.
Move with the arrow keys.

Figure 9.7: Transfer task of Experiment 2.
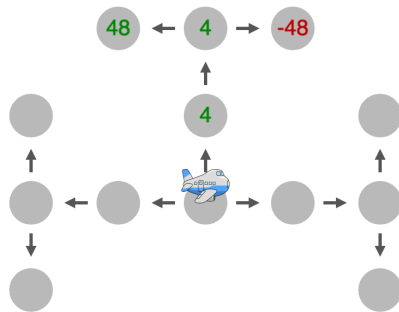
Clicking on a node reveals its value for a $1 fee.
Move with the arrow keys.

Figure 9.8: Training task of Experiment 2.
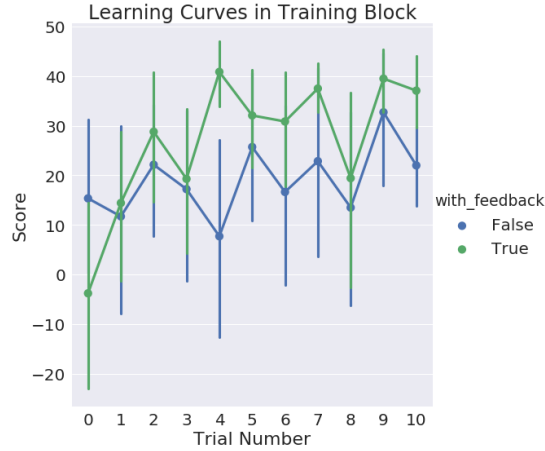
Figure 9.9: Replication of the effect of optimal feedback on the rate of learning in the training block. The error bars are 95% confidence intervals based on bootstrapping.

group received no feedback in the training block. In the post-test block, participants received an endowment of 20 cents ($100 in the game's currency) and were informed that they would receive a bonus of 20 cents for every $100 they made in the game.

9.4.2 RESULTS

The results replicated the finding that the cognitive tutor accelerates learning (see Figure 9.9). Concretely, a linear regression analysis confirmed that the feedback significantly increased the slope of participants' learning curve from 1.21 \$/trial to 2.40 \$/trial ($t(1294) = 2.33, p = 0.020$).

Most importantly, we found that the benefits of practicing 3-step planning with the cognitive tutor in the simple environment shown in Figure 9.8 transferred to the more difficult five-step planning problem in the more complex environment shown in Figure 9.7. Participants who trained with the cognitive tutor improved significantly more on the transfer task from the pretest to the post-test ($+30.7$ \$/trial) than participants who practiced the training task without feedback ($+15.6$ \$/trial, $t(116) = 2.53, p = 0.0127$). Consequently, the experimental group performed much better on the post-test than the control group (37.4 \$/trial vs. 27.4 \$/trial, $t(2358) = 8.8, p < 0.0001$; see Figure 9.10a). As shown in Figure 9.10b) the cognitive tutor appears to have shifted many people's performance from negative and low positive scores to moderately high positive scores.

The transfer effect appears to stem from at least two improvements in people's planning strate-
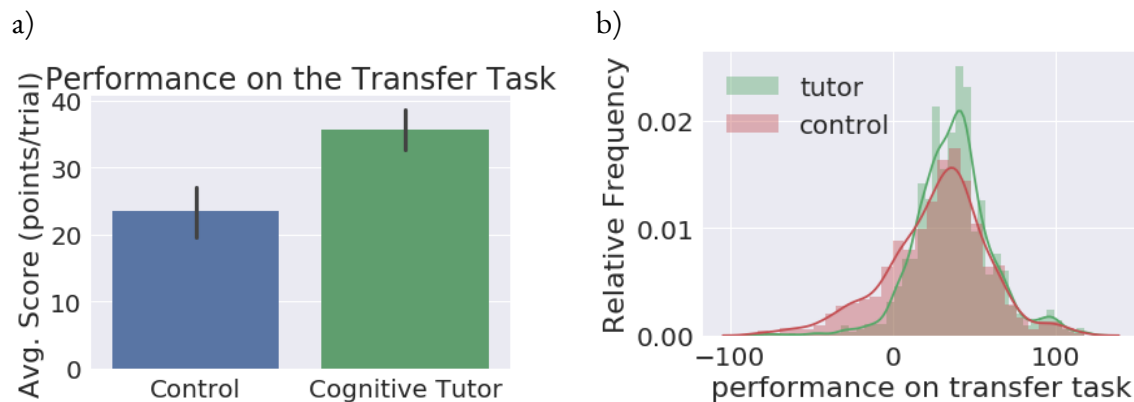
Figure 9.10: Performance in the transfer task. a) Average performance by group. b) Distribution of scores in the transfer task in $ per trial.

gies. First, participants who had practiced with the cognitive tutor employed backward planning more frequently in the transfer task than participants who had practiced on their own (see Figure 9.11): their first clicks fell more frequently on one of the possible final destinations than those of the control group (91.4% vs. 83.1%, $Z = 3.43, p = 0.0006$) and less frequently on one of the rewards in the first step (2.21% vs. 14.8%, $Z = -6.33, p < 0.0001$). Interestingly, having trained with the cognitive tutor made participants slightly more likely to start by inspecting a node at the third step (4.41% vs. 1.51%, $Z = 2.26, p = 0.0239$). Second, training with the cognitive tutor also increased people's propensity to engage in any planning at all: the proportion of trials on which participants inspected at least one location increased from 73% to 96.9% ($Z = 16.1, p < 0.0001$). Finally, practicing with the cognitive tutor also increased the amount of planning participants did on the transfer task, where it increased the average number of clicks from 6.9 ± 7.8 to 9.5 ± 7.5 ($t(2358) = 6.6, p < 0.0001$).

9.4.3 DISCUSSION

While learning-driven improvements in task performance are ubiquitous, those improvements tend to be highly specific to the trained task (C. S. Green & Bavelier, 2008). This specificity of learning is the primary obstacle to improving the human mind through cognitive training (C. S. Green & Bavelier, 2008). The present experiment was designed to test whether the training effects conferred by the cognitive tutor might transfer to the kinds of planning problems people face in everyday life. Those problems differ from the simple 3-step planning task used by the cognitive tutor in several
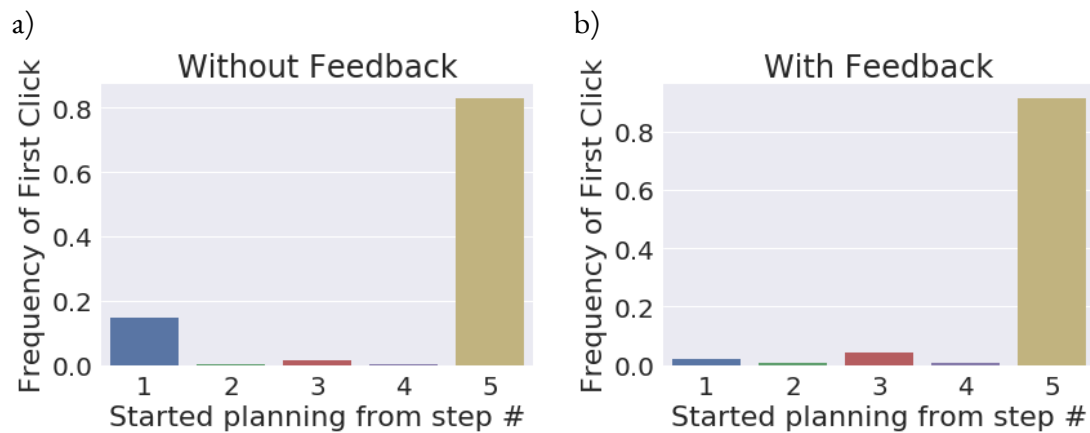
Figure 9.11: Effect of feedback on planning strategies in the transfer task. Participants who were trained with feedback (b) engaged in less forward planning and more backward planning than participants who practiced without feedback (a).

ways. Three key differences are that they typically require planning more than three steps ahead, allow people to choose between many more than six courses of action, and rarely involve flight planning. The transfer task differed from the training task in all three of these regards. The fact that the benefits of the planning training transferred to a more complex problem with a longer planning horizon and a larger number of potential scenarios suggests that transfer from simplified training problems to the more complex problems they were designed to mimic is possible, at least in principle. One of the reasons why the cognitive tutor might have a better chance at conferring transferable benefits than conventional cognitive training is that it explicitly teaches general cognitive strategies. The explicit click-by-click feedback guides people to carry out the target strategy multiple times and it rewards them for doing so. Critically, the training environment was designed such that the corresponding optimal strategy (i.e., identifying a potential goal state and planning backward from it) would also be beneficial in many real-world scenarios (Park, Lu, & Hedgcock, 2017).

How complex, varied, and realistic the tasks used in the planning training have to be to achieve transfer to planning in everyday life is an open question for future research. Next, we turn to a another challenge cognitive training programs have to meet to be useful for improving people's performance in everyday life: the training benefits have to be retained over time.

9.5 Experiment 3: Are the training benefits retained over time?

To test whether the transfer effect observed in Experiment 2 persists over time, we designed a follow-up experiment with an approximately 24 h delay between the training block and the transfer block.

9.5.1 Methods

We recruited a total of 100 adult participants on Amazon Mechanical Turk. We excluded the data from one participant who reported technical problems that caused them to participate twice. Participants were paid a base pay of $2.00 or $2.10 plus a performance-dependent bonus for about 16.3 ± 5.4 minutes of work. The average bonus was 2.29 ± 0.87.

The experimental design was equivalent to the near-transfer experiment (Experiment 2): participants were assigned to either the experimental condition that trained with the cognitive tutor or the control condition that practiced without feedback. Of the 99 remaining participants, 50 were in experimental condition and 49 were in the control condition. Most participants who participated in Stage 1 returned for Stage 2: in the experimental condition 43 of the 50 participants returned and in the control condition 36 of the 49 participants returned.

The experiment was structured into two parts that were separated by a delay of approximately 24 h. This was accomplished by posting separate HITs such that the HIT for Part 1 was available from the morning to the late afternoon of the first day and the HIT for Part 2 became available in the morning of the subsequent day and was only accessible to workers who had completed Part 1 on the previous day. The data was collected in two waves. The first wave took place from March 7 (Part 1) to March 8 2018 (Part 2), and the second wave took place on March 9 (Part 1) to March 10 2018 (Part 2).

The first HIT asked workers to only partake in the first part if they were certain they would also participate in the second part. The monetary incentives were set up to discourage participating only in the first part: The base pay of the first HIT was only $0.10 (first wave of recruitment) or $0.20 (second wave of recruitment). Participants could earn a performance-dependent bonus in part 1 but that bonus would only be paid out if they also completed part 2. Part 2 provided a base pay of $1.90 and an additional performance-dependent bonus. Furthermore, participants could sign up for an email reminder that was sent the following day after the HIT for Part 2 had been posted.
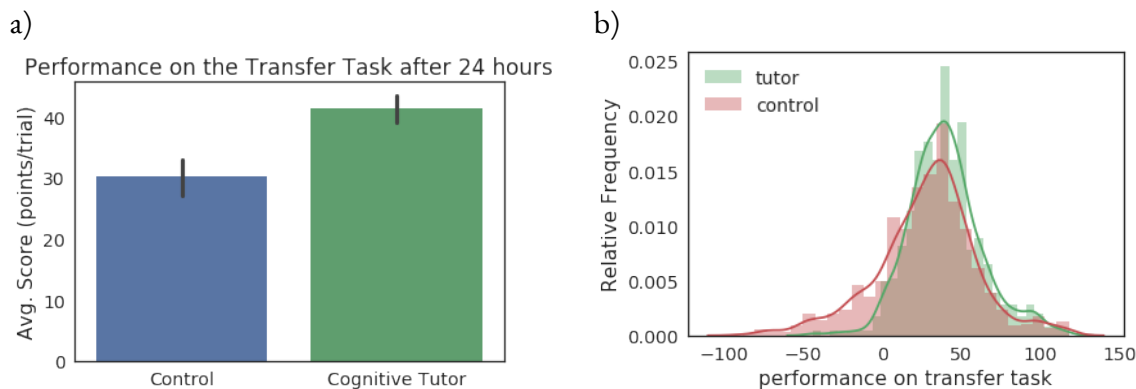
Figure 9.12: Performance in the transfer task after a 24 h delay. a) Average performance by group. b) Distribution of scores in the transfer task in $ per trial.

This experiment employed the pretest block, training block, and transfer from the near-transfer experiment (Experiment 2), but we added additional instructions and a bonus for participants' performance in the training block. The first HIT comprised instructions, the pretest block, and a training block. The second HIT comprised instructions that reminded participants of how the game works, the transfer block, where participants were posed 20 five-step planning problems (see Figure 9.7), and the same closing survey that was used in Experiments 1 and 2.

9.5.2 RESULTS

As shown in Figure 9.12, the results suggested that the transferable benefits of training with the cognitive tutor were retained over time. Concretely, we found that even after the 24 h delay, participants who had received feedback in the training block still performed significantly better on the transfer task than participants who had practiced without feedback (39.9 $/trial vs. 39.1 $/trial, $t(1578) = 7.8, p < 0.0001$). The retained benefit of 11.77 ± 1.88 $/trial was about 97.4% of the immediate benefit of 12.1 ± 3.55 $/trial; these findings are consistent with the null hypothesis that the transfer benefits are fully retained for at least 24 h ($Z = 0.06, p = 0.48$). The benefit of practicing with the cognitive tutor was also visible in participants' improvement from the pre-test to the post-test ($+33.9$ $/trial vs. $+23.9$ $/trial) but due to the high variability of the pretest scores this difference was only borderline-significant ($t(77) = 1.83, p = 0.0684$).

Figure 9.13 suggests that the improved performance of participants who had practiced planning with the cognitive tutor reflects their increased use of the backward-planning strategy in the transfer
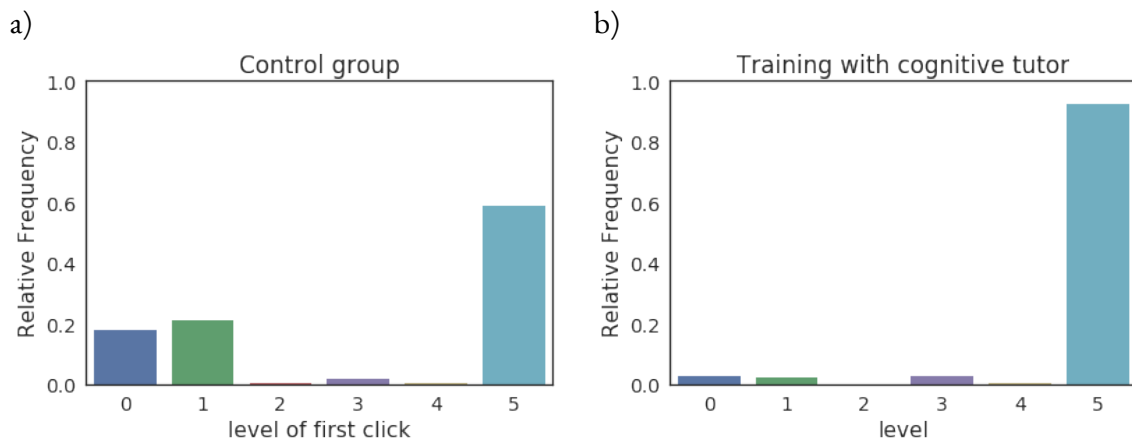
Figure 9.13: Distribution of participants' first clicks in the transfer block over levels. Level 0 means no click, level $1 \leq l \leq$ means clicking on a node that can be reached in l moves. a) Click distribution of the control condition. b) Click distribution of the experimental condition.

task 24 h after the training. Concretely, their first clicks fell more frequently on one of the possible final destinations than those of the control group (94.5% vs. 71.3%, $Z = 12.1, p < 0.0001$) and less frequently on one of the rewards in the first step (2.4% vs. 25.5%, $Z = -13.3, p < 0.0001$). Furthermore, the increased propensity for planning we observed in the near-transfer experiment was also retained after the 24 h delay: The experimental group inspected at least one node on 97.6% of the trials compared to only 82% in the control condition ($Z = 10.4, p < 0.0001$). This increased propensity for planning also manifested in an increased number of clicks (12.7 ± 7.3 vs. 7.8 ± 6.6, $t(1578) = 14.8, p < 0.0001$).

9.5.3 Discussion

These findings suggest that the transferable benefits of training planning with the cognitive tutor are retained for at least 24 hours after the training with very little to no reduction in its magnitude. Having practiced with the cognitive tutor increased people's propensity for planning and taught them to plan backward. Both of these changes could likely help people make better decisions in everyday life. Furthermore the results of this experiment suggest that if the training benefits transfer to decision-making in everyday life, then practicing in the evening would lead to persistent improvements throughout the following day.

9.6 EXPERIMENT 4: BENEFITS OVER PURE INSTRUCTION

The traditional paradigm of education is premised on the assumption that people can learn cognitive skills not only from experience but also through instruction. This raises the question of whether and under which conditions having people practice with the cognitive tutor is more effective than simply instructing them about the optimal strategy. To answer this question, Experiment 4 compared the effectiveness of instruction plus practice with the cognitive tutor versus pure instruction and instruction plus demonstration.

9.6.1 METHODS

We recruited 152 participants on Amazon Mechanical Turk.

Participants were paid a base pay of $2.10 plus a performance-dependent bonus. The average duration of the experiment was about 14.6 ± 9.9 minutes, and the average bonus was $\$2.12 \pm 0.57$.

The experiment used a between-subjects design with three conditions and two stages that were separated by a 12 h–24 h delay that were posted as two separate HITs on Amazon Mechanical Turk. The first HIT taught participants the optimal decision-making strategy for the 3-step planning task with outwardly increasing variance introduced above. Participants in the control were taught this strategy via the instructions shown in Figure 9.14. After having seen this principle, participants were then asked to summarize it in their own words, and then they were shown it again. Participants in the two experimental conditions received the same instructions as participants in the control condition, but afterward they either practiced applying this strategy with the cognitive tutor for 10 trials (Figure 9.8; cognitive tutor condition), or were shown 10 video demonstrations of the optimal strategy applied in this task (demonstration condition).

On the next day, we posted a second HIT in which participants of all three conditions completed 20 trials of the 5-step transfer task introduced in Experiment 2 (see Figure 9.7). In addition, the second HIT included instructions explaining the transfer task a closing survey that included the questions from Experiments 1-3 and also asked people whether they had applied the goal setting principle in their everyday life. This HIT was only accessible to workers who had completed the first part of the experiment. The first HIT asked workers to only partake in the first part if they were certain they would also participate in the second part. The monetary incentives were set up to discourage participating only in the first part: The base pay of the first HIT was only $0.20. Furthermore, the

As we go through our lives we are often drawn to immediate pleasures and avoid doing things that are unpleasant. For instance, we watch a Youtube video because it promises immediate fun, but we put off filing our taxes because that feels difficult.

Highly successful people, like Elon Musk, make their decisions very differently: They *first think about all the things they could achieve in the long-term, pick one of them as their goal*, and then *do what it takes to get there* – even if they are painful in the short-run.

You too can apply this goal-setting principle to make better decisions. Here is how:
1. *Imagine what your life could be like in the future.*
2. *Choose which of those futures you want to create.*
3. *Set yourself the goal to make that happen.*
4. *Plan how to achieve the goal and act accordingly.*

Figure 9.14: Instructions about the goal-setting principle shown in Experiment 4.

$0.55 bonus participants earned in the first part was only be paid out if they also completed Part 2. Part 2 provided a base pay of $1.90 and an additional performance-dependent bonus. Participants could sign up for an email reminder that was sent the following day after the HIT for Part 2 had been posted. About 81% of the participants from Part 1 returned for Part 2, and the retention rate was relatively even across the three conditions (43/50 in the control condition, 42/51 in the feedback condition, and 38/51 in the demonstration).

9.6.2 RESULTS

We found that participants who had practiced with the cognitive tutor performed significantly better on the transfer task than participants who were only told about the principle (38.0 $/trial vs. 24.2 $/trial, $t(83) = 10.5, p = 0.0000$; Figure 9.15). Participants who had seen a demonstration of the optimal strategy performed at the same level as participants who had practiced with the cognitive tutor (38.8 $/trial vs. 38.0 $/trial, $t(78) = -0.7, p = 0.49$; Figure 9.15). Furthermore, the performance of participants in the condition with instructions only was similar to the performance of participants who had practiced the training task without feedback in Experiment 2 (24.2 vs. 27.4 $/trial).
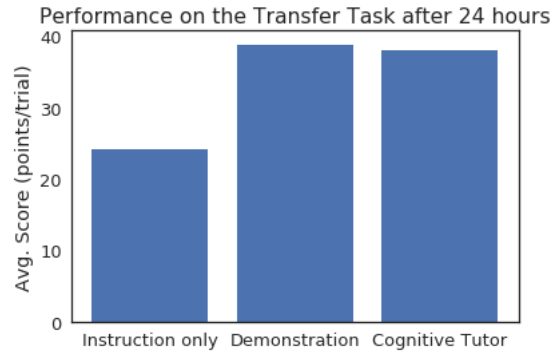
Figure 9.15: Transfer task performance in Experiment 4 by group.
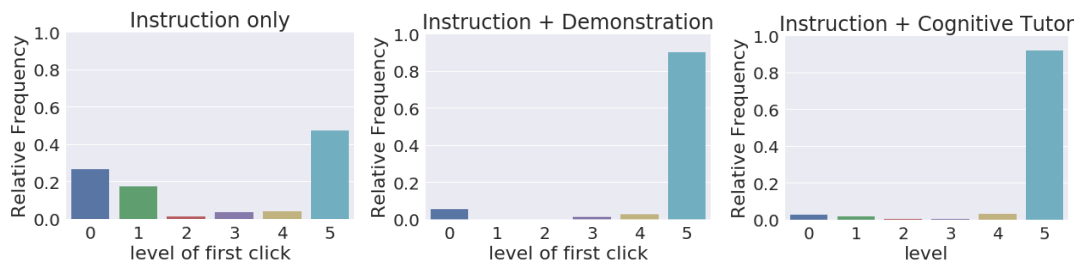
Figure 9.16: Transfer task strategies in Experiment 4 by condition.

As shown in Figure 9.16 the difference in performance between the three experimental conditions arose because participants who had trained with the cognitive tutor appeared to use the backward planning strategy significantly more frequently than participants who had received instructions only: Participants in the cognitive tutor condition started significantly more often by inspecting one of the potential final destinations than participants in the instructions only condition (92% vs. 47%, $Z = 20.3, p < 0.0001$). Conversely, participants who had been instructed about the optimal planning strategy only were significantly more likely to start by inspecting an immediate outcome (26% vs. 2.4%, $Z = +14.2, p < 0.0001$) or an outcome that was only 1 step away from their initial location (17% vs. 1.7%, $Z = +11.0, p < 0.0001$).

9.6.3 DISCUSSION

These results suggest that supplementing instruction by practice with a cognitive tutor or demonstrations has transferable benefits that are retained over time. These benefits might arise because the both practice and demonstration translate the abstract principle into a concrete decision strat-

egy. Furthermore, practice and demonstrations might help people internalize the abstract principles of good decision-making so that they become more likely to use them. Overall, the findings of this experiment suggest that our computational approach to strategy discovery enables training interventions that are significantly more effective than pure instruction.

9.7   Summary and Conclusion

This chapter brought together the theory of resource-rationality (Chapter 1), a computational approach to discovering rational heuristics (Chapter 6), and insights into how people learn how to think and decide (Chapter 5) to develop a cognitive tutor that assists people in learning a resource-rational planning strategy. The success of this approach illustrates that having a normative theory of bounded rationality and model of metacognitive learning make it possible to improve the human mind in ways that eluded previous attempts to improve judgment and decision-making, such as debiasing and cognitive training.

Unlike classical approaches to cognitive training that seek to strengthen basic cognitive capacities, such as working memory, the approach presented here teaches people to make clever use of the cognitive resources they already have. In other words, while most approaches to cognitive training are like running or weightlifting, this cognitive tutor is more like a karate instructor. In this sense, the approach presented here is similar to boosting and debiasing interventions that seek to teach people normative decision strategies. However, there are at least three critical differences: First, the cognitive tutor teaches resource-rational strategies. Second, classic approaches relied on verbalizing the strategy to be taught, but most of human expertise is very difficult or impossible to verbalize, and optimal strategies might not have a succinct and memorable verbal description. The cognitive tutor can teach arbitrary cognitive strategies via feedback without having to compromise the quality of the strategy for succinctness and memorability. Third, the cognitive tutor's pedagogy is rooted in the principle of learning by doing. Participants solve the cognitive task on their own and receive immediate, high-quality feedback that is designed to guide their metacognitive learning towards the optimal strategy as quickly as possible.

To summarize the approach, we have developed the cognitive tutor starting from a mathematical model of the kinds of sequential decision problems people face in everyday life. We then compute the bounded-optimal strategy for making such decisions. The optimal decision strategy is then taught to people by a cognitive training program that gives them optimal metacognitive feedback

on how they plan in a process tracing paradigm. we have instantiated this approach in a cognitive tutor that teaches people to plan backward.

Experiments 1-3 have shown that a) this approach significantly improves people's planning strategies in the training task compared to practicing without feedback, b) those benefits transfer to a more challenging planning task in a more complex environment, and c) those transfer effects are retained over time. Taken together, these findings encourage the interpretation that the approach developed here has the potential to improve people's decision-making in everyday life. However, previous research on cognitive training has consistently found that most transfer effects tend to be narrow (C. S. Green & Bavelier, 2008). Therefore, future experiments will test whether the benefits of practicing with the cognitive tutor also transfer to superficially dissimilar tasks where performance benefits from backward planning. It will be informative to determine under which conditions such far-transfer effects can be attained. If it turns out that far-transfer can be achieved with (a potentially modified version of) the cognitive tutor presented in this chapter, then subsequent work will investigate transfer to decision-making in everyday life.

Given that the benefits of training with the cognitive tutor will be limited to situations where the taught strategy is advantageous, practical applications of the cognitive tutoring approach presented here should be based on realistic models of decision-problems people face in everyday life. Modeling the choices people face in everyday life and deriving optimal strategies for making them is an important direction for future research. A cognitive tutor based on the resulting strategies might be able to teach people practical life skills that have been neglected by traditional approaches to education. I believe that one of the critical obstacles to teaching students how to think and decide well is that experts can rarely articulate their cognitive strategies. For instance, most mathematicians cannot verbalize their general strategies for mathematical problem solving.I hope that my approach to discovering and teaching resource-rational cognitive strategies will alleviate this bottleneck to learning sophisticated cognitive strategies.

In contrast to conventional cognitive training that targets basic cognitive capacities, such as working memory or processing speed, the approach devised in this chapter teaches people to more effectively use the cognitive capacities they already have. The underlying assumption is that being taught a rational heuristic will be more beneficial than conventional cognitive training for people who already have the cognitive capacities to execute that heuristic but do not already know it. I believe that in most situations the bottleneck to success is not that we lack basic cognitive capacities but that we have yet to learn how to use them effectively. For instance, if somebody's programming skills need to be improved, then pointing them to a programming class or tutorial is generally a much more

appropriate intervention than working memory training. I believe that the same argument applies to improving reasoning, decision-making, and problem solving more generally. Most people may already possess the cognitive capacities to perform well in most situations but may still lack effective cognitive strategies for at least some of them. This highlights the potential of cognitive tutors that teach effective cognitive strategies.

The approach to improving human cognition presented in this chapter is very general. While we have instantiated it to teach people to plan backward, it could also be applied to discover and teach optimal strategies for solving algebra problems, or to improve the executive functions of people struggling with attention deficit disorder. It can, in fact, be applied to virtually any cognitive ability, and it can be tailored to specific populations and task environments. This suggests a broad spectrum of potential future applications ranging from psychiatry and cognitive rehabilitation to education and cognitive enhancement. The line of research begun in this chapter might give rise to innovative approaches to teaching cognitive skills that might eventually revolutionize how mental disorders are treated, how seniors keep themselves mentally fit, and how cognitive skills are taught in schools. Beyond providing new tools for existing applications, the cognitive tutoring approach presented here might also give rise to entirely new applications, such as accelerating cognitive development and the cognitive enhancement of healthy young adults. These novel applications might enable us to push the boundaries of human rationality and reach unprecedented levels of cognitive performance.

10
Conclusion

My dissertation work took us beyond bounded rationality in two ways: In Part I, I developed a mathematically precise and very general theoretical framework for understanding human cognition that surpasses the vague notion of "bounded rationality". In Part II, I devised a cognitive tutor and a cognitive prosthesis which have the potential to extend the boundaries of human rationality. The cognitive tutor teaches people resource-rational planning strategies, enabling them to use their limited cognitive resources more effectively. By contrast, the cognitive prosthesis automatically restructures the environment, so that the heuristics people already employ lead to better decisions. These two complementary approaches correspond to the two blades of Herbert Simon's 'scissors' of bounded rationality (Simon, 1972, 1982).

By sharpening the blades of Simon's scissors of bounded rationality, the research presented in Part I has enabled a more clear-cut understanding of what it means to be rational, allowing us to identify the heuristics people use with mathematical precision. Resource-rationality reconciles normative principles that can account for mind's most impressive feats with people's most embarassing errors and cognitive biases. The resource-rational framework integrates the strengths of modeling approaches based on general principles with the descriptive accuracy of theories derived from empirical observations. I hope that the resource-rational framework will find widespread applications in cognitive modeling, since its methodology is relevant to all domains of human cognition. Resource-

rational analysis brings the methodological benefits of rational modeling to the algorithmic level of analysis that is of crucial interest to cognitive psychology. Furthermore, it connects the algorithmic level of analysis to the computational level of analysis so that computational level theories can constrain process models and vice versa. I am optimistic that these advances will enable significant progress in cognitive modeling. The literature reviewed in Chapter 1 and the findings presented in Chapters 2-6 are a testament to its potential.

Taken together, the findings of Part I and Part II suggest that resource-rationality is a promising theoretical framework for understanding and improving human cognition.

10.1 RESOURCE-RATIONALITY AS A SCIENTIFIC FOUNDATION FOR IMPROVING THE HUMAN MIND

The cognitive tutor and cognitive prosthesis presented in Part II illustrate that both 'blades' of bounded rationality can be forged so as to increase people's rationality overall. These successes suggest that resource-rationality has the potential to provide a scientific foundation for improving the human mind. Resource rationality can support this endeavour in at least three qualitatively different ways: First, it can be used to derive prescriptions for clearer thinking and better decision-making. This could, for instance, mean applying the method developed in Chapter 6 to discover optimal heuristics for decision-making in everyday life. These strategies could then be taught to people through verbal instruction or via a cognitive tutor. Second, resource-rational models of heuristics and biases can be used to reason about how we should restructure the environment to make good decision-making easier for people. Third, resource-rationality can be used as a normative standard to identify genuine sub-optimalities in human cognition; those can then be addressed, so as to help people avoid costly mistakes. It is remarkable that resource-rationality can serve all three of these functions, given that normative, descriptive, and prescriptive theories have historically been only loosely connected.

I am optimistic that by tackling the problem of improving the human mind from a resource-rational perspective, we will develop strategies, tools, and interventions significantly more effective than those we have so far. The cognitive prosthesis presented in Chapter 8 is a proof-of-concept that we can translate our understanding of people's bounded-rational decision-mechanisms into software tools that can help them overcome their cognitive limitations. Last but not least, the cognitive tutor presented in Chapter 9 serves as a proof-of-concept, demonstrating that we can improve human de-

cision making by defining and teaching resource-rational heuristics. The way in which the cognitive tutor teaches these strategies is informed by our preliminary model of how people learn to become increasingly more resource rational (Chapter 5). Future work will build on these foundations to define rational heuristics for progressively more realistic problems, devise increasingly more effective ways of teaching those heuristics, and develop more advanced cognitive prostheses.

Furthermore, comparing human performance against the metric of resource-rationality can be used as a principled way to identify which interventions are most appropriate in any particular situation. For instance, teaching people a resource-rational cognitive strategy (Chapter 9) appears most appropriate in situations where their original heuristics are vastly sub-optimal. In other situations, people might already be using resource-rational strategies, but the complexity of the problem might exceed their computational capacities. In cases like this, one might restructure the environment to simplify the computational problems it poses (Chapter 8), or augment people's cognitive capacities. One example of the former is to re-frame probabilistic reasoning problems in terms of natural frequencies rather than conditional probabilities (Gigerenzer & Hoffrage, 1995; Sedlmeier & Gigerenzer, 2001). Alternatively, resource constraints could be addressed through cognitive training or cognitive prostheses, like the one developed in Chapter 8. Resource-rational analysis could also be used to inform the prescription of cognitive training programs. For instance, a resource-rational analysis of a person's performance in various tests might reveal that their performance is mainly limited by their verbal working memory. In that case, working memory training using the verbal n-back task might be effective. In further situations, people's inferences or decisions might be fully rational, given their reasonable assumptions about the structure of the environment. But errors may arise because the current situation violates those assumptions. This appears to be the case for several cognitive biases that have been successfully explained through rational analysis. In these cases, it might be reasonable to accept these inferences as optimal, or to align the presentation of those particular problems with the implicit assumptions of the strategies that people use to solve them.

I am optimistic that this line of research will enable us to push the limits of human rationality in synergy with answering fundamental questions about bounded rationality and metacognitive learning.

References

Abbott, J. T., Austerweil, J., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. In F. Pereira, C. J. C. Burges, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 3050–3058). La Jolla, CA: Neural Information Processing Systems.

Abbott, J. T., & Griffiths, T. L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In L. Carlson (Ed.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. E. (2006). Reward-motivated learning: mesolimbic activation precedes memory formation. *Neuron*, *50*(3), 507–517. doi: 10.1016/j.neuron.2006.03.036

Ainslie, G., & Haslam, N. (1992). Hyperbolic discounting. In G. Loewenstein & J. Elster (Eds.), *Choice over time* (pp. 57–92). New York, NY: Russell Sage Foundation.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, *21*(4), 503–546. doi: 10.2307/1907921

Allais, M. (1979). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School (1952). In M. Allais & O. Hagen (Eds.), *Expected utility hypotheses and the Allais paradox: Contemporary discussions of the decisions under uncertainty with Allais' rejoinder* (pp. 27–145). Dordrecht, Netherlands: Springer Netherlands.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485. doi: 10.1017/S0140525X00070801

Anderson, J. R., & Milson, R. (1989). Human memory: an adaptive perspective. *Psychological Review*, *96*(4), 703–719. doi: 10.1037/0033-295X.96.4.703

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408. doi: 10.1111/j.1467-9280.1991.tb00174.x

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., ... Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature*, *501*(7465), 97–101. doi: 10.1038/nature12486

Ariely, D. (2009). *Predictably irrational.* New York, NY: HarperCollins.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106. doi: 10.1162/00335530360535153

Arkes, H. R. (1991). Costs and benefits of judgment errors: implications for debiasing. *Psychological Bulletin*, *110*(3), 486–498. doi: 10.1037/0033-2909.110.3.486

Aronson, J. E., Liang, T.-P., & Turban, E. (2004). *Decision support systems and intelligent systems* (7th ed.). Upper Saddle River, NJ: Pearson Prentice-Hall.

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, *46*(2), 81–86. doi: 10.1037/h0057955

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*, 397–422. Retrieved from http://www.jmlr.org/papers/v3/auer02a.html

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*(3), 499–526. doi: 10.1111/j.1551-6709.2010.01161.x

Aviv, Y., & Pazgal, A. (2005). A partially observed Markov decision process for dynamic pricing. *Management Science*, *51*(9), 1400–1416. doi: 10.1287/mnsc.1050.0393

Ball, K., Edwards, J. D., & Ross, L. A. (2007). The impact of speed of processing training on cognitive and everyday functions. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *62*(Special Issue 1), 19–31. doi: 10.1093/geronb/62.special_issue_1.19

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*(3), 215–233. doi: 10.1002/bdm.443

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*(4), 341–379. doi: 10.1023/a:1025696116075

Barto, A. G., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In J. Triech & T. Jebara (Eds.), *Proceedings of the 3rd International Conference on Development and Learning (ICDL 2004)* (pp. 112–119). La Jolla, CA: UCSD Institute for Neural Computation.

Bavelier, D., Green, C. S., Pouget, A., & Schrater, P. (2012). Brain plasticity through the life span: learning to learn and action video games. *Annual Review of Neuroscience*, *35*(1), 391–416. doi: 10.1146/annurev-neuro-060909-152832

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society London*, *53*, 370–418. doi: 10.1098/rstl.1763.0053

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, *3*(3), 439–449. doi: 10.5465/AMR.1978.4305717

Beck, J., Ma, W., Pitkow, X., Latham, P., & Pouget, A. (2012). Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39. doi: 10.1016/j.neuron.2012.03.016

Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations Research*, *33*(1), 1–27. doi: 10.1287/opre.33.1.1

Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review*, *120*(3), 522–543. doi: 10.1037/a0032457

Bhatnagar, S., Fernández-Gaucherand, E., Fu, M. C., He, Y., & Marcus, S. I. (1999). A Markov decision process model for capacity expansion and allocation. In *Proceedings of the 38th IEEE Conference on Decision and Control* (Vol. 2, pp. 1380–1385). Red Hook, NY: Curran Associates, Inc. doi: 10.1109/CDC.1999.830146

Bhui, R., & Gershman, S. J. (2017). Decision by sampling implements efficient coding of psychoeconomic functions. *bioRxiv*, 220277. Retrieved from https://www.biorxiv.org/content/early/2017/11/16/220277

Bjorklund, D. F., & Douglas, R. N. (1997). The development of memory strategies. In N. Cowan & C. Hulme (Eds.), *The development of memory in childhood* (pp. 201–246). Hove, United Kingdom: Psychology Press.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. doi: 10.1037/0033-295X.113.4.700

Bogacz, R., Hu, P., Holmes, P., & Cohen, J. (2010). Do humans produce the speed-accuracy trade-off that maximizes reward rate? *Quarterly Journal of Experimental Psychology*, *63*(5), 863–891. doi: 10.1080/17470210903091643

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65. doi: 10.1016/j.cogpsych.2014.06.003

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497–500. doi: 10.1016/j.tics.2014.06.006

Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Salience theory of choice under risk. *Quarterly Journal of Economics*, *127*(3), 1243–1285. doi: 10.1093/qje/qjs018

Bordalo, P., Gennaioli, N., & Shleifer, A. (2017). *Memory, attention, and choice.* Retrieved from https://scholar.harvard.edu/files/shleifer/files/mac.march9_.pdf (Working paper)

Bossaerts, P., & Murawski, C. (2017). Computational complexity and human decision-making. *Trends in Cognitive Sciences*, *21*(12), 917–929. doi: 10.1016/j.tics.2017.09.005

Bostan, B., & Öğüt, S. (2009). Game challenges and difficulty levels: lessons learned from RPGs. In G. K. Yeo & Y. Cai (Eds.), *Learn to Game, Game to Learn: Proceedings of the 40th Conference of the International Simulation and Gaming Association.* Singapore: Society of Simulation and Gaming of Singapore.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262–280. doi: 10.1016/j.cognition.2008.08.011

Botvinick, M. M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130480. doi: 10.1098/rstb.2013.0480

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: self-control and meta-decision making. *Trends in Cognitive Sciences*, *19*(11), 700–710. doi: 10.1016/j.tics.2015.08.013

Bourgin, D. D., Abbott, J. T., Griffiths, T. L., Smith, K. A., & Vul, E. (2014). Empirical evidence for Markov Chain Monte Carlo in memory search. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 224–229). Austin, TX: Cognitive Science Society.

Bowers, J. S., & Davis, C. J. (2012). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin*, *138*, 423–426. doi: 10.1037/a0027750

Brafman, R. I., & Tennenholtz, M. (2002). R-MAX – A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, *3*, 213–231.

Braine, M. D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*(1), 1–21. doi: 10.1037/0033-295X.85.1.1

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological Review*, *113*(2), 409–432. doi: 10.1037/0033-295X.113.2.409

Braver, T. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. doi: 10.1016/j.tics.2011.12.010

Brewer, N. T., & Chapman, G. B. (2002). The fragile basic anchoring effect. *Journal of Behavioral Decision Making*, *15*(1), 65–77. doi: 10.1002/bdm.403

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 611–625. doi: 10.1037/0278-7393.29.4.611

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition*, *5*(1), 73–99. doi: 10.1016/0010-0277(77)90018-X

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, *7*(11), e1002211. doi: 10.1371/journal.pcbi.1002211

Cador, M., Robbins, T. W., & Everitt, B. J. (1989). Involvement of the amygdala in stimulus-reward associations: interaction with the ventral striatum. *Neuroscience*, *30*(1), 77–86. doi: 10.1016/0306-4522(89)90354-0

Callan, R. C., Bauer, K. N., & Landers, R. N. (2015). How to avoid the dark side of gamification: Ten business scenarios and their unintended consequences. In T. Reiners & L. C. Wood (Eds.), *Gamification in education and business* (pp. 553–568). Cham, Switzerland: Springer. doi: 10.1007/978-3-319-10208-5_28

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. L. (2018). A resource-rational analysis of human planning. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Callaway, F., Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). Mouselab-MDP: A new paradigm for tracing how people plan. In *The 3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making, Ann Arbor, MI*. Retrieved from https://osf.io/vmkrq/

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1–3), 7–42. doi: 10.1023/A:1007850605129

Caplin, A., & Dean, M. (2013). *Behavioral implications of rational inattention with Shannon entropy* (NBER Working Paper No. 19318). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19318

Caplin, A., & Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, *105*(7), 2183–2203. doi: 10.1257/aer.20140117

Caplin, A., Dean, M., & Leahy, J. (2017). *Rationally inattentive behavior: Characterizing and generalizing Shannon entropy* (NBER Working Paper No. 23652). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w23652

Caplin, A., Dean, M., & Martin, D. (2011). Search and satisficing. *American Economic Review*, *101*(7), 2899–2922. doi: 10.1257/aer.101.7.2899

Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24* (pp. 2249–2257). Red Hook, NY: Curran Associates, Inc.

Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making*, *7*(4), 223–242. doi: 10.1002/bdm.3960070402

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment.* Cambridge, England: Cambridge University Press.

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences*, *34*(4), 194–196. doi: 10.1017/S0140525X11000239

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65. doi: 10.1016/S1364-6613(98)01273-X

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*(1–2), 93–131. doi: 10.1023/a:1005272027245

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*, 287–-291. doi: 10.1016/j.tics.2006.05.007

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive psychology*, *18*(3), 293–328. doi: 10.1016/0010-0285(86)90002-2

Christian, B., & Griffiths, T. L. (2016). *Algorithms to live by: the computer science of human decisions.* New York: Henry Holt and Company.

Christianson, S. Å., & Loftus, E. F. (1987). Memory for traumatic events. *Applied Cognitive Psychology*, *1*(4), 225–239. doi: 10.1002/acp.2350010402

Colombo, L., Femminis, G., & Pavan, A. (2014). Information acquisition and welfare. *The Review of Economic Studies*, *81*(4), 1438–1483. doi: 10.1093/restud/rdu015

Corr, P. J. (2004). Reinforcement sensitivity theory and personality. *Neuroscience & Biobehavioral Reviews*, *28*(3), 317–332. doi: 10.1016/j.neubiorev.2004.01.005

Courville, A., Daw, N., & Touretzky, D. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. doi: 10.1016/j.tics.2006.05.004

Cruciani, F., Berardi, A., Cabib, S., & Conversi, D. (2011). Positive and negative emotional arousal increases duration of memory traces: common and independent mechanisms. *Frontiers in Behavioral Neuroscience*, *5*, 86. doi: 10.3389/fnbeh.2011.00086

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, *112*(45), 13817–13822. doi: 10.1073/pnas.1506367112

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. doi: 10.1038/nn1560

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. doi: 10.1126/science.2648573

Dean, M., & Neligh, N. (2017). *Experimental tests of rational inattention* (Working paper). New York, NY: Columbia University. Retrieved from http://www.columbia.edu/~md3405/Working_Paper_21.pdf

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI)* (pp. 761–768). Palo Alto, CA: AAAI Press.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. (2013). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition*, *126*(2), 285–300. doi: 10.1016/j.cognition.2012.10.010

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15). New York, NY. doi: 10.1145/2181037.2181040

Devers, C. J., & Gurung, R. A. R. (2015). Critical perspective on gamification in education. In T. Reiners & L. C. Wood (Eds.), *Gamification in education and business* (pp. 417–430). Cham, Switzerland: Springer. doi: 10.1007/978-3-319-10208-5

Dickhaut, J., Rustichini, A., & Smith, V. (2009). A neuroeconomic theory of the decision process. *Proceedings of the National Academy of Sciences*, *106*(52), 22145–22150. doi: 10.1073/pnas.0912500106

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *308*(1135), 67–78. doi: 10.1098/rstb.1985.0010

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. doi: 10.1016/j.neuron.2013.09.007

Doucet, A., De Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. New York, NY: Springer. doi: 10.1007/978-1-4757-3437-9

Edwards, W. (1962). Subjective probabilities inferred from decisions. *Psychological Review*, *69*(2), 109–135. doi: 10.1037/h0038674

Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual Review of Psychology*, *52*(1), 581–606. doi: 10.1146/annurev.psych.52.1.581

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. doi: 10.1037/h0044139

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*(2), 188–200. doi: 10.1177/0146167205282152

Epley, N. (2004). A tale of tuned decks? Anchoring as accessibility and anchoring as adjustment. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 240–257). Oxford, England: Blackwell.

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin*, *30*(4), 447–460. doi: 10.1177/0146167203261889

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, *18*(3), 199–212. doi: 10.1002/bdm.495

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic. *Psychological Science*, *17*(4), 311–318. doi: 10.1111/j.1467-9280.2006.01704.x

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339. doi: 10.1037/0022-3514.87.3.327

Erev, I., & Barron, G. (2005, October). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*(4), 912–931. doi: 10.1037/0033-295x.112.4.912

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., … Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15–47. doi: 10.1002/bdm.683

Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. doi: 10.1016/j.tics.2003.08.012

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. doi: 10.1177/1745691612460685

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130. doi: 10.1016/j.tics.2010.01.003

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. doi: 10.1037/0003-066x.34.10.906

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

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, *120*(1), 34–45. doi: 10.1037/0096-3445.120.1.34

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. doi: 10.1126/science.1218633

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, *56*(4), 279–304. doi: 10.1086/256692

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. doi: 10.1098/rstb.2008.0300

Fudenberg, D., Strack, P., & Strzalecki, T. (2018). *Speed, accuracy, and the optimal timing of choices* (Working paper). Cambridge, MA: Massachusetts Institute of Technology. Retrieved from http://economics.mit.edu/files/11902

Fum, D., & Del Missier, F. (2001). Adaptive selection of problem solving strategies. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Second Annual Meeting of the Cognitive Science Society* (pp. 313–318). Mahwah, NJ: Lawrence Erlbaum Associates.

Gabaix, X. (2014). A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, *129*(4), 1661–1710. doi: 10.1093/qje/qju024

Gabaix, X., & Laibson, D. (2005). *Bounded rationality and directed cognition* (Working paper). Cambridge, MA: Harvard University. Retrieved from https://scholar.harvard.edu/xgabaix/publications/bounded-rationality-and-directed-cognition-working-paper

Gabaix, X., Laibson, D., Moloche, G., & Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *The American Economic Review*, *96*(4), 1043–1068. doi: 10.1257/aer.96.4.1043

Gadomski, A. M., Bologna, S., Costanzo, G. D., Perini, A., & Schaerf, M. (2001). Towards intelligent decision support systems for emergency managers: the IDA approach. *International Journal of Risk Assessment and Management*, *2*(3–4), 224–242. doi: 10.1504/IJRAM.2001.001507

Gagné, M., & Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior*, *26*(4), 331–362. doi: 10.1002/job.322

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: the role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, *81*(4), 657–669. doi: 10.1037/0022-3514.81.4.657

Geary, D. C., Brown, S. C., & Samaranayake, V. A. (1991). Cognitive addition: A short longitudinal study of strategy choice and speed-of-processing differences in normal and mathematically disabled children. *Developmental Psychology*, *27*(5), 787–797. doi: 10.1037/0012-1649.27.5.787

Gelfand, I. M., & Fomin, S. V. (2000). *Calculus of variations*. Mineola, NY: Courier Corporation.

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. doi: 10.1126/science.aac6076

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1–24. doi: 10.1162/NECO_a_00226

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*(6), 1317–1339. doi: 10.2307/1913710

Gigerenzer, G. (2008a). *Rationality for mortals: How people cope with uncertainty*. New York, NY: Oxford University Press.

Gigerenzer, G. (2008b). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29. doi: 10.1111/j.1745-6916.2008.00058.x

Gigerenzer, G. (2015). *Simply rational: Decision making in the real world*. New York, NY: Oxford University Press.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ: British Medical Journal*, *327*(7417), 741–744. doi: 10.1136/bmj.327.7417.741

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482. doi: 10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650–669. doi: 10.1037/0033-295X.103.4.650

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. doi: 10.1037/0033-295X.102.4.684

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

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London, England: Chapman & Hall.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York, NY: Cambridge University Press.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595. doi: 10.1016/j.neuron.2010.04.016

Gluth, S., Rieskamp, J., & Büchel, C. (2013). Neural evidence for adaptive strategy selection in value-based decision-making. *Cerebral Cortex*, *24*(8), 2009–2021. doi: 10.1093/cercor/bht049

Gold, J., & Shadlen, M. (2007). The neural basis of decision making. *Annual Review of Neuro- science*, *30*, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, *118*(4), 523–551. doi: 10.1037/a0024558

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635. doi: 10.1016/S0364-0213(03)00031-4

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166. doi: 10.1006/cogp.1998.0710

Good, I. J. (1983). *Good thinking: the foundations of probability and its applications*. Minneapolis, MN: University of Minnesota Press. doi: 10.5749/j.ctttsn6g

Green, C. S., & Bavelier, D. (2008). Exercising your brain: a review of human brain plasticity and training-induced learning. *Psychology and Aging*, *23*(4), 692–701. doi: 10.1037/a0014345

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, *26*(3), 499–510. doi: 10.1207/s15327906mbr2603_7

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. doi: 10.1016/j.tics.2010.05.004

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*, 415–422. doi: 10.1037/a0026884

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. doi: 10.1111/tops.12142

Griffiths, T. L., & Tenenbaum, J. (2011). Predicting the future as Bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *140*(4), 725–743. doi: 10.1037/a0024899

Griffiths, T. L., & Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 370–375). Mahwah, NJ: Lawrence Erlbaum.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773. doi: 10.1111/j.1467-9280.2006.01780.x

Griffiths, T. L., Vul, E., & Sanborn, A. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. doi: 10.1177/0963721412447619

Gunzelmann, G., & Anderson, J. R. (2001). An ACT-R model of the evolution of strategy use and problem difficulty. In E. M. Altmann, A. Cleermans, C. D. Schunn, & W. D. Gray (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling* (pp. 109–114). Mahwah, NJ: Lawrence Erlbaum.

Gunzelmann, G., & Anderson, J. R. (2003). Problem solving: Increased planning with practice. *Cognitive Systems Research*, *4*(1), 57–76. doi: 10.1016/s1389-0417(02)00073-6

Habenschuss, S., Jonke, Z., & Maass, W. (2013). Stochastic computations in cortical microcircuit models. *PLoS Computational Biology*, *9*(11), e1003311. doi: 10.1371/journal.pcbi.1003311

Hagen, O. (1979). Towards a positive theory of preferences under risk. In M. Allais & O. Hagen (Eds.), *Expected utility hypotheses and the Allais Paradox: Contemporary discussions of the decisions under uncertainty with Allais' rejoinder* (pp. 271–302). Dordrecht, Netherlands: Springer. doi: 10.1007/978-94-015-7629-1_13

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: a Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732. doi: 10.1037/0033-295X.114.3.704

Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: why three heads are better than four. *Psychological Review*, *116*(2), 454–461. doi: 10.1037/a0015241

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? – A literature review of empirical studies on gamification. In R. Sprague (Ed.), *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025–3034). Piscataway, NJ: IEEE. doi: 10.1109/HICSS.2014.377

Hammersley, D. C., & Handscomb, J. M. (1964). *Monte Carlo methods*. London, England: Methuen & Co. Ltd.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Harada, D., & Russell, S. (1998). *Meta-level reinforcement learning*. (Paper presented at NIPS'98 Workshop on Abstraction and Hierarchy in Reinforcement Learning)

Hardt, O., & Pohl, R. (2003). Hindsight bias as a function of anchor distance and anchor plausibility. *Memory, 11*(4–5), 379–394. doi: 10.1080/09658210244000504

Harman, G. (2013). Rationality. In H. LaFollette, J. Deigh, & S. Stroud (Eds.), *International Encyclopedia of Ethics*. Hoboken, NJ: Blackwell Publishing Ltd. doi: 10.1002/9781444367072.wbiee181

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York, NY: Springer. doi: 10.1007/978-0-387-84858-7

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97–109. doi: 10.1093/biomet/57.1.97

Hawkins, G. E., Camilleri, A. R., Heathcote, A., Newell, B. R., & Brown, S. D. (2014). Modeling probability knowledge and choice in decisions from experience. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Hay, N., Russell, S., Tolpin, D., & Shimony, S. (2012). Selecting computations: Theory and applications. In N. de Freitas & K. Murphy (Eds.), *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*. Corvallis, OR: AUAI Press.

Hedström, P., & Stern, C. (2008). Rational choice and sociology. In S. N. Durlauf & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). Basingstoke, England: Palgrave Macmillan.

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of Experimental Analysis of Behavior*, *24*(1), 107–116. doi: 10.1901/jeab.1975.24-107

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539. doi: 10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523. doi: 10.1016/j.tics.2009.09.004

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. doi: 10.1177/1745691617702496

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 621–642. doi: 10.1037/0278-7393.31.4.621

Hobson, C. J., & Delunas, L. (2001). National norms and life-event frequencies for the revised social readjustment rating scale. *International Journal of Stress Management*, *8*(4), 299–314. doi: 10.1023/A:1017565632657

Hobson, C. J., Kamen, J., Szostek, J., Nethercut, C. M., Tiedmann, J. W., & Wojnarowicz, S. (1998). Stressful life events: A revision and update of the social readjustment rating scale. *International Journal of Stress Management*, *5*(1), 1–23. doi: 10.1023/A:1022978019315

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, *290*(5500), 2261–2262. doi: 10.1126/science.290.5500.2261

Holmes, P., & Cohen, J. D. (2014). Optimality and some of its discontents: successes and shortcomings of existing models for binary decisions. *Topics in Cognitive Science*, *6*(2), 258–278. doi: 10.1111/tops.12084

Horvitz, E. J. (1987). Reasoning about beliefs and actions under computational resource constraints. In J. F. Lemmer, T. Levitt, & L. N. Kanal (Eds.), *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence* (pp. 301–324). Arlington, VA: AUAI Press.

Horvitz, E. J., Cooper, G. F., & Heckerman, D. E. (1989). Reflection and action under scarce resources: Theoretical principles and empirical study. In *Proceedings of the eleventh international joint conference on artificial intelligence* (pp. 1121–1127). San Mateo, CA: Morgan Kaufmann.

Horvitz, E. J., Suermondt, H., & Cooper, G. (1989). Bounded conditioning: Flexible inference for decisions under scarce resources. In M. Henrion (Ed.), *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence* (p. 182-193). New York, NY: Elsevier.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, *2*(1), 22–26. doi: 10.1109/TSSC.1966.300074

Howard, R. A. (1988). Decision analysis: practice and promise. *Management Science*, *34*(6), 679–695. doi: 10.1287/mnsc.34.6.679

Howes, A., Warren, P. A., Farmer, G., El-Deredy, W., & Lewis, R. L. (2016). Why contextual preference reversals maximize expected value. *Psychological Review*, *123*(4), 368–391. doi: 10.1037/a0039996

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*(3), e1002410. doi: 10.1371/journal.pcbi.1002410

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103. doi: 10.1073/pnas.1414219112

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166. doi: 10.1177/01461672952111004

Jaeggi, S., Buschkuehl, M., Jonides, J., & Perrig, W. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833. doi: 10.1073/pnas.0801268105

Jarvstad, A., Hahn, U., Rushton, S. K., & Warren, P. A. (2013). Perceptuo-motor, cognitive, and description-based decision-making seem equally good. *Proceedings of the National Academy of Sciences*, *110*(40), 16271–16276. doi: 10.1073/pnas.1300239110

Jiang, N., Kulesza, A., Singh, S., & Lewis, R. (2015). The dependence of effective planning horizon on model accuracy. In R. Bordini, E. Elkind, G. Weiss, & P. Yolum (Eds.), *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems* (pp. 1181–1189). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: a query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 461–474. doi: 10.1037/0278-7393.33.3.461

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, *31*(4), 395–414. doi: 10.1287/mnsc.31.4.395

Johnson, E. J., Payne, J. W., Bettman, J. R., & Schkade, D. A. (1989). *Monitoring information processing and decisions: The MouseLab system* (DTIC Document ADA205963). Fort Belvoir, VA: Defense Technical Information Center. Retrieved from http://dtic.mil/dtic/tr/fulltext/u2/a205963.pdf

Johnson, E. J., Shu, S. B., Dellaert, B. G. C., Fox, C., Goldstein, D. G., Häubl, G., ... Weber, E. U. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, *23*(2), 487–504. doi: 10.1007/s11002-012-9186-1

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*(4), 169–188. doi: 10.1017/S0140525X10003134

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed.). New York, NY: Farrar, Strauss and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. doi: 10.2307/1914185

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science*, *12*(6), 978–990. doi: 10.1111/j.1467-7687.2009.00846.x

Kass, R. E., & Raftery, A. E. (1995, June). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.2307/2291091

Kaufmann, E., Korda, N., & Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In N. H. Bshouty, G. Stoltz, N. Vayatis, & T. Zeugmann (Eds.), *Proceedings of the 23rd International Conference on Algorithmic Learning Theory, ALT 2012* (Vol. 12, pp. 199–213). Berlin, Germany: Springer.

Kawaguchi, K., Kaelbling, L. P., & Lozano-Pérez, T. (2015). Bayesian optimization with exponential convergence. In *Advances in Neural Information Processing Systems* (pp. 2809–2817). Red Hook, NY: Curran Associates.

Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards? *Cognition*, *157*, 126–138. doi: 10.1016/j.cognition.2016.08.020

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5), e1002055. doi: 10.1371/journal.pcbi.1002055

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873. doi: 10.1073/pnas.1609094113

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*(1), 271–304. doi: 10.1146/annurev.psych.55.090902.142005

Khaw, M. W., Li, Z., & Woodford, M. (2017). *Risk aversion as a perceptual bias* (NBER Working Paper No. 23294). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w23294

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences*, *14*(7), 317–324. doi: 10.1016/j.tics.2010.05.002

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. doi: 10.1016/j.tins.2004.10.007

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge, England: Cambridge University Press.

Knuth, D. E. (1998). *The art of computer programming: Sorting and searching* (2nd ed., Vol. 3). Boston, MA: Pearson Education.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247. doi: 10.1038/nature02169

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. doi: 10.1038/nn.2635

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857. doi: 10.1073/pnas.1101328108

Krueger, P. M., Lieder, F., & Griffiths, T. L. (2017). Enhancing metacognitive reinforcement learning using reward structures and feedback. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kunz, S. (2009). *The Bayesian linear model with unknown variance* (Seminar for Statistics). Zurich, Switzerland: ETH Zurich.

Laplace, P. S., & Simon, P. (1951). *A philosophical essay on probabilities*. New York, NY: Dover Publications. (Translated from the 6th French edition by Truscott, F. W. and Emory, F. L.. Original work published in 1814)

Larrick, R. P. (2002). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making*. Malden: Blackwell Publishing.

Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the use of cost-benefit reasoning in everyday life. *Psychological Science*, *1*(6), 362–370. doi: 10.1111/j.1467-9280.1990.tb00243.x

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *20*(7), 1434–1448. doi: 10.1364/JOSAA.20.001434

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153. doi: 10.1002/bdm.722

Lengyel, M., Koblinger, Á., Popović, M., & Fiser, J. (2015). On the role of time in perceptual decision making. *ArXiv e-prints*, 1502.03135. Retrieved from https://arxiv.org/abs/1502.03135

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*(6), 493–497. doi: 10.1016/S0960-9822(03)00135-0

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. doi: 10.1037/0033-2909.125.2.255

Levy, W. B., & Baxter, R. A. (1996). Energy efficient neural codes. *Neural Computation*, *8*(3), 531–543. doi: 10.1162/neco.1996.8.3.531

Levy, W. B., & Baxter, R. A. (2002). Energy-efficient neuronal computation via quantal synaptic failures. *Journal of Neuroscience*, *22*(11), 4746–4755. Retrieved from http://www.jneurosci.org/content/22/11/4746.long

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311. doi: 10.1111/tops.12086

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*(1), 46–55. doi: 10.1037/h0031207

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551–578. doi: 10.1016/S0096-1515(07)60316-8

Lieder, F., Callaway, F., Gul, S., Krueger, P. M., & Griffiths, T. L. (2017). Learning to select computations. *NIPS workshop on Cognitively Informed AI*, *abs/1711.06892*. Retrieved from http://arxiv.org/abs/1711.06892

Lieder, F., Chen, O. X., & Griffiths, T. L. (2018). *Cognitive prostheses for goal achievement.* (Manuscript submitted for publication)

Lieder, F., Goodman, N. D., & Griffiths, T. L. (2013). *Reverse-engineering resource-efficient algorithms.* (Paper presented at NIPS-2013 Workshop Resource-Efficient ML, Lake Tahoe, USA)

Lieder, F., Goodman, N. D., & Huys, Q. J. M. (2013). Controllability and resource-rational planning. In J. Pillow, N. Rust, M. Cohen, & P. Latham (Eds.), *Cosyne Abstracts 2013*. Denver, CO.

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Lieder, F., & Griffiths, T. L. (2016). Helping people make better decisions using optimal gamification. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 2075–2080). Austin, TX: Cognitive Science Society.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762–794. doi: 10.1037/rev0000075

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. In P. Bartlett, F. C. N. Pereira, L. Bottou, C. J. C. Burges, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 2690–2798). Red Hook, NY: Curran Associates, Inc.

Lieder, F., Griffiths, T. L., & Hsu, M. (2017). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1–32. doi: 10.1037/rev0000074

Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2017, January). Testing models of anchoring and adjustment. *PsyArXiv Preprint*. doi: 10.17605/OSF.IO/94YVZ

Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2018a). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322–349. doi: 10.3758/s13423-017-1286-8

Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2018b). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*.

Lieder, F., Hsu, M., & Griffiths, T. L. (2014). The high availability of extreme events serves resource-rational decision-making. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 2567–2572). Austin, TX: Cognitive Science Society.

Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). An automatic method for discovering rational heuristics for risky choice. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 742–747). Austin, TX: Cognitive Science Society.

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2870–2878). Red Hook, NY: Curran Associates, Inc.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*.

Lin, C. H., Kolobov, A., Kamar, E., & Horvitz, E. J. (2015). Metareasoning for planning under uncertainty. In *Proceedings of the 24th International Conference on Artificial Intelligence* (pp. 1601–1609). Palo Alto, CA: AAAI Press.

Lin, D., Donkin, C., & Newell, B. R. (2015). The exemplar confusion model: An account of biased probability estimates in decisions from description. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *34*(1), 1–41. doi: 10.2307/2985048

Littman, M. L., Dean, T. L., & Kaelbling, L. P. (1995). On the complexity of solving markov decision problems. In P. Besnard & S. Hanks (Eds.), *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 394–402). San Francisco, CA.

Lohman, S. (2008). Rational choice and political science. In S. N. Durlauf & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). Basingstoke, England: Palgrave Macmillan. doi: 10.1007/978-1-349-58802-2_1383

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805–824. doi: 10.2307/2232669

Loomes, G., & Sugden, R. (1984). The importance of what might have been. In O. Hagen & F. Wenstøp (Eds.), *Progress in utility and risk theory* (Vol. 42, pp. 219–235). Dordrecht, Netherlands: Springer. doi: 10.1007/978-94-009-6351-1_7

Loomes, G., & Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *The Review of Economic Studies*, *53*(2), 271–282. doi: 10.2307/2297651

Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward value-based gain control: divisive normalization in parietal cortex. *Journal of Neuroscience*, *31*(29), 10627–10639. doi: 10.1523/JNEUROSCI.1237-11.2011

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences*, *110*(15), 6139–6144. doi: 10.1073/pnas.1217854110

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making*, *27*(2), 146–156. doi: 10.1002/bdm.1792

Maass, W. (2000). On the computational power of winner-take-all. *Neural Computation*, *12*(11), 2519–2535. doi: 10.1162/089976600300014827

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2014). Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin & Review*, *21*(3), 629–636. doi: 10.3758/s13423-013-0542-9

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2016). The role of memory in distinguishing risky decisions from experience and description. *The Quarterly Journal of Experimental Psychology*, *70*(10), 2048–2059. doi: 10.1080/17470218.2016.1220608

Marchiori, D., Di Guida, S., & Erev, I. (2015). Noisy retrieval models of over-and undersensitivity to rare events. *Decision*, *2*(2), 82–106. doi: 10.1037/dec0000023

Marcus, G. (2009). *Kluge: The haphazard evolution of the human mind*. Boston, MA: Houghton Mifflin Harcourt.

Marewski, J. N., & Link, D. (2014). Strategy selection: An introduction to the modeling challenge. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 39–59. doi: 10.1002/wcs.1265

Marewski, J. N., & Schooler, L. (2011). Cognitive niches: an ecological model of strategy selection. *Psychological Review*, *118*(3), 393–437. doi: 10.1037/a0024143

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

May, B. C., Korda, N., Lee, A., & Leslie, D. S. (2012). Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, *13*, 2069–2106. Retrieved from http://www.jmlr.org/papers/volume13/may12a/may12a.pdf

McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience*, *27*(1), 1–28. doi: 10.1146/annurev.neuro.27.070203.144157

McGaugh, J. L., McIntyre, C. K., & Power, A. E. (2002). Amygdala modulation of memory consolidation: Interaction with other brain systems. *Neurobiology of Learning and Memory*, *78*(3), 539–552. doi: 10.1006/nlme.2002.4082

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, *49*(2), 270–291. doi: 10.1037/a0028228

Mengersen, K. L., & Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, *24*(1), 101–121. doi: 10.1214/aos/1033066201

Mill, J. S. (1882). *A system of logic, ratiocinative and inductive* (8th ed.). New York, NY: Harper and Brothers.

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. doi: 10.1037/h0043158

Milli, S., Lieder, F., & Griffiths, T. L. (2017). When does bounded-optimal metareasoning favor few cognitive systems? In S. P. Singh & S. Markovitch (Eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (Vol. 31). Palo Alto, CA: AAAI Press.

Milli, S., Lieder, F., & Griffiths, T. L. (2018). *A rational reinterpretation of dual-process theories.* (Manuscript submitted for publication) doi: 10.13140/RG.2.2.14956.46722

Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science*, *244*(4907), 933–938. doi: 10.1126/science.2658056

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. doi: 10.1038/nature14236

Mockus, J. (2012). *Bayesian approach to global optimization: theory and applications* (Vol. 37). New York, NY: Springer Science & Business Media. doi: 10.1007/978-94-009-0909-0

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*(30), 12491–12496. doi: 10.1073/pnas.1101430108

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*, *18*(1), 46–60. doi: 10.3758/s13423-010-0034-0

Mulder, G. (1986). The concept and measurement of mental effort. In G. R. J. Hockey, A. W. K. Gaillard, & M. G. H. Coles (Eds.), *Energetics and human information processing* (pp. 175–198). Dordrecht, Netherlands: Springer. doi: 10.1007/978-94-009-4448-0_12

Mullett, T. L., & Tunney, R. J. (2013). Value representations by rank order in a distributed network of varying context dependency. *Brain and Cognition*, *82*(1), 76–83. doi: 10.1016/j.bandc.2013.02.010

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*(2), 136–164. doi: 10.1006/jesp.1998.1364

Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology*, *78*(6), 1038–1052. doi: 10.1037/0022-3514.78.6.1038

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, *64*(3), 263–276. doi: 10.1901/jeab.1995.64-263

Neal, R. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X. L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (Vol. 2, pp. 113–162). Boca Raton, FL: CRC Press.

Nessler, B., Pfeiffer, M., Buesing, L., & Maass, W. (2013). Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS Computational Biology*, *9*(4), e1003037. doi: 10.1371/journal.pcbi.1003037

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151–166. doi: 10.1037/h0048495

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In I. Bratko & S. Dzeroski (Eds.), *Proceedings of the 16th Annual International Conference on Machine Learning* (pp. 278–287). San Francisco, CA: Morgan Kaufmann.

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*(1), 84–93. doi: 10.1016/j.jmp.2010.08.006

Nisbett, R. E. (1993). *Rules for reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. doi: 10.1016/j.jmp.2008.12.005

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–520. doi: 10.1007/s00213-006-0502-4

Niven, J. E., & Laughlin, S. B. (2008). Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, *211*(11), 1792–1804. doi: 10.1242/jeb.017574

Noguchi, T., & Stewart, N. (in press). Multialternative decision by sampling: a model of decision making constrained by process data. *Psychological Review*. doi: 10.1037/rev0000102

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes*, *39*(1), 84–97. doi: 10.1016/0749-5978(87)90046-X

Nouchi, R., Taki, Y., Takeuchi, H., Hashizume, H., Akitsuki, Y., Shigemune, Y., ... Kawashima, R. (2012). Brain training game improves executive functions and processing speed in the elderly: a randomized controlled trial. *PLoS One*, *7*(1), e29676. doi: 10.1371/journal.pone.0029676

Nouchi, R., Taki, Y., Takeuchi, H., Hashizume, H., Nozawa, T., Kambara, T., ... Kawashima, R. (2013). Brain training game boosts executive functions, working memory and processing speed in the young adults: a randomized controlled trial. *PLoS One*, *8*(2), e55518. doi: 10.1371/journal.pone.0055518

Nunes, L. G. N., de Carvalho, S. V., & Rodrigues, R. d. C. M. (2009). Markov decision process applied to the control of hospital elective admissions. *Artificial Intelligence in Medicine*, *47*(2), 159–171. doi: 10.1016/j.artmed.2009.07.003

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. doi: 10.1037/0033-295X.101.4.608

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

Oh, M.-S., & Berger, J. O. (1992). Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, *41*(3–4), 143–168. doi: 10.1080/00949659208810398

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. doi: 10.1038/381607a0

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325. doi: 10.1016/S0042-6989(97)00169-7

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487. doi: 10.1016/j.conb.2004.07.007

Orhan, A. E., Sims, C. R., Jacobs, R. A., & Knill, D. C. (2014). The adaptive nature of visual working memory. *Current Directions in Psychological Science*, *23*(3), 164–170. doi: 10.1177/0963721414529144

Oster, M., Douglas, R., & Liu, S.-C. (2009). Computation with spikes in a winner-take-all network. *Neural Computation*, *21*(9), 2437–2465. doi: 10.1162/neco.2009.07-08-829

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., ... Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*(7299), 775–778. doi: 10.1038/nature09042

Pachur, T., Hertwig, R., & Steinmann, F. (2012). How do people judge risks: availability heuristic, affect heuristic, or both? *Journal of Experimental Psychology: Applied*, *18*(3), 314–330. doi: 10.1037/a0028279

Park, J., Lu, F.-C., & Hedgcock, W. M. (2017). Relative effects of forward and backward planning on goal pursuit. *Psychological Science*, *28*(11), 1620–1630. doi: 10.1177/0956797617715510

Payne, J. W. (1982). Contingent decision behavior. *Psychological Bulletin*, *92*(2), 382–402. doi: 10.1037/0033-2909.92.2.382

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. doi: 10.1037/0278-7393.14.3.534

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, *43*(1), 87–131. doi: 10.1146/annurev.ps.43.020192.000511

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, England: Cambridge University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. doi: 10.1037/0033-295X.87.6.532

Penny, W. D., & Ridgway, G. R. (2013). Efficient posterior probability mapping using Savage-Dickey ratios. *PLoS One*, *8*(3), e59655. doi: 10.1371/journal.pone.0059655

Penny, W. D., Stephan, K., Daunizeau, J., Rosa, M., Friston, K., Schofield, T., & Leff, A. (2010). Comparing families of dynamic causal models. *PLoS Computational Biology*, *6*(3), e1000709. doi: 10.1371/journal.pcbi.1000709

Piaget, J. (1952). *The origins of intelligence in children*. New York, NY: International Universities Press.

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*(5), 2000–2019. doi: 10.1037/xge0000013

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review*, *122*(4), 621–647. doi: 10.1037/a0039413

Pohl, R. F. (1998). The effects of feedback source and plausibility of hindsight bias. *European Journal of Cognitive Psychology*, *10*(2), 191–212. doi: 10.1080/713752272

Post, T., van den Assem, M. J., Baltussen, G., & Thaler, R. H. (2008). Deal or No Deal? Decision making under risk in a large-payoff game show. *The American Economic Review*, *98*, 38–71. doi: 10.1257/aer.98.1.38

Power, D. J., Sharda, R., & Burstein, F. (2015). Decision support systems. In D. Straub & R. Welke (Eds.), *Wiley Encyclopedia of Management* (Vol. 7). Oxford, England: John Wiley & Sons.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. Hoboken, NJ: John Wiley & Sons.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, *3*(4), 323–343. doi: 10.1016/0167-2681(82)90008-7

Rasmussen, B. K., Jensen, R., Schroll, M., & Olesen, J. (1991). Epidemiology of headache in a general population—a prevalence study. *Journal of Clinical Epidemiology*, *44*(11), 1147–1157. doi: 10.1016/0895-4356(91)90147-2

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., … Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*(2), 359–379. doi: 10.1037/a0029082

Reichenbach, H. (1947). *Elements of symbolic logic.* New York, NY: Macmillan Co.

Reis, R. (2006). Inattentive consumers. *Journal of Monetary Economics*, *53*(8), 1761–1800. doi: 10.1016/j.jmoneco.2006.03.001

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1446–1465. doi: 10.1037/a0013646

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207–236. doi: 10.1037/0096-3445.135.2.207

Robert, C., & Casella, G. (2009). *Introducing Monte Carlo methods with R.* New York, NY: Springer Science & Business Media.

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. *European Journal of Neuroscience*, *35*(7), 1190–1200. doi: 10.1111/j.1460-9568.2011.07986.x

Rothman, A. J., Klein, W. M., & Weinstein, N. D. (1996). Absolute and relative biases in estimations of personal risk. *Journal of Applied Social Psychology*, *26*(14), 1213–1236. doi: 10.1111/j.1559-1816.1996.tb01778.x

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence*, *94*(1–2), 57–77. doi: 10.1016/s0004-3702(97)00026-x

Russell, S. J., & Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, *2*, 575–609.

Russell, S. J., & Wefald, E. (1991a). *Do the right thing: studies in limited rationality.* Cambridge, MA: MIT Press.

Russell, S. J., & Wefald, E. (1991b). Principles of metareasoning. *Artificial Intelligence*, *49*(1–3), 361–395. doi: 10.1016/0004-3702(91)90015-C

Russo, J. E., & Schoemaker, P. J. H. (1989). *Decision traps: Ten barriers to brilliant decision-making and how to overcome them.* New York, NY: Simon & Schuster.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. doi: 10.1037/a0020511

Sanjurjo, A. (2017). Search with multiple attributes: Theory and empirics. *Games and Economic Behavior*, *104*, 535–562. doi: 10.1016/j.geb.2017.05.009

Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801. doi: 10.2307/2284229

Schaul, T., & Schmidhuber, J. (2010). Metalearning. *Scholarpedia*, *5*(6), 4650.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, *120*(1), 39–64. doi: 10.1037/a0030777

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. doi: 10.1126/science.275.5306.1593

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Schwarz, N. (2014). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation.* New York, NY: Psychology Press.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380–400. doi: 10.1037/0096-3445.130.3.380

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939. doi: 10.1016/j.neuron.2016.04.036

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*, 99–124. doi: 10.1146/annurev-neuro-072116-031526

Shi, L., & Griffiths, T. (2009). Neural implementation of hierarchical Bayesian inference by importance sampling. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1669–1677). Red Hook, NY: Curran Associates, Inc.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464. doi: 0.3758/PBR.17.4.443

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*(4), 628–654. doi: 10.1037/a0027473

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science*, *9*(5), 405–410. doi: 10.1111/1467-9280.00076

Shteingart, H., Neiman, T., & Loewenstein, Y. (2013). The role of first impression in operant learning. *Journal of Experimental Psychology: General*, *142*(2), 476–488. doi: 10.1037/a0029550

Shugan, S. M. (1980). The cost of thinking. *Journal of Consumer Research*, *7*(2), 99–111. doi: 10.1086/208799

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*(3), 258–275. doi: 10.1037/0096-3445.117.3.258

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. Oxford, England: Oxford University Press.

Siegler, R. S. (1999). Strategic development. *Trends in Cognitive Sciences*, *3*(11), 430–435. doi: 10.1016/s1364-6613(99)01372-8

Siegler, R. S., & Jeff, S. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of Cognitive Skills: The 18th Annual Carnegie Mellon Symposium on Cognition* (pp. 229–293). Hillsdale, NJ: Lawrence Erlbaum Associates.

Siegler, R. S., & Jenkins, E. A. (1989). *How children discover new strategies*. New York, NY: Psychology Press.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. J. Simon & G. S. Graeme (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31–76). Hillsdale, NJ: Lawrence Erlbaum Associates.

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: do people adjust from provided anchors? *Journal of Personality and Social Psychology*, *99*(6), 917–932. doi: 10.1037/a0021540

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. doi: 10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. doi: 10.1037/h0042769

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, *1*, 161–176. Retrieved from http://innovbfa.viabloga.com/files/Herbert_Simon___theories_of_bounded _rationality___1972.pdf

Simon, H. A. (1982). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). Cambridge, MA: MIT Press.

Simonson, I., & Drolet, A. (2004). Anchoring effects on consumers' willingness-to-pay and willingness-to-accept. *Journal of Consumer Research*, *31*(3), 681–690. doi: 10.1086/425103

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690. doi: 10.1016/S0304-3932(03)00029-1

Sims, C. A. (2006). Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, *96*(2), 158–163. doi: 10.1257/000282806777212431

Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision*, *15*(3), 1–27. doi: 10.1167/15.3.2

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198. doi: 10.1016/j.cognition.2016.03.020

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–930. doi: 10.1037/a0029856

Slagter, H. A., Lutz, A., Greischar, L. L., Francis, A. D., Nieuwenhuis, S., Davis, J. M., & Davidson, R. J. (2007). Mental training affects distribution of limited brain resources. *PLoS Biology*, *5*(6), e138. doi: 10.1371/journal.pbio.0050138

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. doi: 10.1016/j.tins.2004.01.006

Smith, R. (2017). Aristotle's logic. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017 ed.). Stanford, CA: Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2017/entries/aristotle-logic/

Smith-Miles, K. A. (2009). Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Computing Surveys*, *41*(1), 1–25. doi: 10.1145/1456650.1456656

Song, H., Liu, C.-C., Lawarrée, J., & Dahlgren, R. W. (2000). Optimal electricity supply bidding by Markov decision process. *IEEE Transactions on Power Systems*, *15*(2), 618–624. doi: 10.1109/59.867150

Sosis, C., & Bishop, M. (2013). Rationality. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 27–37. doi: 10.1002/wcs.1263

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., & Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, *30*(3), 512–523. doi: 10.1111/j.1539-6924.2009.01337.x

Stanovich, K. E. (2009). *Decision making and rationality in the modern world*. Oxford, England: Oxford University Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford, England: Oxford University Press.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*(2), 332–382. doi: 10.1257/jel.38.2.332

Starmer, C., & Sugden, R. (1989). Probability and juxtaposition effects: An experimental investigation of the common ratio effect. *Journal of Risk and Uncertainty*, *2*(2), 159–178. doi: 10.1007/BF00056135

Steel, P. (2007). The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, *133*(1), 65–94. doi: 10.1037/0033-2909.133.1.65

Steel, P., & König, C. J. (2006). Integrating theories of motivation. *Academy of Management Review*, *31*(4), 889–913. doi: 10.5465/AMR.2006.22527462

Stephan, K., Penny, W., Daunizeau, J., Moran, R., & Friston, K. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*(4), 1004–1017. doi: 10.1016/j.neuroimage.2009.03.025

Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. Cambridge, MA: MIT Press.

Stewart, L., Overath, T., Warren, J., Foxton, J., & Griffiths, T. (2008). fMRI evidence for a cortical hierarchy of pitch pattern processing. *PLoS One*, *3*(1), e1470. doi: 10.1371/journal.pone.0001470

Stewart, N. (2009). Decision by sampling: the role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology*, *62*(6), 1041–1062. doi: 10.1080/17470210902747112

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. doi: 10.1016/j.cogpsych.2005.10.003

Stewart, N., Reimers, S., & Harris, A. J. L. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, *61*(3), 687–705. doi: 10.1287/mnsc.2013.1853

Stewart, T. C., West, R., & Lebiere, C. (2009). Applying cognitive architectures to decision making: How cognitive theory and the equivalence measure triumphed in the Technion Prediction Tournament. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 561–566). Houston, TX: Cognitive Science Society.

Stocker, A., Simoncelli, E., & Hughes, H. (2006). Sensory adaptation within a Bayesian framework for perception. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems* (Vol. 18, pp. 1291–1298). Cambridge, MA: MIT Press.

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*(2), 101–130. doi: 10.1007/s11166-006-8289-6

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*(3), 437–446. doi: 10.1037/0022-3514.73.3.437

Suchow, J. W. (2014). *Measuring, monitoring, and maintaining memories in a partially observable mind* (Doctoral dissertation). Retrieved from https://dash.harvard.edu/handle/1/12274120

Suchow, J. W., & Griffiths, T. L. (2016). Deciding to remember: memory maintenance as a Markov decision process. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2063–2068). Austin, TX: Cognitive Science Society.

Summerfield, C., & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences*, *19*(1), 27–34. doi: 10.1016/j.tics.2014.11.005

Sunstein, C. R., & Zeckhauser, R. (2011). Overreaction to fearsome risks. *Environmental and Resource Economics*, *48*(3), 435–449. doi: 10.1007/s10640-010-9449-3

Sutherland, S. (1992). *Irrationality: The enemy within*. London, England: Constable and Company.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1–2), 181–211. doi: 10.1016/s0004-3702(99)00052-1

Svenson, O., & Sjöberg, K. (1983). Evolution of cognitive processes for solving simple additions during the first three school years. *Scandinavian Journal of Psychology*, *24*(1), 117–124. doi: 10.1111/j.1467-9450.1983.tb00483.x

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, *7*, 12400–12411. doi: 10.1038/ncomms12400

Tang, Y.-Y., & Posner, M. I. (2009). Attention training and attention state training. *Trends in Cognitive Sciences*, *13*(5), 222–227. doi: 10.1016/j.tics.2009.01.009

Tenenbaum, J. B., Griffiths, T. L., et al. (2001). The rational basis of representativeness. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the Cognitive Science Society* (pp. 1036–41). Austin, TX: Cognitive Science Society.

Thaler, R. H., & Johnson, E. J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science*, *36*(6), 643–660. doi: 10.1287/mnsc.36.6.643

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 847–855). New York, NY: ACM. doi: 10.1145/2487575.2487629

Tierney, L., & Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86. doi: 10.2307/2287970

Todd, P. M., & Brighton, H. (2015). Building the theory of ecological rationality. *Minds and Machines*, *26*(1–2), 9–30. doi: 10.1007/s11023-015-9371-0

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, *16*(3), 167–171. doi: 10.1111/j.1467-8721.2007.00497.x

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world.* New York, NY: Oxford University Press.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*(9), 907–915. doi: 10.1038/nn1309

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. doi: 10.1080/13546783.2013.844729

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, *113*(11), 3102–3107. doi: 10.1073/pnas.1519157113

Tsotsos, J. K. (1988). How does human vision beat the computational complexity of visual perception. In Z. W. Pylyshyn (Ed.), *Computational processes in human vision: An interdisciplinary perspective* (pp. 286–338). Norwood, NJ: Ablex Press.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48. doi: 10.1037/h0026750

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281–299. doi: 10.1037/h0032955

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. doi: 10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi: 10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. doi: 10.1007/BF00122574

van den Berg, R., & Ma, W. J. (2017). A rational theory of the limitations of working memory and attention. *bioRxiv*, 151365. doi: 10.1101/151365

Verrecchia, R. E. (1982). Information acquisition in a noisy rational expectations economy. *Econometrica*, *50*(6), 1415–1430. doi: 10.2307/1913389

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

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. doi: 10.1111/cogs.12101

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101–118. doi: 10.1111/1467-6419.00106

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... Botvinick, M. (2017). Learning to reinforcement learn. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. London, England.

Wang, Z., Wei, X.-X., Stocker, A. A., & Lee, D. D. (2016). Efficient neural codes under metabolic constraints. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 4619–4627). Red Hook, NY: Curran Associates, Inc.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. doi: 10.1080/14640746808400161

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292. doi: 10.1007/bf00992698

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice a query-theory account. *Psychological Science*, *18*(6), 516–523. doi: 10.1111/j.1467-9280.2007.01932.x

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, *18*(10), 1509–1517. doi: 10.1038/nn.4105

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244–10249. doi: 10.1073/pnas.1619153114

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, *125*(4), 387–402. doi: 10.1037/0096-3445.125.4.387

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*(11s), 1212–1217. doi: 10.1038/81497

Woodford, M. (2012). *Inattentive valuation and reference-dependent choice* (Technical report). Columbia University Creative Commons. Retrieved from https://academiccommons.columbia.edu/catalog/ac:189071 doi: 10.7916/D8VD6XVK

Woodford, M. (2014). Stochastic choice: An optimizing neuroeconomic model. *American Economic Review*, *104*(5), 495–500. doi: 10.1257/aer.104.5.495

Woodford, M. (2016). *Optimal evidence accumulation and stochastic choice* (Technical report). New York, NY: Columbia University. Retrieved from http://www.columbia.edu/~mw2230/DDMASSA2.pdf

Woodford, M. (2017). *Utility-weighted sampling and salience theory.* (Unpublished technical note)

Wright, W. F., & Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organizational Behavior and Human Decision Processes*, *44*(1), 68–82. doi: 10.1016/0749-5978(89)90035-6

Yang, L., Toubia, O., & De Jong, M. G. (2015). A bounded rationality model of information search and choice in preference measurement. *Journal of Marketing Research*, *52*(2), 166–183. doi: 10.1509/jmr.13.0288

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. doi: 10.1016/j.tics.2006.05.002

Zabaras, N. (2010). *Importance sampling* (Technical report). Ithaca, NY: Cornell University.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1–14. doi: 10.3389/fnins.2012.00001

Zhang, Y. C., & Schwarz, N. (2013). The power of precise numbers: A conversational logic analysis. *Journal of Experimental Social Psychology*, *49*(5), 944–946. doi: 10.1016/j.jesp.2013.04.002

A

Resource-Rational Anchoring-and-Adjustment

A.1 NOTATION

X:	numerical quantity to be estimated
$\hat{X}$:	people's estimates of quantity X
n:	number of adjustments
$\hat{X}_n$:	people's estimates of quantity X after n adjustments
K or y:	knowledge or information about X
$P(X|K), P(X|y)$:	posterior belief about X
$P(R|y)$:	distribution of people's responses to observation y
m:	probabilistic model of participants' responses
$\mathrm{cost}(\hat{x}, x)$:	error cost of reporting estimate $\hat{x}$ when the true value is x
$n^{\star}$:	resource-rational number of adjustments
γ:	relative time cost per iteration
c_e, c_t:	cost of time, cost of error
ε:	measurement error
σ_{ε}:	standard deviation of the measurement error ε
Q:	approximate posterior belief
$\mathcal{H}$:	hypothesis space
ψ:	stopping criterion
μ_{prop}:	average size of proposed adjustments
$\mu_{\mathrm{prop}}^{\star}$:	resource-rational step-size of proposed adjustments
a:	anchor

Together, a person's knowledge K about a quantity X, the cost function $\text{cost}(\hat{x}, x)$, and the correct value x define an estimation problem. However, in most environments people are faced with many different estimation problems rather than just a single one, and the true values are unknown. We therefore define a task environment E by the relative frequency $P(X, K, \text{cost}|E)$ with which different estimation problems occur in it. Within each of the experiments that we are going to simulate, the utilities, and the participant's knowledge are constant. Thus, those task environments are fully characterized by $P(X, K|E)$ and $\text{cost}(\hat{x}, x)$.

The optimal speed-accuracy tradeoff weights the costs in different estimation problems according to their prevalence in the agent's environment. Formally, the agent should minimize the expected error cost in Equation 2.2 with respect to the distribution of estimation problems $P(X, K|E)$ in its environment E:

$$t^\star = \arg\max_t \mathbb{E}_{P(X,K|E)} \left[\mathbb{E}_{Q(\hat{x}_t|K)} \left[u(x, \hat{x}_t) - \gamma \cdot t \right] \right]. \tag{A.1}$$

Thus, the number of adjustments is chosen to optimize the agent's average reward rate across the problem distribution of the task environment (cf. Lewis et al., 2014). If the task environment is an experiment with multiple questions, then the expected value is the average across those questions.

A.3   Estimating beliefs

For each simulated experiment we conducted one short online survey for each quantity X that its participants were asked to estimate. For each survey we recruited 30 participants on Amazon Mechanical Turk and asked the four questions Speirs-Bridge et al. (2010) advocate for the elicitation of subjective confidence intervals: "Realistically, what do you think is the lowest value that the ... could be?", "Realistically, what do you think is the highest value that the ... could be?", "Realistically, what is your best guess (i.e. most likely estimate) of the ... ?", and "How confident are you that your interval from the lowest to the highest value could contain the true value o the ... ? Please enter a number between 0 and 100%.". These questions elicit a lower bound (l_s) and an upper bound (h_s) on the value of X, an estimate (m_s), and the subjective probability p_s that X lies between the lower and the upper bound ($P(X \in [l_s, h_s]|K)$ respectively, for each participant s. To estimate peo-

ple's knowledge about each quantity from the reported confidence intervals, we modeled their belief $P(X|K)$ by a normal distribution $\mathcal{N}(\mu_s, \sigma_s)$. We used the empirical estimate m_s as μ_s, and set σ_s to $\frac{h_s - l_s}{\Phi^{-1}((1+p_s)/2) - \Phi^{-1}(1-(p_s+1)/2)}$, where Φ is the cumulative distribution function of the standard normal distribution. Finally, we took the medians of these estimates as the values of μ and σ used in our simulations. We applied this procedure separately for each quantity from each experiment that will be simulated below. The quantities and the estimated beliefs are summarized in Appendix C.

The hypothesis space $\mathcal{H}$ for each quantity was assumed to contain all evenly spaced values (interval $= \frac{\sigma}{20}$) in the range spanned by the 0.5th and the 99.5th percentile of the belief distribution $P(X|K)$ and the anchor(s) plus or minus one standard deviation. We simulated the adjustments people consider by samples from a Poisson distribution, that is $P(\delta = h_k - h_j) = \text{Poisson}(|k - j|; \mu_{\text{prop}})$, where h_k and h_j are the kth and the jth value in the hypothesis space $\mathcal{H}$, and μ_{prop} is the expected step-size of the proposal distribution $P(\delta)$. This captures the intuition that people consider only a finite number of discrete hypotheses and that the adjustments a person will consider have a characteristic size that depends on the resolution of her hypothesis space.

The following tables summarize our estimates of people's beliefs about the quantities used in the simulated anchoring experiments. Since the estimated probabilistic beliefs are normal distributions, we summarize each of them by a mean μ and a standard deviation σ.

Table A.1: Estimated Beliefs: Insufficient adjustment from provided anchors

Study	Quantity	μ	σ	Correct
Tversky, & Kahneman (1974)	African countries in UN (in %)	22.5	11.12	28
Jacowitz, & Kahneman (1995)	length of Mississippi River (in miles)	1,525	770	2,320
Jacowitz, & Kahneman (1995)	height of mount Everest (in feet)	27,500	3,902	29,029
Jacowitz, & Kahneman (1995)	amount of meet eaten by average American (in pounds)	238	210	220
Jacowitz, & Kahneman (1995)	distance from San Francisco to New York (in miles)	3000	718	2,900
Jacowitz, & Kahneman (1995)	height of tallest redwood tree (in feet)	325	278	379.3
Jacowitz, & Kahneman (1995)	number of United Nations members	111	46	193
Jacowitz, & Kahneman (1995)	number of female professors at the University of California, Berkeley	83	251	805
Jacowitz, & Kahneman (1995)	population of Chicago (in millions)	5	3	2.715
Jacowitz, & Kahneman (1995)	year telephone was invented	1885	35	1876
Jacowitz, & Kahneman (1995)	average number of babies born per day in the United States	8,750	15,916	3,952,841
Jacowitz, & Kahneman (1995)	maximum speed of house cat (in mph)	17	10	29.8
Jacowitz, & Kahneman (1995)	amount of gas used per month by average American (in gallons)	55	84	35.2
Jacowitz, & Kahneman (1995)	number of bars in Berkeley, CA	43	55	101
Jacowitz, & Kahneman (1995)	number of state colleges and universities in California	57	112	248
Jacowitz, & Kahneman (1995)	number of Lincoln's presidency	6	2	16

Table A.2: Estimated beliefs: Insufficient Adjustment from self-generated anchors

Study by Epley, & Gilovich (2006)	Quantity	Mean	SD	Correct
Study 1a	Washington's election year	1786.5	7.69	1789
Study 1a	Boiling Point on Mount Everest in F	158.8	36.82	160
Study 1a	Freezing Point of vodka in F	3.7	17.052	-20
Study 1a	lowest recorded human body temperature in F	86	14.83	55.4
Study 1a	highest recorded human body temperature in F	108	3.39	115.7
Study 1b	Washington's election year	1786.5	7.69	1789
Study 1b	Boiling point in Denver in F	201.3	9.93	203
Study 1b	Number of US states in 1880	33.5	8.52	38
Study 1b	year 2nd European explorer reached West Indies	1533.3	33.93	1501
Study 1b	Freezing point of vodka in F	3.7	17.05	-20

Table A.3: Estimated beliefs: Effect of cognitive load

Study by Epley, & Gilovich (2006)	Quantity	Mean	SD	Correct
Study 2b	Washington's election year	1786.5	7.69	1789
Study 2b	second explorer	1533.3	33.93	1501
Study 2c	Washington's election year	1786.5	7.69	1789
Study 2c	second explorer	1533.3	33.93	1501
Study 2c	Highest body temperature	108	3.39	115.7
Study 2c	boiling point on Mt. Everest	158.8	36.82	160
Study 2c	Lowest body temperature	86	14.83	55.4
Study 2c	freezing point of vodka	3.7	17.05	-20
Study 2c	number of U.S. states in 1880	33.5	8.52	38

Table A.4: Estimated beliefs: effects of distance and knowledge

Study	Quantity	Mean	SD	Correct
Russo, & Shoemaker (1989)	year of Atilla's defeat	953.5	398.42	451
Wilson et al. (1996); less knowledgeable group	Number of countries in the world	46.25	45.18	196
Wilson et al. (1996); knowledgeable group	Number of countries in the world	185	35.11	196

Table A.5: Estimated beliefs: Anchor type moderates effect of accuracy motivation; Abbreviations: EG– Epley & Gilovich (2005), TK– Tversky & Kahneman (1974)

Study	Quantity	Mean	SD	Correct
EG, Study 1	population of Chicago	5,000,000	2,995,797.04	2,719,000
EG, Study 1	height of tallest redwood tree	200	76.58	379.3
EG, Study 1	length of Mississippi river (in miles)	1875	594.88	2,320
EG, Study 1	height of Mt. Everest (in feet)	15400	4657.90	29,029
EG, Study 1	Washington's election year	1788	6.77	1789
EG, Study 1	year the 2nd explorer after Columbus reached the West Indies	1507.75	34.34	1501
EG, Study 1	boiling point on Everest (in F)	150.25	36.82	160
EG, Study 1	freezing point of vodka (in F)	-1.25	14.73	-20
EG, Study 2	Washington election year	1788	6.77	1789
EG, Study 2	2nd explorer	1507.75	34.34	1501
EG, Study 2	boiling point on Mt. Everest (in F)	150.25	36.82	160
EG, Study 2	number of US states in 1880	33.5	8.52	38
EG, Study 2	freezing point of vodka (in F)	-1.25	14.73	-20
EG, Study 2	population of Chicago	3000000	1257981.51	2,719,000
EG, Study 2	height of tallest redwood tree (in feet)	200	76.58	379.3
EG, Study 2	length of Mississippi river (in miles)	1875	594.88	2320
EG, Study 2	height of Mt. Everest	15400	4657.90	29,029
EG, Study 2	invention of telephone	1870	54.48	1876
EG, Study 2	babies born in US per day	7875	8118.58	3,952,841
TK	African countries in UN	22.5	11.12	28

Table A.6: Estimated beliefs: effects of direction uncertainty

Simmons et al. (2010), ...	Quantity	Mean	SD	Correct
Study 2	length of Mississippi river (in miles)	1625	752.3	2,320
Study 2	average annual rainfall in Philadelphia (in inches)	36.5	23.80	41
Study 2	Polk's election year	1857.5	45.42	1845
Study 2	Maximum speed of a house cat (miles per hour)	16	9.40	30
Study 2	Avg. annual temperature in Phoenix (in F)	82.75	13.82	73
Study 2	Population of Chicago	2,700,000	1,560,608	2,719,000
Study 2	Height of Mount Everest (in feet)	23,750	7,519.70	29,032
Study 2	Avg. lifespan of a bullfrog (in years)	5.75	6.68	16
Study 2	Number of countries in the world	216.25	77.21	192
Study 2	Distance between San Francisco and Kansas city (in miles)	1,425	547.86	1,800
Study 3b	Year Seinfeld first aired	1991	2.23	1989
Study 3b	Average temperature in Boston in January	26.5	14.86	36
Study 3b	Year JFK began his term as U.S. president	1961.25	2.26	1961
Study 3b	Avg. temperature in Phoenix in Aug.	96	10.21	105
Study 3b	Year Back to the Future appeared in theaters	1985	1.54	1985
Study 3b	Avg. temperature in NY in Sept.	70	10.51	74

A.4   Mathematical models of anchoring-and-adjustment

We developed six probabilistic models of how people estimate numerical quantities. Each model consists of two parts: the hypothesized mechanism and an error distribution.

A.4.1   Bayes-optimal estimation

The first model (m_{BDT}) formalizes the hypothesis that people's estimates are Bayes optimal. According to Bayesian decision theory, the optimal estimate of a quantity X given observation y is

$$\hat{x} = \arg\min_{\hat{x}} \mathbb{E}[\mathrm{cost}(X, \hat{x}) \mid y]. \tag{A.2}$$

The error distribution accounts for both errors in reporting the intended estimate as well as trials in which people do not comply with the task and guess randomly. The model combines these two types of errors with the Bayes-optimal estimate as follows:

$$R = \begin{cases} \hat{x} + \varepsilon, \ \hat{x} = \arg\min_{\hat{x}} \mathbb{E}\left[\mathrm{cost}(x, \hat{x})|y\right], \ \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon), & \text{with prob.} 1 - p_{\mathrm{cost}} \\ R \sim \mathrm{Uniform}(\mathcal{H}), & \text{with prob. } p_{\mathrm{cost}} \end{cases}, \tag{A.3}$$

where R denotes people's responses based on y, p_{cost} is the probability that people guess randomly, $\mathcal{H}$ is their hypothesis space, and ε is people's error in reporting their intended estimate. This model has two free parameters: the probability p_{cost} that people guess randomly on a given trial and the standard deviation of the response error σ_ε. The model's prior distributions on these parameters are

$$p(\sigma_\varepsilon) = \mathcal{U}([0, \max_{h_i, h_j \in \mathcal{H}} |h_i - h_j|]) \tag{A.4}$$

$$p_{\mathrm{cost}} \sim \mathrm{Uniform}([0, 1]). \tag{A.5}$$

A.4.2   Posterior probability matching

Posterior probability matching (m_{PPM}) assumes that people approximate Bayes-optimal estimation by drawing one sample from the posterior distribution $P(X|y)$:

$$\hat{X}_y \sim P(X|y). \tag{A.6}$$

The error model assumes that with probability p_{cost} people guess at random on given trial:

$$P(R = x) = (1 - p_{\text{cost}}) \cdot P(X = x|y) + p_{\text{cost}} \cdot \frac{1}{|\mathcal{H}|} \tag{A.7}$$

This model has only one free parameter: the error probability p_{cost}. The prior on this parameter is the standard uniform distribution:

$$p_{\text{cost}} \sim U([0, 1]). \tag{A.8}$$

A.4.3 ANCHORING-AND-ADJUSTMENT WITH A SIMPLE STOPPING RULE

The anchoring-and-adjustment model with a simple stopping rule (m_{AAs}) starts from an anchor a and adjusts the estimate until its plausibility (i.e. posterior probability) reaches a threshold ψ. We model adjustment as a Markov chain that converges the posterior distribution $P(X|y)$. Consequently, the estimate $\hat{X}_n$ becomes a random variable whose distribution changes $Q(\hat{X}_n)$ depends on the number of adjustments n. The initial distribution assigns all of its probability mass to the anchor a: $Q_0(x) = \delta(x - a)$. The probability $P(\hat{X}_n = h_l|\hat{X}_{i-1} = h_k)$ of adjusting estimate $\hat{X}_n = h_k$ to estimate $\hat{X}_{i+1} = h_l$ is defined as the probability that this adjustment is proposed ($P(X_n^{\text{prop}}|\hat{X}_{n-1})$) times the probability that it will be accepted according to the Metropolis-Hastings algorithm (Hastings, 1970):

$$P(\hat{X}_n = h_l|\hat{X}_{n-1} = h_k) = P(X_n^{\text{prop}} = h_l|\hat{X}_{n-1} = h_k) \cdot \min\left\{1, \frac{p(X = h_l|y)}{p(X = h_k|y)}\right\} \tag{A.9}$$

$$P(X_n^{\text{prop}} = h_k|\hat{X}_{n-1} = h_l) \propto \text{Poisson}(|k - l|; \mu_{\text{prop}}), \tag{A.10}$$

where μ_{prop} is the expected step-size of a proposed adjustment. If the current estimate's plausibility is above the threshold ψ then adjustment terminates. The set of states in which adjustment would terminate is

$$\mathcal{S} = \{h \in \mathcal{H} : P(X = h|y) > \psi\}. \tag{A.11}$$

If the current estimate is not in this set, then adjustment continues. Consequently, the number of adjustments is a random variable and we have to sum over its realizations to computed the distribu-

tion of the estimate $\hat{X}$:

$$Q_{AAs}(\hat{X} = h) = \sum_n Q_{AAs}(\hat{X}_n \in \mathcal{S} \wedge \forall m < n : \hat{X}_m \notin \mathcal{S}) \cdot Q_{AAs}(\hat{X}_n = h | \hat{X}_n \in \mathcal{S}) \tag{A.12}$$

$$Q_{AAs}(\hat{X}_n = x) = \sum_{k=1}^{|\mathcal{H}|} Q_{AAs}(\hat{X}_{n-1} = h_k | \hat{X}_{n-1} \notin \mathcal{S}) \cdot P(\hat{X}_n = x | \hat{X}_{n-1} = h_k). \tag{A.13}$$

As in the posterior probability matching model the response distribution combines takes into account that people guess randomly on some of the trials:

$$P(R = x) = (1 - p_{\text{cost}}) \cdot Q_{AAs}(\hat{X} = x) + p_{\text{cost}} \cdot \frac{1}{|\mathcal{H}|} \tag{A.14}$$

The prior distributions on the models' free parameters are given below:

$$p(\psi) = \exp(-\psi) \tag{A.15}$$

$$p(\mu_{\text{prop}}) = \mathcal{U}([\min_{h_i,h_j \in \mathcal{H}} |i - j|, \max_{h_i,h_j \in \mathcal{H}} |i - j|]) \tag{A.16}$$

$$p_{\text{cost}} \sim \text{Uniform}([0, 1]) \tag{A.17}$$

A.4.4 Anchoring-and-adjustment with a fixed number of adjustments

The anchoring-and-adjustment model with a fixed number of adjustments (m_{AA}) differs from the previous model in that adjustment stops after a fixed, but unknown, number of adjustments (N) regardless of the plausibility of the current estimate:

$$Q_{\text{AA}}(\hat{X}) = Q_{\text{AA}}(\hat{X}_n) \tag{A.18}$$

$$Q_{\text{AA}}(\hat{X}_0 = x) = \delta(x - a) \tag{A.19}$$

$$Q_{\text{AA}}(\hat{X}_n = h_l | \hat{X}_{n-1} = h_k) = P(X_n^{\text{prop}} = h_l | \hat{X}_{i-1} = h_k) \cdot \min\left\{1, \frac{P(X = h_l | y)}{P(X = h_k | y)}\right\} \tag{A.20}$$

$$P(X_n^{\text{prop}} = h_l | \hat{X}_{i-1} = h_k) \propto \text{Poisson}(|l - k|; \mu_{\text{prop}}) \tag{A.21}$$

The error model is the same as before:

$$P(R = x) = (1 - p_{\text{cost}}) \cdot Q_{\text{AA}}(\hat{X} = x) + p_{\text{cost}} \cdot \frac{1}{|\mathcal{H}|} \tag{A.22}$$

The prior distributions on the model parameters are given below:

$$P(N) = \mathcal{U}(\{0, 100\}) \tag{A.23}$$

$$p(\mu_{\text{prop}}) = \mathcal{U}([\min_{h_i, h_j \in \mathcal{H}} |i - j|, \max_{h_i, h_j \in \mathcal{H}} |i - j|]) \tag{A.24}$$

$$p_{\text{cost}} \sim \text{Uniform}([0, 1]) \tag{A.25}$$

A.4.5 ADAPTIVE ANCHORING-AND-ADJUSTMENT

According to the adaptive anchoring-and-adjustment model (m_{aAA}), the mind adapts the expected step-size of its adjustments μ_{prop} and the number of adjustments n. Concretely, the model chooses the optimal combination $(n^\star, \mu_{\text{prop}}^\star)$ of adjustments and the step-size such as to minimize the expected sum of time cost and error cost given given the relative time cost per adjustment γ and the posterior variance σ:

$$Q_{\text{aAA}}(\hat{X} = x) = Q_{\text{aAA}}(\hat{X}_{n^\star} = x) \tag{A.26}$$

$$(n^\star, \mu_{\text{prop}}^\star) = \arg\min_{n, \mu_{\text{prop}}} \mathbb{E}_{P(\mu), P(\sigma)} \left[\mathbb{E}_{\mathcal{N}(\tilde{X}; \mu, \sigma)} \left[\mathbb{E}_{\tilde{Q}(\hat{X}_n; \mu, \sigma)} \left[\text{cost}(\tilde{X}, \hat{X}) \right] \right] \right] + \gamma \cdot n, \tag{A.27}$$

where $\tilde{Q}(\hat{X}_n | \hat{X}_{i-1})$ is the probability to transition from one estimate to the next, if the posterior distribution is a normal distribution with mean μ and standard deviation σ:

$$\tilde{Q}(\hat{X}_n = h_l | \hat{X}_{i-1} = h_k; \mu, \sigma) = P(X_n^{\text{prop}} = h_l | \hat{X}_{i-1} = h_k) \cdot \min\left\{1, \frac{\mathcal{N}(h_l; \mu, \sigma)}{\mathcal{N}(h_k; \mu, \sigma)}\right\} \tag{A.28}$$

$$P(\mu) = P(X), P(\sigma) = \mathcal{U}\left(\sigma; \min_y \sqrt{\text{Var}(X|y)}, \max_y \sqrt{\text{Var}(X|y)}\right). \tag{A.29}$$

The relative iteration cost γ is determined by the time cost c_t, the error cost c_e, and the time $\tau_{\text{adjustment}}$ it takes to perform one adjustment

$$\gamma = \frac{\tau_{\text{adjustment}} \cdot c_t}{c_e}. \tag{A.30}$$

Note that the choice of the number of iterations and the step-size of the proposal distribution is not informed by the distance from the anchor to the posterior mean since this would presume that the answer was already known. Instead, the model minimizes the expected value of the cost under the assumption that the posterior mean will be drawn from the prior distribution. The model also does not presume the shape of the posterior distribution was known a priori; instead it makes a Gaussian approximation with matching mean and variance. Given the number of adjustment and the step-size of the proposal distribution, the adjustment process and response generation work as in the previous model:

$$P(R = x|y) = (1 - p_{\text{cost}}) \cdot Q_{\text{aAA}}(\hat{X}_{n^\star} = x) + p_{\text{cost}} \cdot \frac{1}{|\mathcal{H}|} \tag{A.31}$$

$$Q_{\text{aAA}}(\hat{X}_0 = x) = \delta(x - a) \tag{A.32}$$

$$Q_{\text{aAA}}(\hat{X}_n = h_l|\hat{X}_{i-1} = h_k) = P(X_n^{\text{prop}} = h_l|\hat{X}_{i-1} = h_k) \cdot \min\left\{1, \frac{P(X = h_l|y)}{P(X = h_k|y)}\right\} \tag{A.33}$$

$$P(X_n^{\text{prop}} = h_l|\hat{X}_{i-1} = h_k) \propto \text{Poisson}(|l - k|; \ \mu_{\text{prop}}^\star) \tag{A.34}$$

The prior distributions on the model's parameters are given below:

$$p(\tau_{\text{adjustment}}) = \text{Exp}(\tau_{\text{adjustment}}; \mu = 50\text{ms}) \tag{A.35}$$

$$p(\sigma_\varepsilon) = \mathcal{U}([0, \max_{h_i, h_j \in \mathcal{H}} |h_i - h_j|]) \tag{A.36}$$

$$p_{\text{cost}} \sim \text{Uniform}([0, 1]) \tag{A.37}$$

A.4.6 ADAPTIVE ANCHORING-AND-ADJUSTMENT WITH INTRINSIC ERROR COST

The adaptive anchoring-and-adjustment model with intrinsic error cost (m_{aAAi}) extends the adaptive model m_{aAA} by one parameter: a constant $c_{\text{intrinsic}}$ that is added to the error cost:

$$\gamma = \frac{\tau_{\text{adjustment}} \cdot c_t}{c_e + c_{\text{intrinsic}}} \tag{A.38}$$

The prior over $c_{\text{intrinsic}}$ was

$$p(c_{\text{intrinsic}}) = \text{Uniform}([0, 100]) \tag{A.39}$$

A.4.7 RANDOM CHOICE

According to the random choice model, people's responses are independent of the task and uniformly distributed over the range of all possible responses:

$$R \sim \text{Uniform}(\mathcal{H}) \tag{A.40}$$

B

Utility-Weighted Sampling

B.1 Derivation of the optimal importance distribution for self-normalized importance sampling

One way to derive the optimal importance distribution q for estimating the expected value of f with respect to p, that is $\mathbb{E}_p\left[f(x)\right]$, is to minimize the asymptotic variance (Equation 3.7) of the self-normalized importance sampling estimator (Equation 3.5) subject to the constraints that $\int q(x)\ dx = 1$ and $q(x) > 0$ for all x using variational calculus (Gelfand & Fomin, 2000). To solve this constrained optimization problem we minimize its Lagrangian

$$L(q) = \frac{1}{s} \cdot \int \frac{p(x)^2}{q(x)} \cdot (f(x) - \mathbb{E}_p[X])^2\ dx - \lambda \cdot \int q(x)\ dx, \tag{B.1}$$

where λ is the Lagrange multiplier. To minimize the Lagrangian $L(q)$ we compute its functional derivative

$$\frac{\delta}{\delta q}L(q) = \frac{1}{s}\frac{p(x)^2}{q(x)^2} \cdot (f(x) - \mathbb{E}_p[X])^2 - \lambda, \tag{B.2}$$

403

and set it to zero. Solving that equation for q yields

$$q(x) = \frac{1}{\lambda \cdot s} \cdot p(x) \cdot |f(x) - \mathbb{E}_p[X]|. \tag{B.3}$$

Therefore, the optimal importance distribution for self-normalized importance sampling is proportional to $p(x) \cdot |f(x) - \mathbb{E}_p[X]|$.

B.2 Worked Example of UWS applied to binary decisions from description

Here we provide a worked example of how UWS makes the decision whether or not to accept a gamble. We consider the choice between a gamble with a 90% chance of losing \$1 ($o_1 = -1$) and a 10% chance of winning \$99 ($o_2 = 99$) versus \$1 for sure. Thus, the largest and the smallest possible outcome are $o_{max} = 99$ and $o_{min} = -1$. For the sake of illustration, let's assume that the utility function is like the one defined in Equation 3.16, but deterministic:

$$u(o) = \frac{o}{o_{max} - o_{min}}. \tag{B.4}$$

Hence, the utility of the sure gain is $u(1) = 0.01$, the probability of the gamble's likely loss is $u(-1) = -0.01$ and utility of the gamble's unlikely gain is $u(99) = 0.99$.

If the gamble is chosen, then its first outcome $o_1 = -1$ has a differential utility of $\Delta U(o_1) = u(-1) - u(1) = -0.02$ whereas its second outcome has a large positive differential utility of $\Delta U(o_2) = u(99) - u(1) = 0.98$. Given these differential utilities, we can now compute the distribution the decision-maker should sample from to decide whether or not to take the sure gain by applying Equation 3.20:

$$\tilde{q}(\Delta U = -0.02) \propto p(o_1) \cdot |\Delta u(o_1)| = 0.9 \cdot 0.02 = 0.018 \tag{B.5}$$
$$\tilde{q}(\Delta U = +0.98) \propto p(o_2) \cdot |\Delta u(o_2)| = 0.1 \cdot 0.98 = 0.098. \tag{B.6}$$

To normalize this probability distribution we divide each value by their sum. This yields

$$\tilde{q}(o_1) = \frac{p(o_1) \cdot |\Delta u(o_1)|}{p(o_1) \cdot |\Delta u(o_1)| + p(o_2) \cdot |\Delta u(o_2)|} = 0.1552 \approx 0.16 \tag{B.7}$$
$$\tilde{q}(o_2) = \frac{p(o_2) \cdot |\Delta u(o_2)|}{p(o_1) \cdot |\Delta u(o_1)| + p(o_2) \cdot |\Delta u(o_2)|} = 0.8448 \approx 0.84. \tag{B.8}$$

Table B.1: UWS applied to the decision between a gamble with a 90% chance of losing $1 and a 10% chance of winning $99 versus a sure gain of $1.

Simulated Outcomes	Utilities	Count	Decision	Frequency
(o_1, o_1)	$(-0.02, -0.02)$	-2	decline gamble	$0.1552 \cdot 0.1552 = 2.41\%$
(o_1, o_2)	$(-0.02, +0.98)$	0	accept with prob. 0.5	$0.1552 \cdot 0.8448 = 13.11\%$
(o_2, o_1)	$(+0.98, -0.02)$	0	accept with prob. 0.5	$0.8448 \cdot 0.1552 = 13.11\%$
(o_2, o_2)	$(+0.98, +0.98)$	$+2$	accept gamble	$0.8448 \cdot 0.8448 = 71.37\%$
$P(\text{choose gamble})$				84.48%

This means that UWS would simulate the possibility of losing out on the $99 prize more than 80% of the time even though its probability is only 10%. If the decision-maker generates two samples, then there are four possible simulations results: $(o_1, o_1), (o_1, o_2), (o_2, o_1), (o_2, o_2)$. After the outcomes have been simulated, the UWS heuristic for binary decisions from description determines their utilities and tallies how often the utility is positive minus how often it is negative. If the resulting *count* is positive, then UWS accepts the gamble. If the count is negative, then it declines the gamble, and if the count is zero, then UWS has no preference and chooses at random. All possible outcomes of this process and their respective probabilities are summarized in Table B.1. Summing up the probability of the simulations that lead UWS to accept the gamble reveals that it predicts that about 84.48% of people who are offered the gamble should accept it. This illustrates that UWS can identify the correct decision with high probability using only two simulations.

B.3 Detailed explanation of how UWS explains the fourfold pattern of risk preferences

In this appendix we explain how UWS chooses between a two-outcome gamble and its expected value and show how this gives rise to the fourfold-pattern of risk preferences. These decisions can be formalized as the choice between a $p \cdot 100\%$ chance of winning x and winning nothing otherwise versus the gamble's expected value $p \cdot x$ dollars for sure. As a first step towards explaining UWS we assume that each outcome's utility was equal to its monetary value, that is $u(x) = x$.[*] In this case, the differential utility of choosing a gamble that yields x with probability p over its expected value

---

[*]We will soon return to the stochastic, normalized utility function we used for the simulations reported in Chapter 3.

$p \cdot x$ is

$$\Delta U = \begin{cases} x - p \cdot x & \text{with probability } p \\ -p \cdot x & \text{with probability } 1 - p \end{cases}. \tag{B.9}$$

Thus, the utility-weighted sampling distribution $\tilde{q}$ becomes

$$\tilde{q}(x - p \cdot x) \propto p \cdot |x - p \cdot x| = p \cdot (1 - p) \cdot |x| \tag{B.10}$$

$$\tilde{q}(0 - p \cdot x) \propto (1 - p) \cdot |-p \cdot x| = (1 - p) \cdot p \cdot |x|. \tag{B.11}$$

Note that the two terms are equal. Therefore, if we normalize the distribution we find that

$$\tilde{q}(x - p \cdot x) = \tilde{q}(0 - p \cdot x) = 0.5. \tag{B.12}$$

As our first concrete example, let's consider the choice between a 1% chance of winning $100 versus $1 for sure. In this case, the differential utility of winning is $99 and the differential utility of losing is $-\$1$. Hence, the differential utility of winning the gamble is 99 times as extreme as the differential utility of losing the gamble. Thus, we would intuitively expect UWS to over-simulate winning relative to losing. This is indeed the case since UWS will simulate winning and losing as if they were equally probable (Equation B.12). In this example UWS over-simulates winning because the differential utility of winning ($99 dollars) is more extreme than the disutility of losing $(-\$1)$. As our second concrete example, let's consider the choice between a 99% chance of winning $100 versus $99 for sure. Now the differential utility of winning is $1 whereas the differential utility of losing is minus $99. The sampling distribution is still 50/50. Thus, now UWS over-simulates losing the gamble because the differential utility of losing is 99 times as extreme as the utility of winning. This illustrates that UWS always over-simulates the event whose differential utility is most extreme.

Next, let's work through how the simulations are translated into decisions. For simplicity, let's assume that the decision-maker generates only two samples. In our examples there are two possible outcomes of each of the two simulations. So there are four possibilities in total. Intuitively, these possibilities correspond to (lose, lose), (lose, win), (win, lose), and (win, win). In the first case, the decision-maker would decline the gamble and choose the sure outcome instead. In the second and the third case the decision-maker would not have a systematic preference and their decision would be determined by noise. In the fourth case, that is (win,win), the decision-maker would choose the gamble. Critically, these four simulation results occur with different probabilities. These probabili-

ties depend on the simulation distribution $\tilde{q}$, which in turn depends on the probability p of winning the gamble. Concretely, the probability that UWS will choose the gamble over the sure payoff is the probability of sampling (win,win) plus one half of the probability of sampling (win,lose) or (lose,win).

Table B.2 summarizes the probabilities of the four possible outcomes and the resulting choice frequencies for the general case and the two examples. As this table shows, the probabilities of the four scenarios add up such that the probability of choosing the gamble based on two simulations is equal to the probability to simulate winning the gamble. Consequently, when offered the choice between a 1% chance of winning \$100 versus \$1 for sure, UWS is risk neutral because it chooses the gamble 50% of the time. When offered the choice between a 99% chance of winning \$100 versus \$99 for sure, UWS is also risk neutral and chooses the gamble only 50% of the time. However, when the utility function is non-linear or noisy then the resulting judgments appear to be risk-seeking or risk-averse depending on the problem posed to the decision-maker.

To illustrate this, let's see what happens when we take into account that the brain's representation of value is noisy so that $u(x) = \frac{x}{x_{\max} - x_{\min}} + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 0.17)$. The utility affects two stages of the decision-process: It biases the probability distribution according to which different outcomes will be simulated ($\tilde{q}$) and it is used to judge the value of the simulated outcomes. Since the utility is noisy, both stages are subject to noise. In this example the noise has no systematic effect on the simulation frequencies because $\mathbb{E}[\tilde{q}(u(x) - u(p \cdot x))] = \tilde{q}(x - p \cdot x)$ and $\mathbb{E}[\tilde{q}(u(0) - u(p \cdot x))] = \tilde{q}(-p \cdot x)$. However, the noise in the utility function does systematically bias how the simulated outcomes are translated into a decision. The reason is that the noise ε is more likely to flip the sign of values that are close to zero than the sign of values that are far from zero.

Concretely, for $p = 0.01$, the differential payoff of winning is \$99 whereas the differential payoff for losing is only $-\$1$. The utility function u divides these differential payoffs by the range of possible payoffs ($x_{\max} - x_{\min} = 100$). This transforms these two differential payoffs into $+0.99$ and -0.01 respectively. Next, the noise ε is sampled from a normal distribution with mean zero and standard deviation $\sigma = 0.17$. Thus, for each simulation of losing there is a roughly 48% chance that the sign of its differential utility will be flipped from negative to positive, but the probability that the sign will flip for a simulated win is less than 2 in one billion. This means that if losing is simulated k times, then the probability that the sign will be flipped for at least one of those simulations is $1 - (1 - 0.48)^k$.

From the Technion data set we estimated that the number of samples is $s = 10$. Winning and los-

ing are simulated with equal probability. So a typical value for k would be 5, and when 5 losses are simulated then there is a 96% chance that the sign flips for at least one of them. When this happens in the example where the person simulated 5 wins and 5 losses, then there will be more simulations in favor of the gamble than against it. So the UWS heuristic for binary decisions from description will choose the gamble. This induces risk seeking in the domain of gains when $p < .5$.

By contrast, when the probability of winning is 99% the differential payoff for winning (i.e. $1) is closer to zero than the differential payoff for losing (i.e. $-$99). Therefore, now the noise has exactly the opposite effect, and this induces risk-aversion. Thus, like people, UWS is risk-seeking for improbable gains but risk averse for probable gains. These effects become less extreme as the probability of winning approaches 50% but they do persist. For instance, for the choice between a 30% chance of winning $100 versus $30 for sure, the normalized differential payoff for losing is -0.3, which is still less than two standard deviations of the noise. Consequently, there is an almost 4% chance that its sign will be flipped for a single simulation of losing. This probability is small but its cumulative effect is non-negligible: it entails that when 5 losses are simulated then there is an 18% chance that the sign will be flipped for at least one of them, and this could be enough to make the decision-maker prefer the risky gamble.

Next, let's see how UWS makes decisions in the domain of losses. Let's start by considering the choice between the 1% risk to lose $100 and losing $1 for sure. In this case, the differential utilities for choosing the gamble are $-$99 when the loss occurs versus $ + 1$ when the loss does not occur. The corresponding normalized differential payoffs are $+0.01$ and -0.99. Thus, it is very likely that the addition of noise will flip the sign of the positive outcome into a minus but very unlikely that it would flip the sign of the negative outcome. Therefore, the noise tilts the balance towards negative outcomes and thereby induces risk aversion. Conversely, if we were choosing between a 99% risk of loosing $100 and a sure loss of $99, then the normalized differential payoffs would be -0.01 for the big loss and 0.99 for its absence. Hence, the noise would be very likely to flip the sign of the negative outcome into a plus, but it would almost never flip the sign of the positive outcome. This tilts the balance towards positive outcomes, and thereby induces risk-seeking. Thus, as for people, the risk preferences of UWS flip when the outcomes are framed in terms of losses instead of gains. These examples illustrate that UWS correctly predicts the fourfold pattern of risk preferences. Note that while the noise in the utility function is necessary to get these effects, none of them would occur if the outcomes were simulated according to their actual frequencies. Therefore, the over-simulation of extreme outcomes plays an important role in utility weighted sampling's explanation of the fourfold pattern of risk preferences.

Table B.2: Two worked examples of UWS applied to the choice between a gamble ($\$x$ with probability p) versus its expected value ($p \cdot x$ dollars for sure) for a linear utility function without noise. These predictions change significantly when UWS takes into account that outcome valuation is noisy, as discussed in the text.

Samples	Decision	Frequency	Freq. if $p = 0.01$	Freq. if $p = 0.99$
(win, win)	gamble	$\tilde{q}(x - p \cdot x)^2$	$0.5 \cdot 0.5 = 0.25$	$0.5 \cdot 0.5 = 0.25$
(lose, lose)	sure option	$\tilde{q}(0 - p \cdot x)^2$	$0.5 \cdot 0.5 = 0.25$	$0.5 \cdot 0.5 = 0.25$
(win, lose)	choose randomly	$\tilde{q}(x - p \cdot x) \cdot \tilde{q}(-p \cdot x)$	$0.5 \cdot 0.5 = 0.25$	$0.5 \cdot 0.5 = 0.25$
(lose, win)	choose randomly	$\tilde{q}(-p \cdot x) \cdot \tilde{q}(x - p \cdot x)$	$0.5 \cdot 0.5 = 0.25$	$0.5 \cdot 0.5 = 0.25$
$P(\text{choose gamble})$:		$\tilde{q}(x - p \cdot x)$	0.5	0.5

Furthermore, our model makes the counterintuitive prediction that for choices between a gamble and its expected value the inconsistencies in people's risk preferences increase with the number of simulations. Thus, although increased stakes seem to increase the number of simulations, our model predicts that this will exacerbate people's inconsistent risk preferences rather than ameliorate them. Therefore, in this particular case incentives should increase 'irrationality' instead of reducing it. This is very counterintuitive because it means that people should become more irrational the more they think, and the way to make them more rational would be to encourage them to think less. Testing this prediction is an interesting direction for future research.

B.4 Deal or No Deal: Overweighting of extreme events in real-life high-stakes economic decisions

In Chapter 3, we found that people overweight extreme outcomes in judgment tasks and hypothetical and low-stakes decisions in the laboratory. Is this cognitive bias restricted to artificial laboratory tasks or does it also pervade the high-stakes economic decisions we make in real life? To answer this question, we analyze the high-stakes decisions of contestants in a popular TV game show called "Deal or No Deal" (Post et al., 2008).

In this gameshow, the contestant is presented with up to 26 briefcases that contain prizes between $\$0.01$ and up to $\$5,000,000$. Knowing which prizes are available but not knowing which briefcase contains which prize, the contestant chooses one of the briefcases. In the first round, six of the remaining briefcases are opened and their contents are revealed. This narrows down the prize that might be in the contestant's briefcase to the 20 remaining prizes. Next, the contestant receives a call from a banker offering to buy the contestant's briefcase for a certain amount of money. If

the contestant accepts the offer ("Deal") the game is over and they receive the banker's offer. If the participant rejects the offer ("No Deal"), then the second round begins. In the second round, five additional briefcases are opened and the participant receives a new offer that reflects the change in the expected value of their chosen briefcase brought about by the new information. Whenever the contestant rejects the offer the game proceeds to the next round and the process repeats. In the subsequent four rounds the number of briefcases opened is four, three, two, and one respectively. From there onward one briefcase will be opened on all subsequent rounds. The contestant's chosen briefcase will be opened last, and when it is opened then the participant receives the prize contained therein and the game ends.

Post et al. (2008) extracted the round-by-round options and decisions of 151 contestants from the Netherlands, Germany, and the United States who were on the show between 2002 and 2007. Here, we reanalyze their data set to determine whether contestants overweighted extremely high prizes, such as $5,000,000$, and extremely low prizes, such as $\$0.01$, as predicted by utility-weighted sampling. To answer this question, we performed a formal model comparison between models that do versus models that do not overweight extreme outcomes. The results by Post et al. (2008) indicated that contestants evaluated prizes relative to a reference point that is adjusted gradually. They formalized this insight in a model called *dynamic prospect theory* (DPT). We therefore compared two dynamic reference point models with versus without utility-weighted sampling. In addition, we considered three models without dynamic reference points: a simple baseline model, a basic utility-weighted sampling model, and a basic representative sampling model. All of these models assume that contestants choose between the banker's offer o_r and their unknown prize $\tilde{X}$ based on which prizes $x^{(r)} = \{x_1^{(r)}, x_1^{(r)}, \cdots\}$ were still available in round r.

B.4.1 Models

The baseline model (m_{Random}) has one free parameter p_{accept}. It accepts the bank's offer o_r with probability p_{accept} and rejects it with probability $1 - p_{\text{accept}}$, that is

$$P(A = 1 | o_r, x^{(r)}, m_{\text{Random}}, p_{\text{accept}}) = p_{\text{accept}}. \tag{B.13}$$

The static representative sampling model (m_{RS}) accepts the offer o_r with probability

$$P(A = 1|o_r, x^{(r)}, m_{\mathrm{RS}}) = \Phi\left(\frac{\mathbb{E}\left[\Delta\hat{U}(o_r, x^{(r)})\right]}{\sigma_{\Delta\hat{U}(o_r, x^{(r)})}}\right), \tag{B.14}$$

where $\mathbb{E}\left[\Delta\hat{U}(o_r, x^{(r)})\right]$ is the expected value of the decision-maker's estimate of the difference between the utility of the offer and the utility of the unknown prize, that is

$$\mathbb{E}\left[\Delta\hat{U}(o_r, x^{(r)})\right] = \frac{1}{\#x^{(r)}} \cdot \sum_{x_i^{(r)} \in x^{(r)}} \left(u(o) - u(x_i^{(r)})\right), \tag{B.15}$$

where $\#x^{(r)}$ denotes the number of elements in the set $x^{(r)}$. $\sigma_{\Delta\hat{U}(o_r, x^{(r)})}$ is the standard deviation of this estimate, that is

$$\sigma_{\Delta\hat{U}(o_r, x^{(r)})} = \sqrt{\frac{1}{\#x^{(r)}} \cdot \sum_{x_i^{(r)} \in x^{(r)}} (u(o) - u(x_k))^2}. \tag{B.16}$$

As above, we assume that the utility-function normalizes each payoff by the range of possible outcomes according to efficient coding (Summerfield & Tsetsos, 2015):

$$u(o) = \frac{o}{\max\{x_1^{(r)}, \cdots, x_k^{(r)}\} - \min\{x_1^{(r)}, \cdots, x_k^{(r)}\}} + \varepsilon;\ \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon). \tag{B.17}$$

The static utility-weighted sampling model (m_{UWS}) is an analytic likelihood model that approximates the utility-weighted sampling model for decisions from description. It captures the central assumption that people approximate the expected utility difference in a stochastic fashion that over-weights extreme outcomes:

$$\Delta\hat{U}_{\mathrm{UWS}}(o_r, x^{(r)}) = \sum_{x_i^{(r)} \in x^{(r)}} w_i \cdot \left(u(o) - u(x_i^{(r)})\right), \tag{B.18}$$

where the weight w_i of the i$^{\text{th}}$ potential value of the contestant's prize is defined by

$$w_i \propto P\left(\tilde{X} = x_i^{(r)}\right) \cdot \left|u(o_t) - u(x_i^{(r)})\right|^\gamma. \tag{B.19}$$

This formulation reflects that UWS over-simulates extreme outcomes and only partially corrects for

it. How strongly UWS corrects for the bias of the sampling distribution depends on the number of samples and is captured by the parameter γ. As before, the utility function $u(o)$ normalizes outcomes by the range of the outcomes and adds normally distributed noise (Equation 3.16 in Chapter 3). In order to obtain an analytic expression for the likelihood function we approximate the distribution of $\Delta \hat{U}_{\text{UWS}}(o_r, x^{(r)})$ by a Gaussian with mean $\mathbb{E}\left[\Delta \hat{U}_{\text{UWS}}(o_r, x^{(r)})\right]$ and variance

$$\sigma^2_{\Delta \hat{U}_{\text{UWS}}(o_r, x^{(r)})} = \frac{1}{s} \cdot \mathbb{E}\left[\left(\Delta \hat{U}_{\text{UWS}}(o_r, x^{(r)}) - \mathbb{E}\left[\Delta \hat{U}_{\text{UWS}}(o_r, x^{(r)})\right]\right)^2\right], \tag{B.20}$$

where s is a free parameter that approximately corresponds to the number of samples. Therefore, the likelihood function is given by

$$P(A = 1 \mid o_t, x^{(r)}, m_{\text{UWS}}) = \Phi\left(\frac{\mathbb{E}\left[\Delta \hat{U}_{\text{UWS}}(o_r, x^{(r)})\right]}{\sigma_{\Delta \hat{U}_{\text{UWS}}(o_r, x^{(r)})}}\right). \tag{B.21}$$

The dynamic prospect theory model by Post et al. (2008) extends the utility-function of prospect theory, that is

$$u_{\text{RP}}(o) = \begin{cases} (o - \text{RP})^\alpha, & \text{if } o \geq \text{RP} \\ -\lambda \cdot (\text{RP} - o)^\alpha, & \text{else} \end{cases}, \tag{B.22}$$

by a dynamic model of its reference point (RP). According to this model, people gradually adjust their reference point RP to reflect the expected value of the possible payoffs in the current round, that is $\bar{x}^{(r)} = \frac{1}{\#x^{(r)}} \cdot \sum_{x_i^{(r)} \in x^{(r)}} x_i$. Because the adjustment is gradual, the reference point in round r is still influenced by the expected outcomes of earlier rounds:

$$\text{RP} = B(x^{(r)}) \cdot (\theta_1 + \theta_2 \cdot d_t^{(t-2)} + \theta_3 \cdot d_t^{(0)}), \tag{B.23}$$

where $d_i^{(k)}$ is the relative difference between the average payoff in round i and the average payoff in round k, i.e. $d_i^{(k)} = \frac{\bar{x}^{(i)} - \bar{x}^{(k)}}{\bar{x}^{(k)}}$. Furthermore, the reference point is thresholded from below by the smallest possible payoff in the current round, and from above by the largest possible payoff in the current round. According to this model, the probability that the contestant will accept the deal is

$$P(A = 1 | o_r, x^{(r)}, x^{(r-1)}, x^{(0)}, m_{\text{DPT}}) = \Phi\left(\frac{u_{\text{RP}}(o_r) - \mathbb{E}\left[u_{\text{RP}}(\tilde{X}) \mid \tilde{X} \in x^{(r)}\right]}{\sigma \cdot \sqrt{\text{Var}\left[u_{\text{RP}}(\tilde{X}) \mid \tilde{X} \in x^{(r)}\right]}}\right), \tag{B.24}$$

where σ is a free parameter that determines the choice variability.

The hybrid model $m_{\text{UWS+DPT}}$ extends utility-weighted sampling by the utility function with a dynamic reference point postulated by dynamic prospect theory. This model is a conceptual analogue of the utility-weighted learning model for decisions from description: The utility-weighted learning model gradually adjusts its estimate of the reward expectancy $\bar{u}(t)$ (Equation 3.36 of Chapter 3) which could be interpreted as the reference point of a utility function $\tilde{u}(o) = r(o) - \bar{u}(t)$. In utility-weighted learning, it is the absolute value of $\tilde{u}$ with its dynamic reference $\bar{u}(t)$ that determines the probability weighting according to $\tilde{q}(o) = p(o) \cdot |\tilde{u}(o)|$ just like in the hybrid model. The hybrid model's decision variable is

$$\Delta \hat{U}_{\text{UWS+DPT}}(o_r, x^{(r)}) = \sum_{x_i^{(r)} \in x^{(r)}} w_i \cdot \left(u_{\text{RP}}(o) - u_{\text{RP}}\left(x_i^{(r)} \right) \right), \tag{B.25}$$

where the weight w_i of the ith possible prize is defined by

$$w_i \propto P(\tilde{X} = x_i^{(r)}) \cdot \left| u_{\text{RP}}(o_r) - u_{\text{RP}}(x_i^{(r)}) \right|^{\gamma}. \tag{B.26}$$

The model's choice probability is thus given by

$$P(A = 1 \mid o_r, x^{(r)}, x^{(r-1)}, x^{(0)}, m_{\text{UWS+DPT}}) = \Phi \left(\frac{\mathbb{E}\left[\Delta \hat{U}_{\text{UWS+DPT}}(o_r, x^{(r)}) \right]}{\sigma_{UWS+DPT}} \right), \tag{B.27}$$

where $\sigma_{UWS+DPT} = \frac{1}{s} \cdot \sqrt{\text{Var}\left[\Delta \hat{U}_{\text{UWS+DPT}}(o_r, x^{(r)}) \right]}$.

B.4.2 Priors distributions on parameters of the models of the Deal No Deal dataset

For parameters that occurred in multiple models, the prior was always the same across all models.

For the random choice model the prior on the choice probability was the standard uniform distribution over the interval $[0, 1]$:

$$p(p_{\text{accept}}) = \begin{cases} 1, & \text{if } 0 \leq p_{\text{accept}} \leq 1, \\ 0, & \text{else} \end{cases}. \tag{B.28}$$

For the representative sampling model the prior on the noise parameter σ_ε of the stochastic utility function was a standard uniform distribution over the range $[0, 1]$ because 0 corresponds to no noise whereas 1 would entails that the magnitude of the noise is as high as the highest possible expected utility gain. The prior on the number of samples s was a uniform distribution over the range $[1, 1000]$ because the minimum number of samples is 1 and 1000 would be more than sufficient to estimate the expected utility gain accurate.

For the basic utility-weighted sampling model the prior on the utility-weighting parameter γ was a standard uniform distribution over the range $[0, 1]$ because 0 corresponds to no bias and 1 correspond to drawing only a single sample. The priors on the variability parameter σ and the number of samples s were the same as for the representative sampling model.

For the dynamic prospect theory models, the prior distribution on the exponent α of the utility was the uniform distribution over this parameters admissible range $[0, 1]$, because the utility function is concave if and only if $0 \leq \alpha < 1$ and linear for $\alpha = 1$. The prior distribution on the slope in the domain of losses λ was defined to be uniform distribution over the range $[0, 5]$ because it cannot be negative and the estimates obtained by Nilsson, Rieskamp, and Wagenmakers (2011) suggested that it is always smaller than five. The prior on the weights $\theta = (\theta_1, \theta_2, \theta_3)$ determining the updates of the reference point was a multivariate standard normal distribution:

$$p(\theta) = \mathcal{N}\left(\theta; \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right). \tag{B.29}$$

The prior on the noise parameter σ was an exponential distribution with mean 1 to express that the expected variability is the variance that is multiplied by σ and that less noise is more likely than more noise:

$$p(\sigma) = \exp(-\sigma). \tag{B.30}$$

For the combined model integrating UWS with dynamic prospect theory the priors on its parameters were the same as those reported above: The priors on the parameters of the utility function $(\alpha, \lambda, \theta)$ were the same as for the DPT model and the priors on the utility weighting parameter (γ) and the choice variability parameter σ were the same as those in the basic UWS model.

B.4.3 Results

We estimated the model parameters from all choices of the 151 contestants from the Netherlands, Germany, and the US using the maximum-a-posteriori method with the priors specified in Appendix D. To find these estimates we used a global optimization algorithm known as infinite-metric Gaussian process optimization (Kawaguchi et al., 2015). For all models this optimization algorithm was run for 1000 iterations. We then use the global maximum found by this derivative-free algorithm as the starting point for the gradient-based quasi-Newton algorithm (fminunc in Matlab 2015b) which was run until convergence. To find out which of these five models best explains people's choices in this high-stakes game show, we performed Bayesian model selection (Kass & Raftery, 1995) with a uniform prior over the five models. This method measures the goodness of each model by the marginal likelihood of the data given that model, which integrates over all possible settings of the model's parameters. The marginal likelihood thereby penalizes each model's fit by a complexity penalty that accurately reflects the model's flexibility and not just its number of parameters. Here, we estimate the marginal likelihood of each model using the Laplace approximation (Tierney & Kadane, 1986). Bayesian model selection then compares pairs of models by computing their Bayes factor (BF), which is the ratio of their posterior probabilities given the data.

Figure B.1 shows the results of the model comparison. Consistent with the results of Post et al. (2008) we found that models with a dynamic reference point explained the contestants' decisions better than models with a fixed utility function. Most importantly, utility-weighted sampling performed better than unweighted decision mechanisms for either type of utility function: For models with the static, normalized stochastic utility function (Equation 3.16 of Chapter 3), we found that our basic utility-weighted sampling model explained the contestants' choices substantially better than random choice ($BF_{UWS, random} = 3.7 \cdot 10^{50}$) or representative sampling ($BF_{UWS, RS} = 2.3 \cdot 10^{18}$). Among the models with dynamic utility functions, utility-weighted sampling with a dynamic reference point explained the contestants' choices substantially better than the unweighted decision mechanism of dynamic prospect theory ($BF_{DPT+UWS, DPT} = 1.1 \cdot 10^7$). In both cases, the data provided decisive evidence for utility-weighted sampling, because the Bayes factors are larger than 100 (Kass & Raftery, 1995). Furthermore, the models with the dynamic utility function captured the data significantly better than their counterparts with the static utility function ($BF_{DPT+UWS, UWS} = 1.4 \cdot 10^{16}$, $BF_{DPT, RS} = 5.6 \cdot 10^{30}$). Post et al. (2008) used a different task analysis according to which contestants choose between the current offer and anticipated next offer. To evaluate this alternative perspective, we adapted all models to their alternative task analysis and recomputed the model evidence scores. Quantitative model comparisons provided very strong
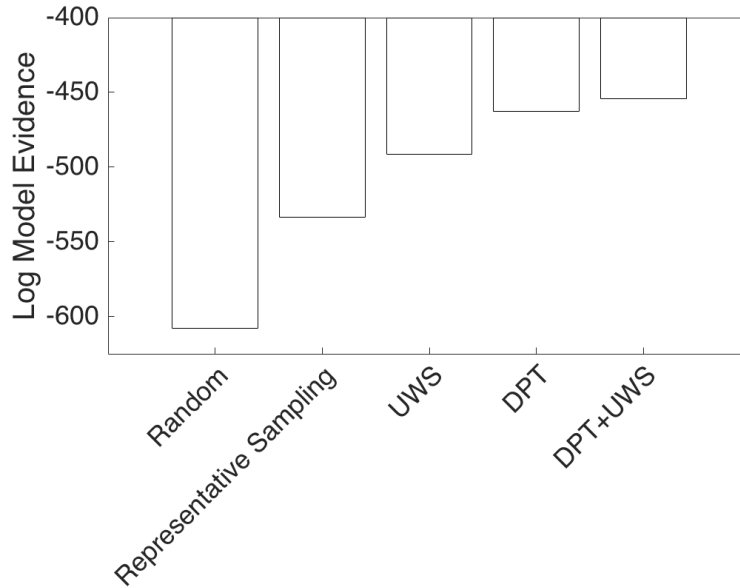
Figure B.1: Model comparison for "Deal No Deal" data set. Better models have a *higher* log-model evidence.

evidence for our task analysis over the one by Post et al. (2008).

Our analyses support the hybrid model ($m_{\text{UWS+DPT}}$) that combines utility-weighted sampling with a utility function with a gradually adjusting reference point. For this model, the estimated utility-weighting coefficient $\hat{\gamma}$ was significantly larger than zero ($\hat{\gamma} = 0.5721$, 95% CI: $[0.5668; 0.5774]$). This is consistent with the hypothesis that contestants performed utility-weighted sampling with an intermediate number of samples. Furthermore, fixing the probability-weighting parameter to 0, which yields the DPT model, led to a significantly worse fit that is not offset by the corresponding gain in parsimony. The estimated value of the number of samples was $s \approx \frac{1}{\sigma^2} = 11.88$ suggesting that contestants simulated their potential prize about 12 times on average. Note that in UWS some of these imagined outcomes would have been identical so that the number of considered prizes can be smaller. Note also, that s only approximately corresponds to the number of samples, because the number of samples is also reflected by the value of γ. The maximum-a-posteriori estimates for the remaining parameters were $\hat{\alpha} = 0.6721$, $\hat{\lambda} = 1.1346$, and $\hat{\theta} = (1.0049, -0.0070, -0.0313)$. For these parameter values, the hybrid model correctly predicts 87.1% of the contestants choices, meaning that 87.1% of the time the predicted probability of the contestant's choice was greater than 0.5.

416

B.4.4 Discussion

In conclusion, we found that people overweight extreme potential outcomes not only in hypothetical and low-stakes laboratory tasks but also in high-stakes real-life decisions whose outcomes do count. This finding is consistent with utility-weighted sampling. In fact utility-weighted sampling predicts that the overweighting of extreme outcomes is larger for high-stakes decisions than for low-stakes decisions, because their highest possible outcomes are more extreme. However, we cannot conclude that the contestants' choices were resource-rational because the normative status of the dynamic reference point of the winning model's utility function is unclear. On the one hand, the reference point can be seen as an estimate of the expected utility gain $\mathbb{E}[\Delta U]$. Therefore, the difference between what would otherwise be the outcome's utility and the reference point can be interpreted as an approximation to the term $u(o) - \mathbb{E}_p[\Delta U]$ used in optimal importance sampling (Equation 3.10 in Chapter 3). Hence, the model's use of the absolute value of the reference-point-dependent utility to weight the probabilities of the corresponding outcomes can be interpreted as an approximation to optimal importance sampling (Equation 3.10 in Chapter 3). Since the reference point is an estimate of the expected utility gain, it is rational to update it when additional outcomes are observed. The update equation for the dynamic reference point emphasizes recent outcomes. This is consistent with estimating the expected utility gain of decisions in dynamic environments like Deal No Deal where the expected utility gain of future decisions changes every time an outcome is observed. Therefore, the winning model could be a rational extension of UWS to dynamic decision environments. However, this rational interpretation has to be taken with a grain of salt, because the update rule for the dynamic reference point was not derived from first principles, and the normative status of other aspects of the utility function is also unclear. Deriving a fully-principled form of UWS for dynamic environments and testing it against the models examined here is a possible direction for future work.

Interestingly, the estimated number of samples (s) was substantially higher for the high-stakes decisions in *Deal or No Deal* than for the low-stakes decisions in the Technion choice prediction competition. This finding is consistent with the hypothesis that people make rational use of their finite time and limited computational resources: Raising the stakes increases the expected gain in reward for performing an additional simulation but its time cost remains the same. Once the expected gain in reward exceeds the time cost, it becomes resource-rational to perform an additional simulation.

The decisions from experience simulated above were very simple in that each option had only two possible outcomes, but in the real world a choice can have very many outcomes. To investigate whether utility-weighted sampling can capture these more complex decisions from experience, we simulated Experiment 1 by Barron and Erev (2003) where outcomes were sampled from normal distributions with different means and variances. Participants were instructed to maximize their earnings by repeatedly choosing between two buttons but received no further information about the task other than that the experiment would last for about 30 minutes. After each decision an outcome was sampled from the chosen option's payoff distribution and shown to the participant. There were three groups who made 200 choices each: In the first condition the outcome of the first option was sampled from a normal distribution with mean 25 and standard deviation 17.7, and the outcomes of the second option were sampled from a normal distribution with mean 100 and standard deviation 354. The second condition was like the first, except that both means were shifted upwards by 1000. The third condition was like the second one except that the standard deviation of the second option was reduced to 17.7. Barron and Erev (2003) found that the high variability of the payoffs in the first and second condition interfered with people's ability to discover that the first option was better than the second option. This is known as the *payoff-variability effect*.

We simulated the experiment with the parameters estimated from the experiment by Madan et al. (2014). The largest and the smallest possible outcome (o_{max}^c and o_{max}^c in Equation 3.16 of Chapter 3) were initialized by ± 10 and continuously updated to always equal the largest and smallest outcome observed so far respectively. We conducted 1000 simulations for each of the 3 conditions. Figure B.2 shows the average frequency with which our model choose the option with the higher expected value as a function of time in the experiment. To evaluate the effect of learning we compared the average choice frequencies between the first 5 trials and the last 100 trials. We found that the model captures the outcome variability effect (see Figure B.2): When the payoff variability of the better option was large compared to the expected values and their difference (Condition 1), then participants came to avoid the better option as their choice frequency dropped from 51.3% to 43.3% ($\chi^2(1) = 123.17, p < 10^{-15}$). When the means were increased to be substantially higher than the payoff variability (Condition 2), then the frequency of the maximizing choice increased slightly to 49.37% ($\chi^2(1) = 740.4, p < 10^{-15}$) but remained below chance level ($p < .0001$), and their choice frequency did not change significantly over time ($\chi^2(1) = 3.07, p = 0.08$). But when the payoff variability was reduced (Condition 3), then people learned to choose the better option:
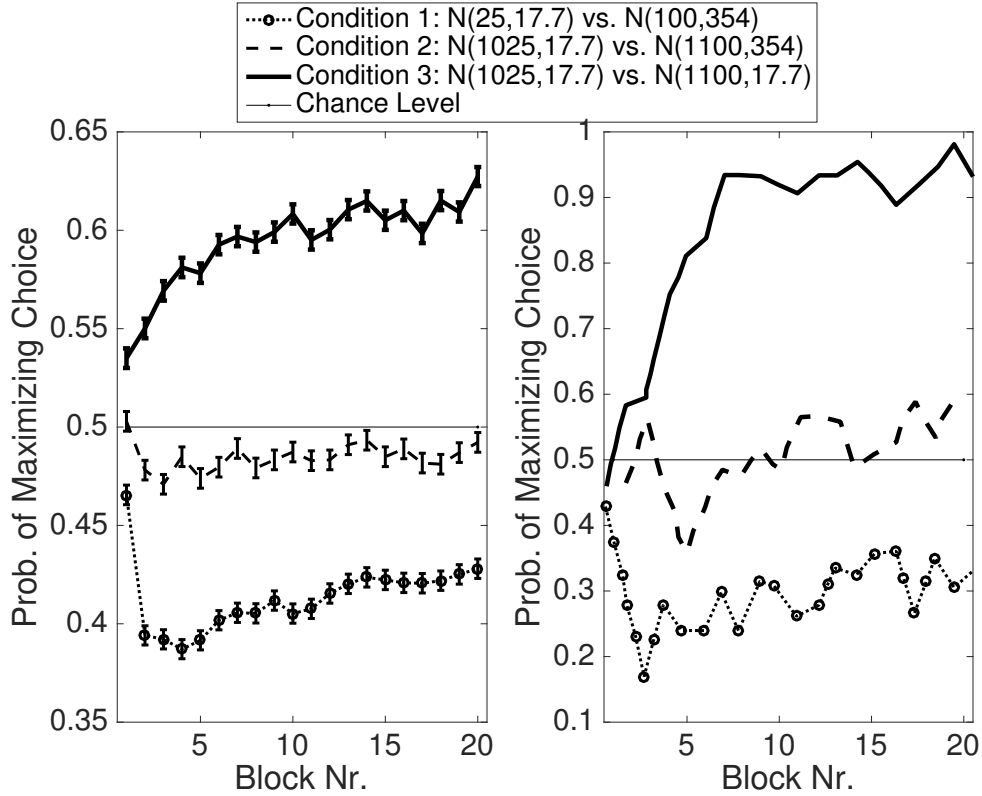
Figure B.2: Simulation of Experiment 1 by Barron and Erev (2003) according to the utility-weighted learning model. Each line represents the frequency of choosing the first option in each of the 20 blocks averaged across 1000 simulations. The error bars indicated standard errors of the mean.

the predicted frequency of choosing the better option rose significantly from 51.5% in the first five trials to 60.8% in the last 100 trials $(\chi^2(1) = 174.5, p < 10^{-15})$ and surpassed the chance level $(p < 10^{-15})$. Thus, our utility-weighted learning model correctly predicted the detrimental effects of payoff variability on decisions from experience.

This illustrates that utility-weighted sampling can capture people's ability to make decisions with (infinitely) many possible outcomes as well as people's biases in the face of high payoff variability. According to utility-weighted sampling, people's apparently irrational aversion to choices with superior expected value but higher payoff variability in decisions from experience arises because people overweight the salient memories of large losses.

We found that the average risky-choice frequency of the UWL model was $41.6 \pm 2.0\%$ whereas the average risky-choice frequency of people was $38.1 \pm 2.2\%$. This shared overall preference for the safe option suggests that utility-weighted learning captures that people underweight rare gains in classic decisions-from-experience paradigms. However, according to a paired t-test, the predictions of UWL were significantly less risk averse than people ($-3.5 \pm 1.4\%, t(59) = 2.59, p = 0.01$). This apparent bias towards risk seeking does, at least in part, result from a regression towards the "mean" frequency of 50%. Consistent with this interpretation, UWS was less risk averse than people primarily when they chose the risky option less than half of the time (36.47% vs. 31.05%; $t(46) = 3.68, p < .0006$), but when they chose it more than 50% of the time, then UWL was less risk-seeking than people (60.14% vs. $63.58\%^{\dagger}$). For this particular data set, regression to the mean increased the overall frequency of choosing the risky option because people were risk averse in 47 of the 60 problems, and chose the risky option only 38% of the time on average.

To understand why the UWL model's risk preferences were less extreme than human risk preferences, we inspected the decision problems on which UWL was much more risk-seeking than people. We found that the two problems where the bias was largest were the only problems in which the risky option was dominated by the safe option. In these problems the outcome of the safe option was slightly higher than the best possible outcome of the risky option that occurred with a frequency of 97%. Here, people chose the dominated risky option only about 15% of the time, whereas UWL chose it 40% of the time. The choice frequency of UWL was closer to 50% because the difference between the safe outcome and the high outcome of the risky option was small relative to the noise of its utility function.

Examining these results, it seems that people can exploit obvious dominance better than UWL. For instance, when people recognize dominance they can switch to a different decision strategy (Lieder & Griffiths, 2015, 2017). People's advantage on problems with obvious dominance contributed to the apparent bias of UWL, because the safe option dominated the risky option twice as often as vice versa. When the three problems with dominance were excluded, the bias decreased to 2.9% but remained statistically significant ($t(56) = -2.29, p = 0.0257$). We therefore also inspected the problem where UWL had the third largest bias towards risk seeking. In this problem

---

†This difference was not statistically significant ($t(12) = -1.37, p = 0.20$), but the test was highly underpowered because people were risk-seeking for only 13 of the 60 problems.

the probability of the high payoff was very low ($p_{\text{high}} = 0.06$), and the low payoff (o_{low}) differed from the sure payoff (o_{sure}) by less than 2% of its value (-20.5 vs. -20.3). For this problem many participants may thus never have sampled the high payoff. This would again create the dominance scenario in which the noisy utility function of UWL induces more random choices, and hence more risk-seeking, than the heuristic that people appear to use for problems with dominance. When we additionally removed the six problems where the safe option was very likely to slightly dominate the risky option according to the sampled outcomes ($p_{\text{high}} < 0.1$ and $0 < \frac{o_{\text{sure}} - o_{\text{low}}}{\max\{|o_{\text{sure}}|, |o_{\text{high}}|\}} < 0.025$), then the average difference in the frequency of risk-seeking dropped to 1.5% and was no longer statistically significant ($t(49) = -1.0, p = 0.32$).

Taking these results into account, it appears that the risk-seeking bias we observed in the predictions of the UWL model may arise from situations where the safe option dominates the risky one according to their observed outcomes. On those trials UWL does not capture people's choices frequencies. One possible reason for this is that the model's assumptions about the normalized, stochastic utility function are invalid when the difference between the observed outcomes is very small relative to the range of possible outcomes. Another possible reason is that people switch to a specialized heuristic when they encounter dominance. Investigating these possibilities is an interesting direction for future research.

B.7 UWS CAPTURES THAT PEOPLE'S PERFORMANCE APPROACHES OPTIMALITY AS THE OPTIONS BECOME MORE DIFFERENT

Resource-rationality predicts that as the stakes increase people should become increasingly more accurate. Consistent with this prediction, Jarvstad, Hahn, Rushton, and Warren (2013) found that as the difference between two gambles' expected values increases people's decision quality increases gradually. To test whether UWS can capture this effect, we simulated the decisions from description experiment from Jarvstad et al. (2013) according to our binary choice model with the parameters estimated from the Technion tournament for decisions from description. As shown in Figure B.3, our model captures that people err primarily when the options' expected values are very close but come to choose the optimal action almost 100% of the time as the difference in expected value increases (Jarvstad et al., 2013). These findings suggest that the biases and suboptimalities that classic laboratory experiments have demonstrated for choices between options with very similar expected values are not representative of decision-making in the real world where the (relative and absolute) difference in the options' expected values tends to be larger. Instead, the fact that the difference in
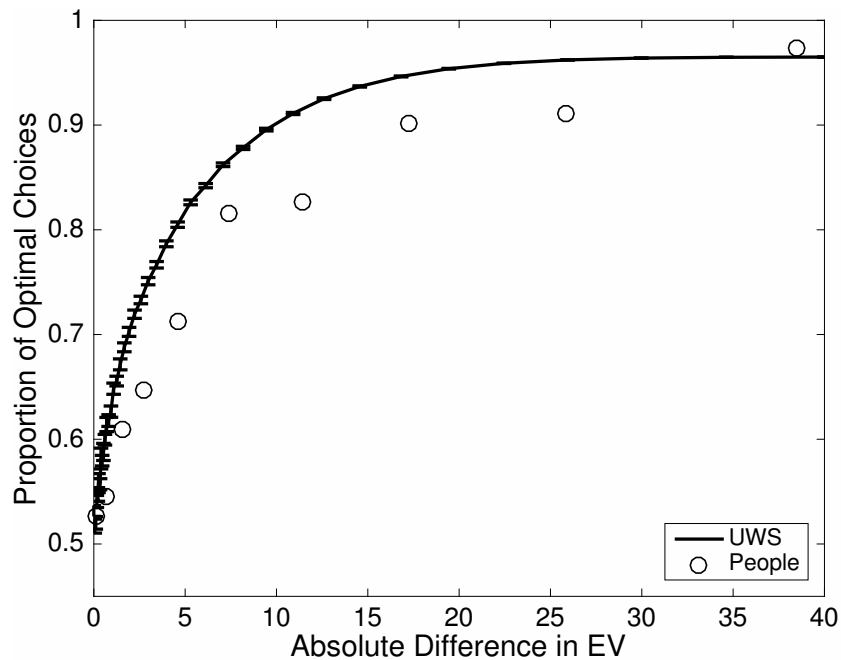
Figure B.3: UWS captures that people reach (near) optimal performance as the difference between the options' expected values increases. The human data was taken from Jarvstad, et al. (2013). Error bars enclose 95% confidence intervals.

expected value has to be small to elicit biases and sub-optimalities in people is consistent with the rational use of limited cognitive resources. Indeed, it is resource-rational to save time and mental effort when the return for investing additional cognitive resources is less than their cost.

B.8 Comparison to previous theories of memory, judgment, and decision-making

Comparison to previous theories of memory and frequency judgment   Anderson's rational analysis of memory demonstrated that the availability of a memory rationally reflects how likely it is going to be needed according to its frequency and recency of occurrence (Anderson, 1990, 1991). Here, we have demonstrated another rational aspect of availability: eventualities that are more important for making a decision are more available in memory than their equally probable counterparts. We have shown that the rational availability of extreme events can account for the memory biases observed by Madan et al. (2014). Our model of frequency judgments is consistent with the availability-by-recall model (Hertwig et al., 2005; Pachur et al., 2012) of the availability heuristic

(Tversky & Kahneman, 1973), but it goes one step further by predicting how many instances of each event people will recall from memory and how this number depends on the event's frequency and extremity. This allowed our model to correctly predict that people overestimate the frequency of extreme events regardless of whether they are rare as in Experiment 1 or frequent as in the Experiment by Madan et al. (2014). Our theory thereby reconciles seemingly irrational availability biases with Anderson's rational analysis of memory, and our results resolved the open question whether biases in frequency estimation are due to availability or regression to the mean (Hertwig et al., 2005) in favor of a rational version of availability.

Comparison to previous theories of decisions from experience   Which events are retrieved from memory is critical to people's decisions from experience. Several models of experience based choice assume that memory recall rationally reflects past experience (Lejarraga et al., 2012; N. Stewart et al., 2006) and this is also true of the exploratory sampler with recency that won the Technion choice prediction competition (Erev et al., 2010). Concretely, instance-based learning theory (C. Gonzalez & Dutt, 2011; C. Gonzalez, Lerch, & Lebiere, 2003; Lejarraga et al., 2012) assumes that previous instances of similar past decisions are recalled with a probability that reflects their frequency and recency according to Anderson's rational analysis of memory (Anderson, 1990, 1991). Our analysis suggests that these models' assumption of rational memory recall implies that events with extreme utilities should be recalled more frequently than would be warranted by how often they have been encountered in the past, whereas equally frequent events with unremarkable utilities should be recalled less often. Other models of decisions from experience emphasize that people's memory is fallible (Hawkins, Camilleri, Heathcote, Newell, & Brown, 2014; Marchiori, Di Guida, & Erev, 2015). The Technion choice prediction tournament also included reinforcement learning models and an ACT-R model of instance-based learning, and Plonsky et al. (2015) have proposed a new model according to which people's decision mechanisms are tuned to dynamic environments. Yet, as far as we know, no previous model of decisions from experience captures the over-weighting of events with extreme utilities. We now compare our utility-weighted learning model to each of these theories in turn

Decision-by-sampling assumes that outcomes are sampled from memory in a manner that reflects the structure of the environment but is also subject to availability biases (N. Stewart et al., 2006). This view is consistent with our model but decision-by-sampling does not explain why some past experiences are more available than others. The explorative sampler with recency (Erev et al., 2010; pp. 29-31) stochastically chooses to explore or to exploit. When it explores, it chooses at random. When

it exploits, it estimates each option's value and chooses the option with the highest value estimate. To estimate an alternative's value the sampler retrieves a randomly generated number of past experiences with that alternative from memory. The retrieved experiences always include the most recent outcome and all earlier experiences are retrieved with equal probability. The retrieved outcomes are regressed towards the mean outcome and passed through a concave utility function. In the Technion choice prediction competition, the performance of the exploratory sampler with recency was not significantly higher, and its lower mean-squared deviation might reflect that it captures that people face an exploration-exploitation dilemma and assume that the environment is changing. Incorporating this idea into the UWL model might lead to even better predictions.

The exemplar-confusion model by Hawkins et al. (2014) assumes that people store a new memory trace every time they experience an outcome. Every time a new memory trace is added to the store of the chosen lottery, every stored memory trace has a small probability that its outcome will be confused. It this happens, then that memory's outcome will be replaced by a value that is sampled uniformly at random from the set of values that have been observed for that lottery so far regardless of how often each value has been observed. This model predicts both choices and probability judgments by assuming that people average over all of their memory traces. When evaluated on the Technion choice prediction tournament for repeated decisions from experience, the exemplar-confusion model's risk preferences agreed with people's risk preferences slightly less often than the risk preferences of our utility-weighted learning model (83.3% vs. 90% agreement). While the exemplar confusion model focuses on errors during encoding, the noisy retrieval models focus on errors during retrieval (Marchiori et al., 2015). Concretely, noisy retrieval models assume that people retrieve only a very small number of experienced outcomes and erroneously recall outcomes of unrelated decisions and use them as if they pertained to the current choice. These models reconcile the under-weighting of rare events in repeated decisions from experience with their being over-weighted in one-shot decisions under risk, and the overestimation of their probabilities. However, none of these models captures the effect of extremity on memory recall, frequency judgments, and choice. Furthermore, in contrast to these theories, UWS is based on a rational model of memory.

The basic reinforcement learning model from the Technion choice prediction tournament (Erev et al., 2010) probabilistically chooses the option with the higher recency-weighted average payoff, and the normalized reinforcement learning model normalizes the options' weighted average values by the variability of their payoffs. The value assessment model by Barron and Erev (2003) and the model by Shteingart, Neiman, and Loewenstein (2013) are similar to the basic reinforcement learning model. The main difference in the model by Shteingart et al. (2013) is that it gives special

weight to the first outcome of each action, and the model by Barron and Erev (2003) includes a separate exploration mechanism and a utility function that captures loss aversion. These models differ from UWL in that they do not simulate potential outcomes and do not overweight extreme outcomes relative to moderate outcomes. As reported above, our model achieved a significantly lower mean-squared deviation than the basic reinforcement learning model. While the normalized reinforcement learning model was about as accurate as our UWL model by assuming that payoff variability has a deterring effect, our model provides a mechanistic explanation for why this was the case for many problems in the choice prediction competition. As reported above, the instance-based learning model by Lejarraga et al. (2012) predicted decisions in the Technion choice prediction tournament significantly better than our UWL model. This might be because it incorporates additional psychological insights such as people's optimism in the face of uncertainty and the implicit assumption that the environment is changing. Incorporating these assumptions into the UWL model or incorporating the heightened availability of extreme events into the instance-based learning model might lead to even better predictions. The ACT-R model of instance-based learning (T. C. Stewart et al., 2009) was very similar to the model by Lejarraga et al. (2012). The main difference was that the ACT-R model learned separately about the contexts established by the history of previous choices and outcomes. Concretely, the model by T. C. Stewart et al. (2009) recalls only those previous outcomes that followed the sequence of choices and outcomes observed in the preceding k trials. Like, the ACT-R model of instance-based learning, the contingent average and trend (CAT) model by Plonsky et al. (2015) postulates that people assume that the same choice will lead to different outcomes depending on the outcomes that preceded it. Concretely, this model assumes that people learn a separate reward expectation for every possible sequence of the k preceding outcomes. In addition, the model probabilistically responds to trends: If the last three outcomes suggest an increase or decrease in an action's average payoff, then the CAT model estimates the expected value of that action by its most recent payoff with some probability. This model captures people's sensitivity to patterns, the underweighting of rare events, and the non-monotonic effect of recency on the weight of previous outcomes (Plonsky et al., 2015). The CAT model is complementary to UWS in the sense that it describes how people learn when they assume that the environment is dynamic whereas UWL describes how people learn when they assume that the environment is static. The two theories could be combined into an integrated model of utility-weighted learning in dynamic environments. Overall, comparing UWL to models of decisions from experience highlights that extending our model to dynamic environments is an important direction for future work.

Comparison to previous theories of decisions from description   Most descriptive theories of decisions from description modify expected utility theory in order to account for some of the ways in which people deviate from its predictions (Starmer, 2000). Some of these theories postulate that people's choices optimize not only the expected utility of their payoffs but also additional experiential qualities like regret (Loomes & Sugden, 1982) or disappointment (Bell, 1985; Loomes & Sugden, 1984, 1986), or assume that people have additional preferences about the variance (Allais, 1979) and skewness (Hagen, 1979) of a prospect's outcome distribution. Other theories maintain that people maximize their subjective expected utility with respected to weighted probabilities (Edwards, 1962; R. Gonzalez & Wu, 1999; Quiggin, 1982). By contrast to all of these theories, utility-weighted sampling is derived from the assumption that people are striving to maximize the expected utility of their outcomes but are constrained by their finite time and limited cognitive resources. Hence, unlike these earlier theories, UWS does not propose that people behave as if they were optimizing a certain preference function. Instead, UWS is a procedural theory. Like prospect theory (Kahneman & Tversky, 1979), cumulative prospect theory (Tversky & Kahneman, 1992), dynamic prospect theory (Post et al., 2008), rational inattention theory (C. A. Sims, 2003), and salience theory (Bordalo et al., 2012), it is informed by people's limited cognitive resources, but it goes beyond these theories by providing a decision strategy that is optimal given the constraints imposed by those limited resources under certain assumptions. There is a profound structural similarity between the application of UWS to binary decisions from description and salience theory that causes both theories to always agree on which option should be chosen most frequently (Woodford, 2017). However, the stochastic component of UWS allows it to additionally explain the effect of choice difficulty on how often the better option is chosen. Most importantly, salience theory is a theory of decisions from description only[‡] whereas UWS also applies to memory recall, judgment, and decisions from experience. While rational inattention (C. A. Sims, 2003) prescribes how much time and attention a decision-maker should allocate to each of their choices, but it does not specify how that decision should be made. Utility-weighted sampling complements rational inattention by specifying a decision strategy that makes the best possible use of the limited amount of attention that has been allocated to a choice. Conversely, rational inattention complements UWS by specifying how many samples it should generate.

What sets UWS apart from all of theories mentioned above, is that it provides a process model.

[‡]While Bordalo, Gennaioli, and Shleifer (2017) developed a model that combines salience theory with memory retrieval, salience plays no role in the assumed memory mechanism. Rather than explaining memory biases, their model uses a standard memory mechanism to explain where the reference point of the utility function that drives salience comes from.

Process models of decisions from description that are similar to UWS include the *priority heuristic* (Brandstätter et al., 2006), *decision-by-sampling* (N. Stewart et al., 2006), the *exemplar-confusion* model (D. Lin, Donkin, & Newell, 2015), query theory (Johnson et al., 2007; Weber et al., 2007), *selective integration* (Tsetsos et al., 2016), *drift-diffusion models* of value-based choice (Krajbich, Armel, & Rangel, 2010; Krajbich & Rangel, 2011; Shadlen & Shohamy, 2016), and the *associative accumulation model* (Bhatia, 2013). We now discuss the similarities and differences between these models and UWS.

The priority heuristic (Brandstätter et al., 2006) is a fast-and-frugal heuristic for binary decisions from description. It sequentially compares the two alternatives on a list of criteria and stops after comparing the options on the first criterion on which they are sufficiently different. In the domain of gains, the priority heuristic first considers the minimum gain. If that does not lead to a decision, then it considers the probability of the minimum gain, and the remaining criteria are the maximum gain and its probability. UWS is similar to the priority heuristic in that it prioritizes important information. On the other hand, the two theories are very different in that UWS uses the probabilities to simulate outcomes whereas the priority heuristic treats the probabilities as just another attribute. Furthermore, the prioritization of UWS is stochastic allowing it to predict choice probabilities. While both the priority heuristic and UWS qualitatively capture the violations of expected utility theory in decisions from description, we found that UWS outperformed the priority heuristic on the Technion choice prediction tournament. Unlike the priority heuristic, UWS was derived from first principles and is more widely applicable.

The heuristic we derived by applying utility-weighted sampling to binary choices from description is similar to decision-by-sampling (N. Stewart et al., 2006) in that both mechanisms rely on drawing samples, comparing them, and counting how often the comparison favored the option to be evaluated. Although the decision-by-sampling model was originally proposed as a model of magnitude judgments, it has since been extended to predict choice probabilities (Noguchi & Stewart, in press; L. Stewart, Overath, Warren, Foxton, & Griffiths, 2008; N. Stewart, 2009; N. Stewart, Reimers, & Harris, 2015). Decision by sampling is consistent with the over-weighting of extreme events predicted by UWS because it assumes that the samples are drawn from memory and thus are subject to the availability bias in memory retrieval (N. Stewart et al., 2006; Tversky & Kahneman, 1973), but unlike UWS it does not specify why extreme events are more available than mundane events and how strong their availability should be.

Like UWS, the exemplar confusion model of decisions from description (D. Lin et al., 2015) assumes that people mentally simulate the outcomes of choosing either option. But unlike UWS, it

simulates outcomes representatively according to their true stated probabilities and for each simulated outcome there is a small chance that its value will be confused. When a confusion occurs a value is chosen uniformly at random from the set of possible values and the sampled value replaces the value of the simulated outcome. This model captures that small probabilities tend to be overweighted whereas large probabilities tend to be underweighted in decisions from description. However, unlike UWS, the exemplar confusion model does not capture that overweighting depends on extremity.

UWS is similar to query theory (Johnson et al., 2007; Weber et al., 2007) in that both assume that preferences are constructed by the sequential consideration of a limited number of aspects or possible outcomes. Both accounts agree that cognitive constraints lead decision-makers to give more weight to the desiderata that are processed first. The main advance of UWS is to provide a rational process model of the order and frequency with which potential outcomes are queried and how the considered outcomes are translated into a decision. The similarity between UWS and query theory suggests that the process tracing methods that provided support for query theory could also be used to test UWS.

The strengths of UWS are complementary to early drift-diffusion models of value-based choice (Krajbich et al., 2010; Krajbich & Rangel, 2011). Both models sequentially accumulate evidence. But while the focus of UWS is on which outcomes should be generated to generate the most informative evidence, the focus of drift-diffusion models is on when the process of evidence generation should be terminated. Furthermore, while most applications of the drift-diffusion model focus on evidence that is generated by the environment, UWS focuses on evidence that is internally generated by memory recall or mental simulation. Recent work has applied to the drift-diffusion model to decisions from memory (Shadlen & Shohamy, 2016) to capture the relationship between response times and choice frequency. Our model is complementary in that it offers a quantitative account of which potential outcomes are sampled from memory. Combining UWS with the drift-diffusion model is one of the directions for future research we will discuss below.

Utility-weighted sampling is similar to selective integration (Tsetsos et al., 2016) in that both provide a rational explanation for violations of expected utility theory. However, the mechanism of utility-weighted sampling is different: In binary choice, utility-weighted sampling overweights attributes on which the alternatives differ by a large amount relative to attributes on which their values are similar. By contrast, selective integration always underweights the weaker attribute value by the same factor regardless of how much larger the stronger attribute value is. Furthermore, while the normative explanation of selective integration emphasizes noise in the decision stage, the normative

justification of utility-weighted sampling is that most real-life decisions have to be made from a small subset of the available information because time is valuable. Our article complements the normative explanation of intransitivity by Tsetsos et al. (2016) by explaining a different set of cognitive biases that might result from a different mechanism.

Utility-weighted sampling is also related to the associative accumulation model by Bhatia (2013) according to which an attribute of a choice alternative will be sampled more frequently if its value is high. This is similar to our model except that in our model the sampling frequency would increases with the extremity of the attribute value's utility instead of its value per se. Utility-weighted sampling provides a strong rational explanation for the importance of extremity whereas the alternative assumption of the associative accumulation model appears to be less principled.

A critical feature of UWS is that it overweights the probability of extreme events. While prospect theory (Kahneman & Tversky, 1979) assumed that the overweighting of outcomes depends only on their probability, utility-weighted sampling predicted that overweighting is driven by the outcome's utility, and the results reported here support this assumption very strongly. Rank-dependent expected utility theories (Quiggin, 1982) like cumulative prospect theory (Tversky & Kahneman, 1992) accommodate the effect of utility on probability weighting by applying the weighting function to the cumulative outcome distribution ($P(O \leq o)$). This captures that the probabilities of the worst and the best outcomes tend to be overweighted. Utility-weighted sampling adds to cumulative prospect theory by identifying cognitive mechanisms that might give rise to this effect. Furthermore, while Kahneman and Tversky (1979) assumed that the overweighting of outcome probabilities in decision-making was independent of the overestimation of event frequencies they attributed to the availability heuristic, we have argued that both originate from the same utility-weighted sampling mechanism.

In addition, utility weighted sampling predicts the distribution of people's choices whereas cumulative prospect theory was created to predict their modal response. This difference allowed utility-weighted sampling to capture people's choice frequencies in the Technion choice prediction competition more accurately than cumulative prospect theory. Our model performed on par with a stochastic extension of cumulative prospect theory that predicts choice distributions (Erev et al., 2010) and other probabilistic extensions of cumulative prospect theory (Rieskamp, 2008; Stott, 2006) might perform similarly. Furthermore, in cumulative prospect theory overweighting only depends on the rank of the outcome's utility. Thus, if the largest outcome is very close to all other outcomes, then it should be overweighted just as much as when it is orders of magnitudes larger than all other outcomes. By contrast, according to utility-weighted sampling, the largest outcome

should be overweighted more heavily in the latter case than in the former.

Previous descriptive theories of choice, including disappointment theory (Bell, 1985; Loomes & Sugden, 1984, 1986), regret theory (Loomes & Sugden, 1982), and salience theory (Bordalo et al., 2012) also assert that people overweight extreme events. Our resource-rational analysis provides a rational justification for this assumption. Despite this commonality, UWS is qualitatively different from all of these previous theories. While all three previous theories are descriptive theories that predict what people will choose, utility-weighted sampling and utility-weighted learning are process models that specify the mechanism of how people decide and how this mechanism changes with learning. Unlike any of the previous theories, these mechanisms predict that people's memory recall and frequency estimates should be biased to overrepresent extreme events and both predictions were confirmed in the experiments reported above. We now discuss the similarities and differences between UWS and each of these three theories in turn.

Our UWS models of frequency estimation and decisions from experience bear a surprising similarity to disappointment theory (Bell, 1985; Loomes & Sugden, 1984, 1986) in that the optimal sampling distribution (Equation 3.10 in Chapter 3) is proportional to the absolute value of the disappointment or elation that the decision maker would experience about the outcome, and according to the UWL model, the absolute value of the disappointment or elation that the decision-maker experiences determines how much the association between an action and its outcome is strengthened. Likewise, our model of binary choice from description is similar to regret theory and salience theory in that it amplifies the impact of large utility differences. Like regret theory and salience theory, this model assumes that decision-makers reason about the difference between the outcomes of the two actions instead of evaluating each action separately. Due to this commonality, our model of binary decisions from description shares some of the strengths and weaknesses of regret theory. On the positive side, this assumption allows all three theories to explain the Allais paradox, the fourfold pattern of risk preferences, and preference reversals. Furthermore, this shared property also predicts violations of weak stochastic transitivity (Tversky, 1969) for some triplets of gambles. For instance, with the parameters estimated from the Technion choice prediction tournament our UWS model of binary choice prefers a 50% chance of \$38 to a 35% chance of \$58 ($p_{2 \succ 1} = 51.1\%$), prefers a 70% chance of \$30 over the 50% chance of \$38 ($p_{3 \succ 2} = 51.7\%$), and yet prefers the 35% chance of \$58 over the 70% chance of \$30 ($p_{1 \succ 3} = 52.1\%$). On the negative side, this commonality entails that, unlike disappointment theory, neither UWS nor regret theory can capture the common-ratio effects in problems that control for regret (Starmer & Sugden, 1989). Nor can UWS capture the specific intransitivity of people's preferences demonstrated by Tversky (1969).

Despite the commonality, the mechanism by which UWS overweights extreme events deviates from regret theory. Specifically, while regret theory and disappointment theory amplify the subjective utility of extreme events, UWS postulates that extremity increases the decision-maker's propensity to consider an outcome and thereby increases its subjective probability without affecting its utility. This entails a non-linear, non-monotonic interaction between probability and extremity: For unlikely outcomes the effect of extremity increases with their probability but for likely outcomes the effect of extremity decreases with their probability because subjective probabilities cannot be larger than 1. Furthermore, in UWS the overweighting of a large difference also depends on the magnitude of the utility differences between other pairs of outcomes. Depending on the magnitude of the differences of those other pairs, UWS can underweight the same pair of outcomes in one context and overweight it in a different context. By contrast, in regret theory, the same difference is always amplified to the same extent and its weight increases linearly with the event's probability. These differences did manifest in our simulations of the common-ratio effects reported by Starmer and Sugden (1989). Concretely, we found that UWS with the parameters estimated from the Technion choice prediction tournament for decisions from description failed to predict the common ratio effects that regret theory did capture (Starmer & Sugden, 1989). Furthermore, disappointment theory and regret theory make the very intuitive prediction that expectations and counterfactual outcomes modulate the satisfaction that people experience when they attain a certain outcome. This too, is not captured by UWS. On the other hand, UWS correctly predicted that extreme events come to mind first and that people overestimate their frequency. Taken together, these findings suggest that extremity affects both subjective utilities and subjective probabilities. UWS thus appears to be complementary to disappointment theory and regret theory because it captures different effects and explains them at a different level of analysis.

Although salience theory and UWS both assume that the subjective probability of extreme events is inflated, our account offers three advances over salience theory. First, we do not only describe the effect of utility on probability-weighting, but we also model the cognitive strategy that generates it. Second, our theory reconciles this seemingly irrational effect with rational information processing. Concretely, the resource-rational basis of the salience of a utility difference $\Delta U = u(O_1) - u(O_2)$ is the relative frequency with which it should be simulated, i.e. the importance distribution $\tilde{q}(\Delta u) \propto p(\Delta u) \cdot |\Delta u|$.[§] This provides a resource-rational justification for salience and a mechanistic account of its effect on decision-making. Third, since our explanation instantiates a more general theoretical framework – resource-rationality – it can also capture many additional phenomena such as decisions

---

[§]This definition satisfies two of Bordalo et al.'s (2012) three axioms of salience.

from experience, memory biases, and biases in frequency estimation.

B.9   Counterintuitive Model Prediction: Inconsistency increases with mental effort

Concretely, the gap between the UWS heuristic's risk seeking in choices between a small chance of winning a lottery versus the lottery's expected value (e.g., a 1% chance of winning $100 vs. $1 for sure) and its risk aversion for choices between a small chance of losing a gamble versus losing its expected value for sure (e.g., a 1% chance of losing $100 vs. losing $1 for sure) widens with the number of simulated outcomes due to the compounding of random errors in the evaluation of the utility of the simulated outcomes (see Figure B.4). This entails that, in this very particular situation, manipulations that reduce mental effort, such as time pressure, should make people appear more rational in these decisions, whereas manipulations that increase mental effort should make them appear less rational.¶ This counter-intuitive relationship could also be used to test whether people allocate their cognitive resources rationally: While incentives for high performance should increase measures of mental effort (Mulder, 1986) on most tasks, people should always exert the minimal amount of cognitive effort on decisions problems where effort fails to improve performance. Regardless of how much mental effort a person exerts on these tasks they should always be biased at least as much as a person who simulates the outcome only once.

¶This prediction is very specific to the particular decisions described here, the normalized, stochastic utility function, and the estimated noise level but not representative of UWS in general.
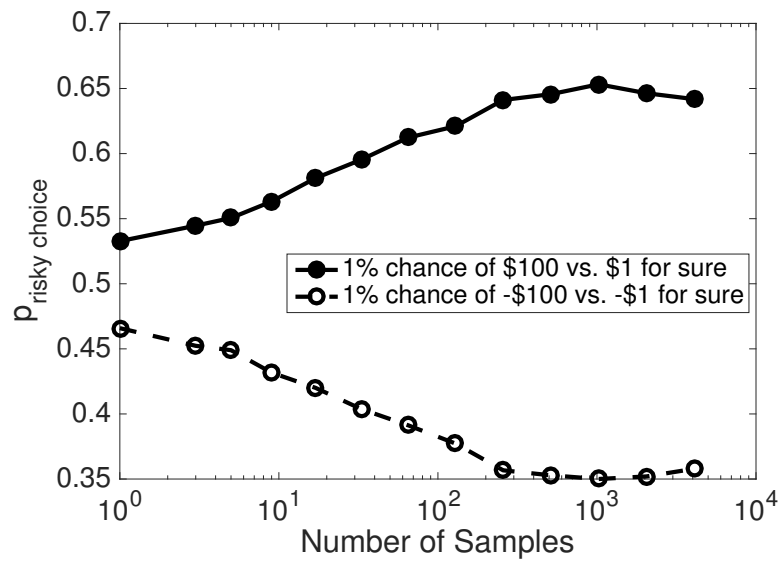
Figure B.4: Counterintuitive prediction of UWS: Investing more mental effort can increase the inconsistency of people's risk preferences in choices between gambles and their expected values. Each line shows the frequency with which the UWS heuristic for binary decisions from description chose the risky option, averaged across 50000 simulations.

C

Strategy Selection

C.1 Technical Details About the Models

C.1.1 SSL and RELACS

According to the SSL model Rieskamp and Otto (2006) the probability that strategy i will be chosen ($P(S_t = i)$) in trial t is proportional to its reward expectancy q_i:

$$P(S_t = i) \propto q_t(i),$$

where $q_t(k)$ is the sum of the rewards obtained when strategy k was chosen prior to trial t plus the initial reward expectancy

$$q_0(k) = r_{max} \cdot w \cdot b_k,$$

where r_{max} is the highest possible reward, w is the strength of the initial reward expectancy, and $b_1, \ldots, b_N \in [0; 1]$ are the agent's initial relative reward expectancies for strategies $1, \ldots, N$ and sum to one.

The RELACS model Erev and Barron (2005) chooses strategies according to their recency-weighted average payoffs

$$
\begin{cases} \alpha \cdot r_t + (1 - \alpha) \cdot w_t(k) & \text{if } S_t = k \\ w_t & \text{else} \end{cases}
$$

$$
P(S_t = k) \propto e^{\lambda \cdot \frac{w_t(k)}{v_t}},
$$

where the parameters α and λ determine the agent's learning rate and decision noise respectively, and V_t is the agent's current estimate of the payoff variability.

C.1.2 Conjugate update equations for the posterior distribution of a Gaussian likelihood and a Gaussian prior

The prior on the reward rate is a normal distribution and the likelihood of the ratio of observed total reward over total time is a standard normal distribution, that is

$$
P(\bar{r}) = \mathcal{N}(1, 1),
$$

$$
P\left(\frac{r_\text{total}}{t_\text{total}} \,\middle|\, \bar{r}\right) = \mathcal{N}\left(\bar{r}, \frac{t_\text{total}}{60\text{sec}}\right).
$$

Consequently, the posterior distribution of the reward rate is

$$
P(\bar{r}|r_\text{total}, t_\text{total}) = \mathcal{N}(\mu_\text{post}, \tau_\text{post}),
$$

with

$$
\tau_\text{post} = \tau_\text{prior} + \tau_\text{likelihood} = 1 + \frac{t_\text{total}}{60\text{sec}}
$$

$$
\mu_\text{post} = \frac{\tau_\text{prior} \cdot \mu_\text{prior} + \tau_\text{likelihood} \cdot \frac{r_\text{total}}{t_\text{total}}}{\tau_\text{prior} + \tau_\text{likelihood}} = \frac{1 + \frac{r_\text{total}}{60\text{sec}}}{1 + \frac{t_\text{total}}{60\text{sec}}}.
$$

C.1.3 Bayesian Regression

For continuous outcomes (i.e., execution time and reward) we performed exact Bayesian inference in a linear regression model Kunz (2009); Lindley and Smith (1972). For binary outcomes (i.e., correct

vs. incorrect) we use Bayesian logistic regression with the Laplace approximation Lieder and Griffiths (2015); Lieder, Hsu, and Griffiths (2014). This approach learns a probability distribution over the amount of time that will pass and the amount of reward that will be obtained.

The Bayesian linear regression Kunz (2009); Lindley and Smith (1972) model for the execution time T was defined as

$$P(T|\mathbf{f}, s, w^{(T)}, \sigma_T^2) = \mathcal{N}(\mu = \sum_i w_{k,s}^{(T)} \cdot f_i, \sigma_T^2),$$

$$P(w_{:,s}^{(T)}) = \mathcal{N}(\mu = 0, \Sigma = \mathrm{Id}),$$

$$P(\sigma_T^2) = \mathrm{InvGamma}(\alpha_0, \beta_0),$$

where $\mathcal{N}$ stands for the normal distribution, InvGamma stands for the inverse gamma distribution, $w_{:,s}^{(T)}$ is the vector of the the weights of all features on the expected execution time of strategy s, and Id stands for the identity matrix. Given observed rewards $\mathbf{r}^{(\mathbf{1},\dots,\mathbf{t})} = (r_1, \dots, r_t)$ in trials $1, \dots, t$ when the strategy was applied to a problem with features $\mathbf{f}^{(1,\dots,t)} = (\mathbf{f}^{(1)}, \dots, \mathbf{f}^{(t)})$, i.e. $P(\alpha^{(s)}|\mathbf{r}, \mathbf{f}^{(1,\dots,t)})$ the model's prior distributions on the regression coefficient and the variance were updated to the respective posterior distributions $P(\alpha^{(T)}|\mathbf{r}^{(1,\dots,t)}, \mathbf{f}^{(1,\dots,t)})$ and $P(\sigma_T^2|\mathbf{r}^{(1,\dots,t)}, \mathbf{f}^{(1,\dots,t)})$. Since the priors are conjugate to the likelihood function, the posterior distributions are in the same family as the prior distributions and their parameters can be computed by the standard update equations for the normal-normal and normal-gamma models Kunz (2009); Lindley and Smith (1972). When the reward was continuous, then the same model was used for learning to predict the reward. But if the reward was binary then we used Bayesian logistic regression with the Laplace approximation (see Section 4.3).

The model's priors on the error variance and the precision of the prior on regression coefficients were set to convey weak domain knowledge. In the sorting simulations, the prior expectation on the variance of the noise was 10 for the execution time in seconds ($\alpha_0 = 1$, $\beta_0 = 10$) and 0.1 for the binary reward ($\alpha_0 = 10$, $\beta_0 = 1$), and the standard deviation of the prior on the regression coefficients was 10 for the execution time and 1 for the binary score. These priors reflect that the execution times in this simulation were one to two orders of magnitude larger than the rewards.

In the simulations of the decision-making experiments by Payne et al. (1988)Payne et al. (1988), the prior expectation of the variance in the execution time was 1 ($\alpha_0 = \beta_0 = 1$), and the variance of the prior on the coefficients predicting the execution time was 1 as well. Since the relative reward was confined to the interval $[-1, 1]$ the prior expectation of its error variance was 0.1 ($\alpha_0 = 1$, $\beta_0 =$

0.1); the precision of the prior on the regression coefficients was 1.

The simulations of the Mouselab experiments assumed a time cost of \$7/h at a rate of 1 computation/sec. The prior on the reward rate corresponded to 1 minute's worth of experience in an environment with a reward rate of \$7/h. The prior distributions on the strategies expected rewards and execution times were a normal distribution with mean zero and precision 0.1. The priors on the error variances of execution time and expected reward were Gamma(1,1).

In the simulations of the Rieskamp experiments the precision of the Gaussian prior on the coefficients of the reward model and the execution time model were estimated according to the maximum likelihood method. The prior on the error variance of the score model was Gamma(1,0.1) and the prior on the error variance of the execution time was Gamma(1,1).

In the simulations of mental arithmetic, the variance of the prior on the regression coefficients was 1 for both the execution time model and the model of accuracy, because the score was binary and single-digit addition takes only a few seconds. The prior on the error variance of the execution time was Gamma(1,1) because the execution time variability of addition strategies is in the order of seconds.

C.1.4 Laplace Approximation to Bayesian Logistic Regression

When the reward is binary (e.g., correct versus incorrect) rather than continuous, then linear regression would be ill-suited to predict it. Hence, in this case our model uses Bayesian logistic regression to predict that probability that the response will be correct ($R = 1$). According the Bayesian logistic regression model, the probability that a strategy s will generate a reward is given by

$$P(R = 1 | s, \mathbf{f}, \alpha) = \frac{1}{1 + \exp(-\sum_k w_{k,s}^{(R)} \cdot f_k)},$$

$$P(\alpha^{(s)}) = \mathcal{N}(^- = \mathbf{0}, \text{œ} = 0.01 \cdot \mathbf{I})$$

The posterior distribution on the regression coefficients $w_{:,s}^{(R)}$ for the expected reward of strategy s given observed rewards $\mathbf{r}^{(1,\dots,t)} = (r_1, \dots, r_t)$ in trials 1, ... ,t when the strategy was applied to a problem with features $\mathbf{f}^{(1,\dots,t)} = (\mathbf{f}^{(1)}, \dots, \mathbf{f}^{(t)})$, i.e. $P(w_{:,s}^{(R)} | r, f^{(1,\dots,t)})$, does no longer have a simply analytic solution. Therefore, we approximate by a normal distribution whose mean is the mode of the posterior distribution and whose precision matrix is the negative Hessian (which is the

matrix of second partial derivatives) of the log-posterior at its mode:

$$P(w_{:,s}^{(R)}|\mathbf{r}^{(1,\dots,t)}, \mathbf{f}^{(1,\dots,t)}) \approx Q(w_{:,s}^{(R)}|\mathbf{r}^{(1,\dots,t)}, \mathbf{f}^{(1,\dots,t)})$$

$$= \mathcal{N}(\mu = w_{\max}, \Sigma^{-1} = -H(\alpha_{\max})),$$

$$w_{\max} = \arg\max_{\alpha} p(w_{k,s}^{(R)} = w|\mathbf{r}, \mathbf{f})$$

$$H_{i,j} = \frac{\partial^2 \log p(w_{:,s}^{(R)}|\mathbf{r}, \mathbf{f})}{\partial \alpha_i^{(s)} \partial \alpha_j^{(s)}}.$$

This is known as the Laplace approximation. It can be derived as a second-order Taylor series expansion of the log-posterior. The posterior mode was determined by numerical optimization using the function *fminunc* from the Matlab 2014b optimization toolbox and the gradients and Hessian were computed analytically.

C.1.5   Feature Selection by Bayesian Model Selection

To model how people discover which features are relevant for predicting a strategy's execution time or reward, our model includes a feature selection mechanism. According to our model, features are selected by Bayesian model selection Kass and Raftery (1995). Concretely, we consider one model for each possible subset of the features and determine the model with the highest posterior probability given the observations. To efficiently compute Bayes factors, we exploit that all models are nested within the full model that includes all of the features by computing Savage-Dickey ratios Penny and Ridgway (2013).

C.2   Quantitative comparison of human performance in Experiment 2 against model prediction

In the pretest, people achieved a reward rate of 62 ± 10.9 points/sec and by the posttest block their reward rate had increased to 156.8 ± 22.9 points/sec. By contrast, the model's reward rate increased from 12.6 points/sec to 142 points/sec. Perhaps the main reason for these differences is that even though people did not receive any feedback in the pretest block, they could already avoid effortful deliberation in favor of skipping decision problems with low expected value about $60.4 \pm 5.7\%$ of the time. By contrast, our model started out deliberating on 62.5% of the problems presented in

the pretest block. This suggests that people adapt their strategies not only based on the experienced outcomes of their choices but also based on predicted outcomes. However, once feedback was presented in the training block our model quickly learned to be as fast and frugal as our participants and achieved an even higher level of adaptive frugality in the posttest. Furthermore, while the model predicted a decrease from about 7.7 to about 0 acquisitions, people's average number of information acquisitions decreased from 3.8 ± 0.8 to 1.6 ± 0.5 acquisitions per trial. When people engaged in effortful decision-making they acquired about 8.3 ± 0.57 pieces of information in the pretest and only 6.4 ± 0.48 pieces of information in the posttest, whereas the predicted number of acquisitions dropped from 13.0 to 5.4. Hence, the changes in people's strategy choices tended to be more moderate than predicted.

C.3 Additional Model Comparisons and related analyses

This section reports the additional model comparisons mentioned in the main text in more detail.

Comparing the lesioned metareasoning models against the full metareasoning model on people's strategy choices in the sorting task of Experiment 1 suggested that feature-based learning and choosing strategies based on their predicted VOC were necessary to capture human performance; see Supplementary Figure C.1. When the features were removed from the rational metareasoning model, adaptive strategy selection was abolished: the frequency of adaptive strategy choice dropped significantly to $4.5\%(t(398) = -18.63, p < 10^{-15})$. Likewise, the models that chose strategies by a criterion other than the VOC were unable to choose strategies adaptively: When the reward function ignored the cost of time, then the frequency of adaptive strategy selection dropped to $23.5 \pm 3.0\%(t(398) = -10.67, p < 10^{-15})$. When the reward function was the reward rate, then the performance dropped to $39.0 \pm 3.5\%(t(398) = -6.67, p < 10^{-10})$. Learning a model-free approximation to the VOC by estimating reward minus cost directly achieved adaptive strategy choice patterns in $68.0 \pm 3.3\%$ of the simulations; this frequency was significantly above chance $(p < 10^{-7})$ and not significantly lower than the frequency achieved by the full rational metareasoning model that learns separate models for predicting execution time and reward $(t(398) = -0.54, p = 0.2941)$. These results suggest that both strategy selection based on the predicted VOC and feature-based learning are important to capture the adaptiveness of people's strategy choices. By contrast, when the exploration component of the model was removed, the frequency of adaptive strategy selection remained at $63.0 \pm 3.4\%$ which is significantly above chance $(p < 10^{-7})$ and not significantly different from the performance of the RM model with explo-

ration ($t(398) = -1.60, p = 0.0555$). This was expected because in the present experiment there was nothing to be gained from exploration since participants had no choice over their strategy on the training trials and did not receive any feedback about the performance of the strategies that they selected in the test trials. Finally, we compared all model's frequencies of adaptive strategy choice against chance level with the Bonferroni correction for multiple comparisons. We found that only the full rational metareasoning model ($p < 10^{-14}$), the lesioned metareasoning model without exploration ($p < 10^{-7}$), and the lesioned metareasoning model that learned to predict reward minus computational cost ($p < 10^{-11}$) chose strategies adaptively significantly more often than what could be expected by chance.

Our simulations of the first experiment by Payne et al. (1988) showed that only the full rational metareasoning model, the lesioned metareasoning model ignoring the cost of time, and the lesioned metareasoning model that approximated the VOC through model-free metacognitive RL were able to capture that people choose fast-and-frugal heuristics more frequently when some outcomes are much more probable than others (Supplementary Figure C.2, panel A) and when their decision time is limited (Supplementary Figure C.2, panel B). The feature-based representation and exploration were necessary to capture people's adaptive strategy choices in this experiment: When the feature-based representation was removed the model no longer preferred fast, attribute-based heuristics on problems with high time pressure versus low time pressure ($t(1998) = 0, p = 0.50$), and it also would not choose them more frequently on non-compensatory problems than on compensatory problems ($t(1998) = 0, p = 0.50$). When the exploration mechanism was removed, the model never tried any of the fast, attribute-based heuristics on any type of decision problem. This illustrates the importance of exploration for strategy selection learning when the number of strategies is large. By contrast, two of the models that did not distinguish between reward and computational cost retained the adaptive strategy choice pattern of the original rational metareasoning model. The model that ignored time costs learned to choose fast, attribute-based heuristics 40.20% more frequently under time pressure ($t(1998) = 14.11, p < 10^{-15}$) and 36.98% more frequently when the dispersion of the outcome probabilities was high ($t(1998) = 12.98, p < 10^{-15}$); the model whose reward function was the difference between payoff and opportunity cost chose fast, attribute-based processing 32.94% more frequently under time pressure ($t(1998) = 10.94, p < 10^{-15}$) and 43.35% more frequently under high dispersion ($t(1998) = 14.39, p < 10^{-15}$). By contrast, the model that learned to predict the reward-rate directly failed to adapt strategy choices to time pressure ($t(1998) = 0, p = 0.50$) and dispersion ($t(1998) = 0, p = 0.50$). The reason may be that the reward rate was determined primarily by the execution time (because the relative reward is con-

strained to lie between 0 and 1) whereas adaptive strategy selection in this task required optimizing accuracy.

Comparing the lesioned metareasoning models against the full rational metareasoning model on Experiment 2 and Experiment 3 suggested that all components of the rational metareasoning model are necessary to capture people's capacity to adapt how much they think to the reward structure of their environment. Supplementary Figure C.3 shows each model's predictions for the reward rate, the average number of acquisitions, and the frequency of engagement in Experiment 2. Supplementary Figure C.4 summarizes the equivalent predictions for Experiment 3. Only the full rational metareasoning model and the lesioned model without features captured the increase in people's reward rate and the decrease in the number of acquisitions and the frequency of engagement in Experiment 2. Similarly, only the full rational metareasoning model and the model-free metareasoning model whose reward function was r=reward-cost were able to capture the increase in people's reward rate, number of acquisitions, and frequency of engagement in Experiment 3. Hence, only the full rational metareasoning model can capture human performance in both experiments jointly.

Supplementary Figure C.1 summarizes the BIC values of the model comparison for the first experiment from Rieskamp and Otto (2006). The model comparison provided strong evidence for SSL and SCADS over the rational metareasoning model and the lesioned metareasoning models. The BIC of the rational metareasoning model was larger than the BICs of the lesioned metareasoning models, and the comparisons between the rational metareasoning model and the lesioned metareasoning model provided strong evidence for exploration and model-based strategy selection.

Supplementary Figure C.5 shows the performance of each model in the mixed multi-attribute decision environment with 50% compensatory problems and 50% non-compensatory problems. The rational metareasoning model predicted that the average performance across the 168 trials would be 61.1% for the rational metareasoning model, 57.9% for the lesioned metareasoning model without exploration, and 62.9% for the lesioned metareasoning model that ignoring time cost. By contrast, the lesioned metareasoning model without features, and the lesioned metareasoning model that performed model-free reinforcement learning from the reward rate, as well as the SCADS models, and the SSL and RELACS models performed at or below chance. This highlights that feature-based strategy selection learning is critical to capture people's adaptive flexibility in heterogeneous environments.

As shown in Supplementary Figure C.3, the learning effects in problem solving reported by Gunzelmann and Anderson provided strong evidence for the full rational metareasoning model

(BIC = 213.42) over the lesioned metareasoning model without exploration (BIC ≥ 221.1). Furthermore, there was very strong evidence for the full rational metareasoning model over the lesioned metareasoning models without features or model-based learning (BIC ≤ 221.1) and all SCADS models (BIC ≥ 293.2). This suggests that feature-based learning, exploration, and model-based strategy selection are all necessary to capture people's capacity to learn how to solve problems.

When simulating children's strategic development in the domain of mental arithmetic, we found that exploration and feature-based learning are critical to capture the overlapping waves of different strategies documented by Svenson et al. (1983)Svenson and Sjöberg (1983): Without exploration the model does not try out the Short-Cut sum strategy, the Min strategy, or the Retrieval strategy often enough to learn that they are superior to the Sum strategy (see Supplementary Figure C.6). Without features, the model transitions directly from the sum strategy to the Retrieval strategy because it cannot learn that the Min strategy is more effective than the Retrieval strategy when experience with the problem is limited and one of the addends is small (see Supplementary Figure C.7). Similarly, the lesioned metareasoning model that ignores the time cost of strategy execution fails to switch to the Min strategy, because it is insensitive to the time saved by the Min strategy (see Supplementary Figure C.8). Interestingly, model-free metacognitive reinforcement learning of the VOC failed to switch to the Retrieval strategy (see Supplementary Figure C.9). Finally, model-free metacognitive reinforcement learning based on the reward rate transitioned abruptly to the Min strategy once it had been discovered and failed to transition to the Retrieval strategy afterwards (see Supplementary Figure C.10). These findings suggest that maintaining separate representations of execution time, opportunity cost, and expected reward enables faster learning and adaptation to changes in the strategies' performance or the reward rate.
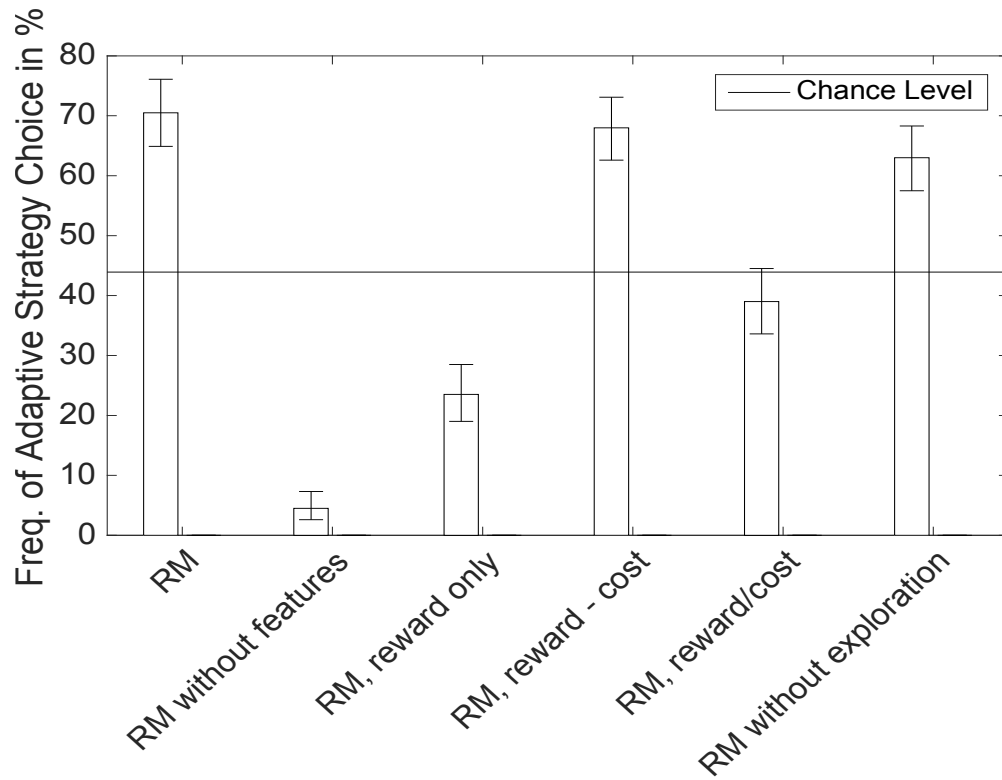
Figure C.1: Evaluation of the full rational metareasoning model of people's choice of sorting strategies in Experiment 1 against sub-models without one of its four components.
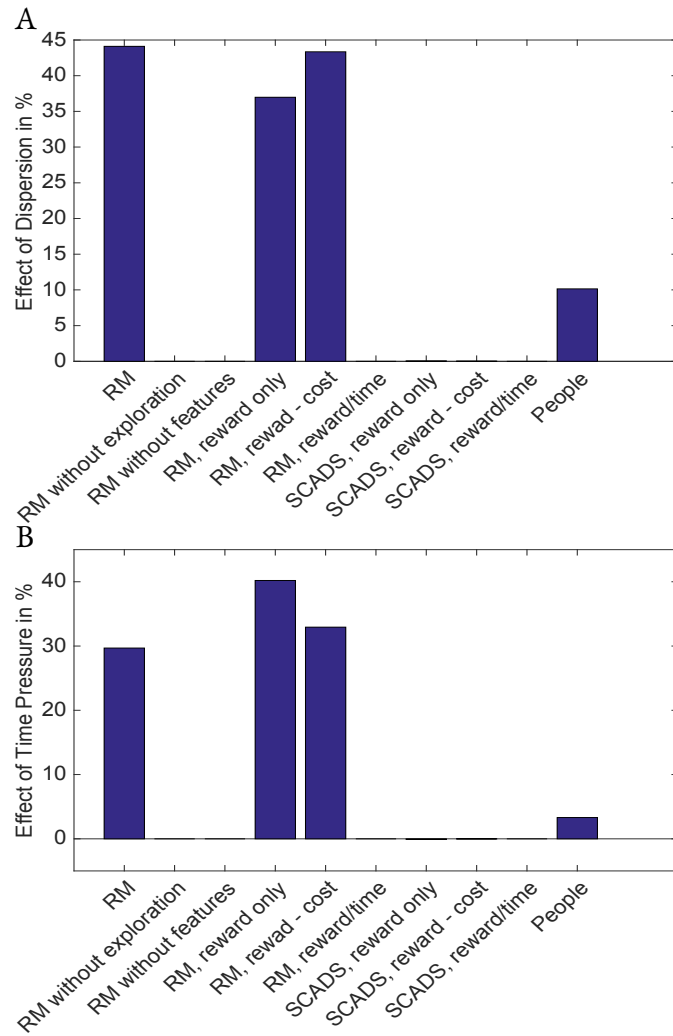
Figure C.2: Simulation of the first experiment from Payne et al. (1988) with lesioned metareasoning models and SCADS models. A: Effect of dispersion on the predicted usage frequency of fast-and-frugal heuristics. B: Effect of time pressure on the predicted usage frequency of fast-and-frugal heuristics.
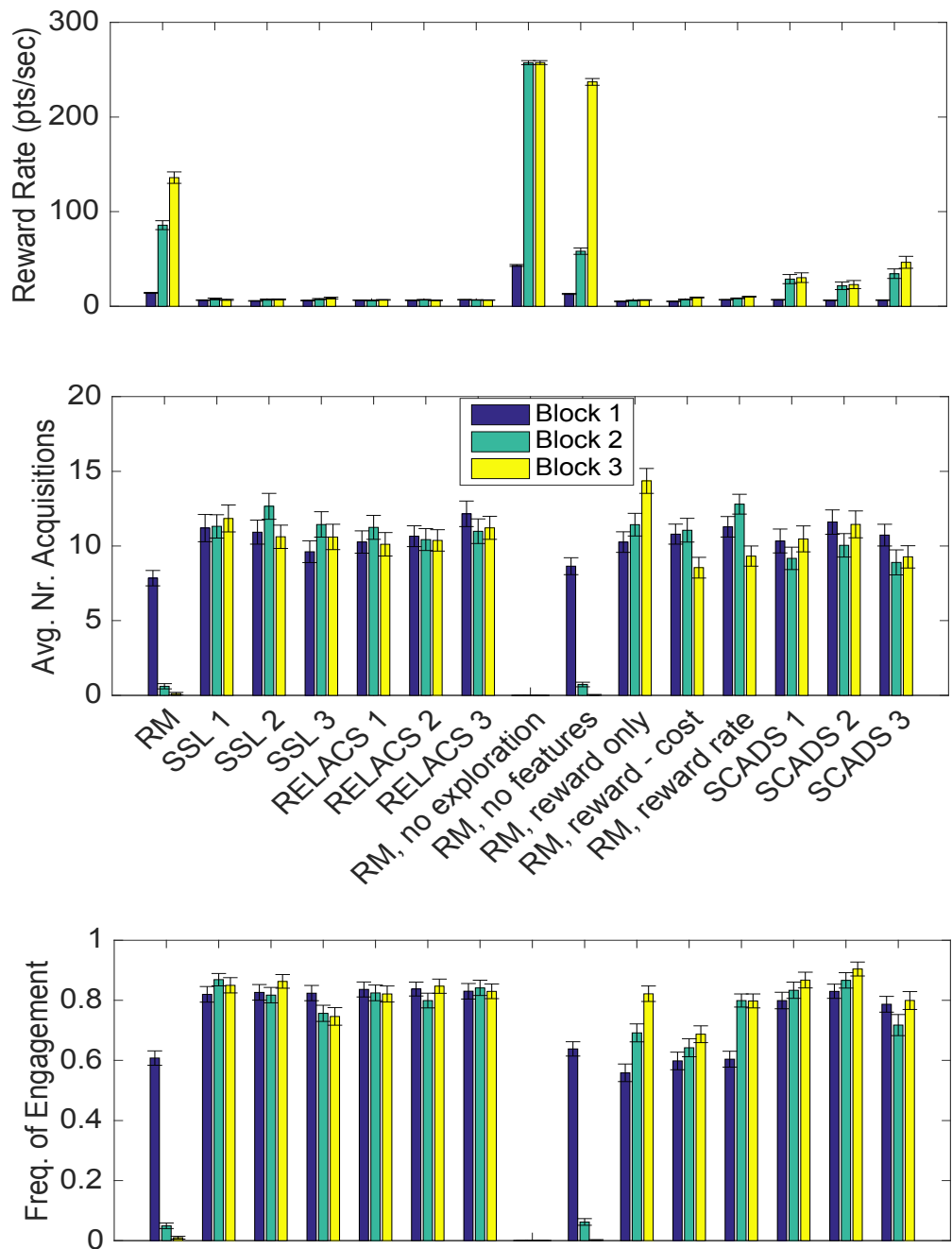
Figure C.3: Model predictions for Experiment 2. Error bars are plus/minus 1 SEM.
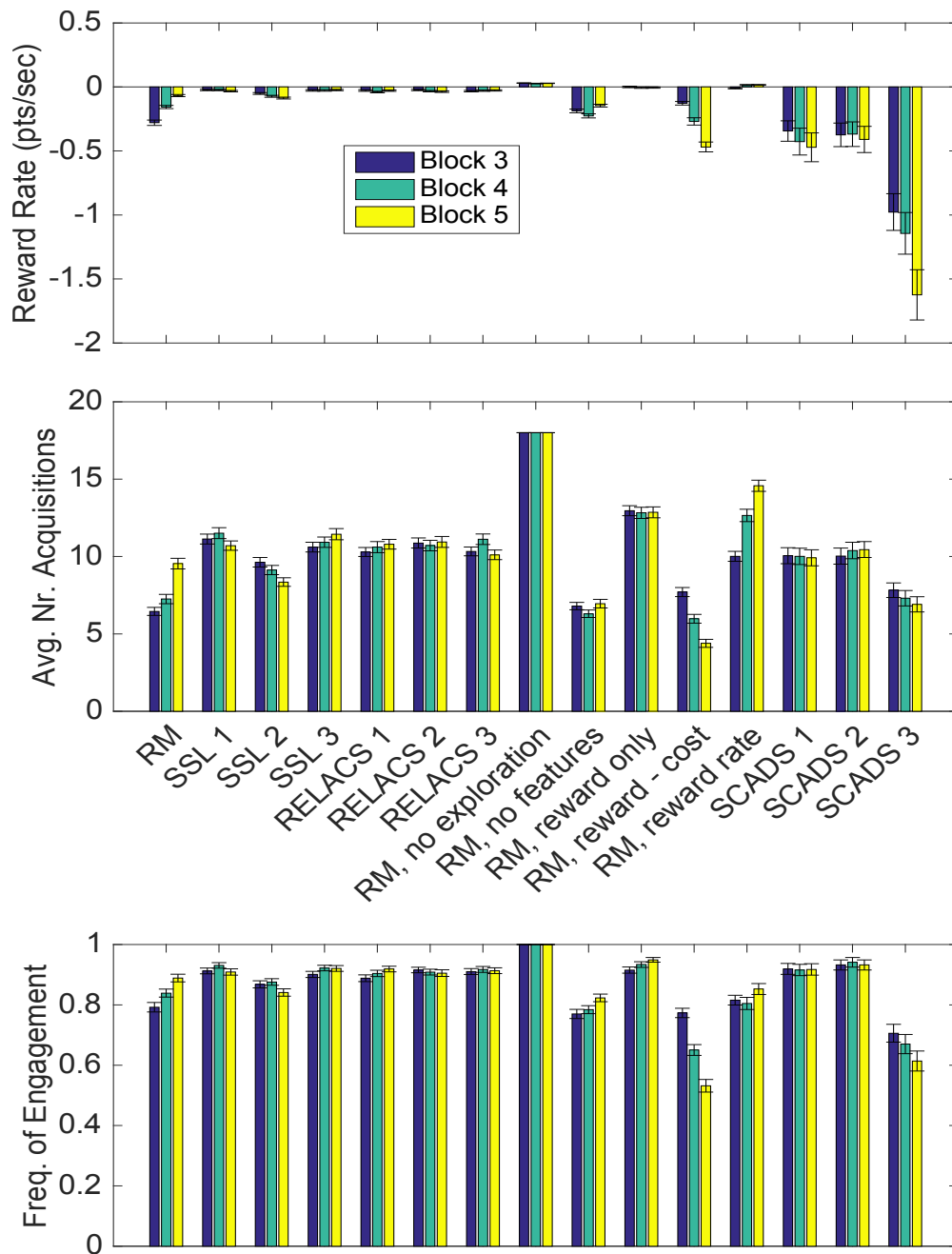
Figure C.4: Predicted reward rates for Experiment 3 Error bars are plus/minus 1 SEM.

Table C.1: Model Comparison on Experiment 1 from Rieskamp and Otto (2006).

Model	BIC	Model	BIC	Model	BIC	Model	BIC
RM	68.9	RM. No cost	94.9	SCADS 1	61.1	SSL	60.6
RM, no exploration	111.7	RM, $r =$ reward $-$ cost	75.3	SCADS 2	62.4		
RM, no features	70.5	RM, $r =$ reward/cost	127.8	SCADS 3	62.0		
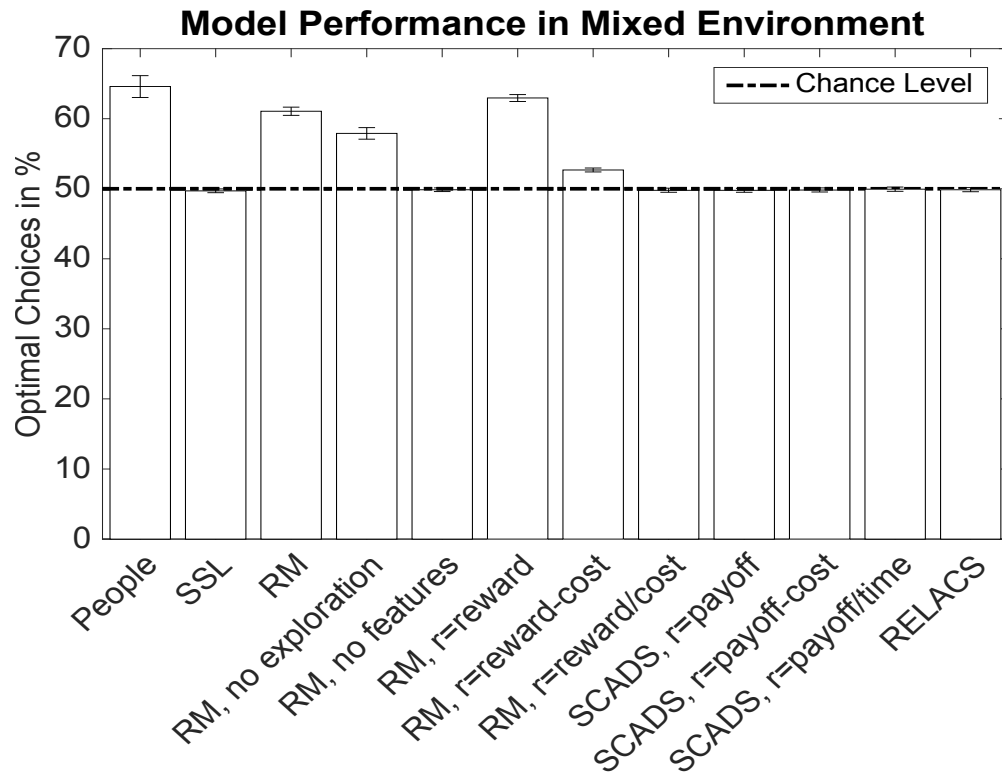
Figure C.5: Performance of model-based versus model-free strategy selection learning mechanisms in a heterogeneous decision environment where 50% of the problems required TTB and 50% required WADD. The error bars denote 95% confidence intervals.
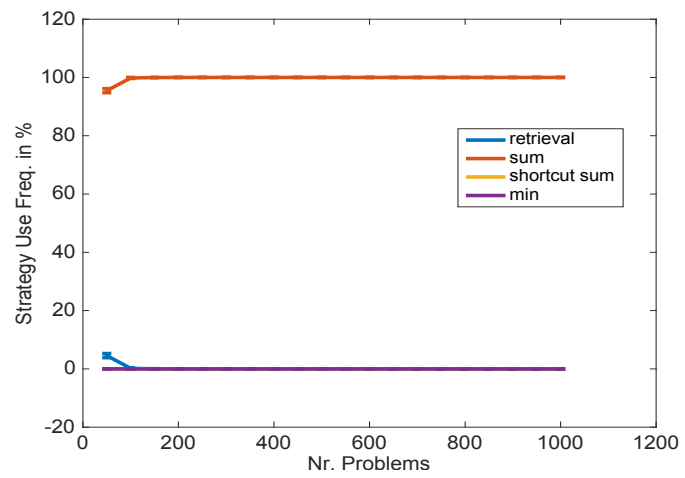
Figure C.6: The lesioned metareasoning model without exploration fails to capture children's strategic development in mental arithmetic.
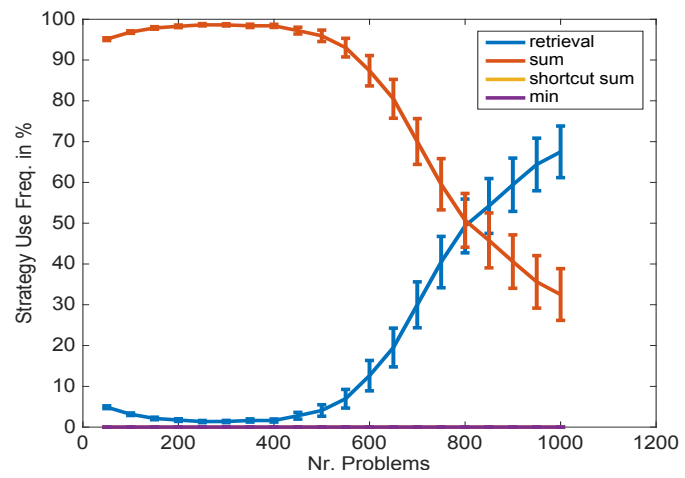
Figure C.7: The lesioned metareasoning model without features fails to capture children's strategic development in mental arithmetic.
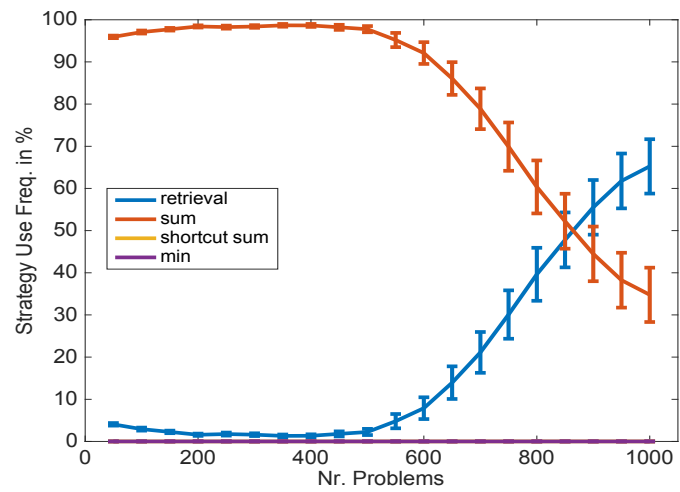
Figure C.8: The lesioned metareasoning model ignoring the cost of time fails to capture children's strategic development.
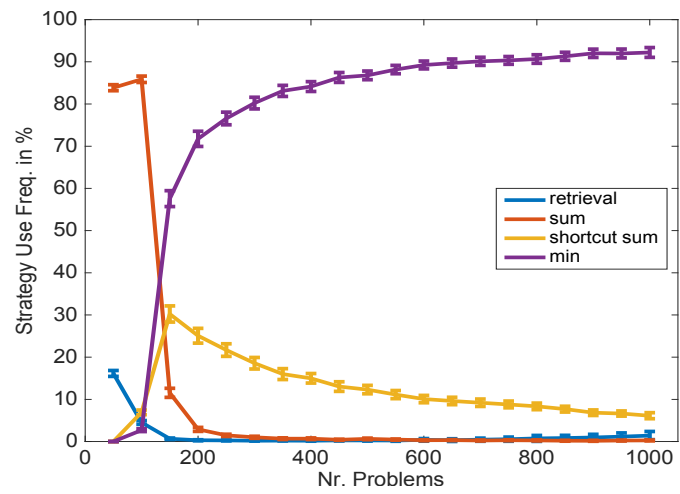
Figure C.9: Model-free metacognitive RL of the VOC ($r = $ rewardcost) predicts that the sum-strategy should fade much earlier than it does in children.
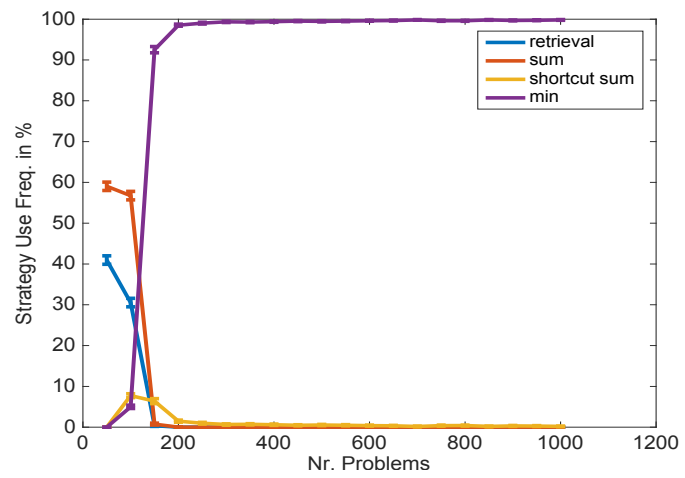
Figure C.10: Model-free metacognitive RL driven by the reward rate converges on the Min strategy much faster than children and fails to transition to the Retrieval strategy.

D

Cognitive prostheses for goal achievement

D.1 Supplementary Methods

D.1.1 Optimal Gamification

Pseudo-rewards can be shifted and scaled without changing the optimal policy, because linear transformations of potential-based pseudo-rewards are also potential-based, that is

$$a \cdot f(s, a, s') + b = \gamma \cdot \Phi'(s') - \Phi'(s), \tag{D.1}$$

$$\text{for } \Phi'(s) = a \cdot \Phi(s) - \frac{b}{1 - \gamma}. \tag{D.2}$$

D.1.2 Experiment 1

The complete experiment can be inspected at http://cocosci.dreamhosters.com/mturk/falk/FlightPlanning/.

MDP MODEL OF THE PLANNING TASK.   The sequential decision-making task of Experiment 1 is isomorphic to a MDP with six states, two actions, deterministic transitions, and a discount factor of
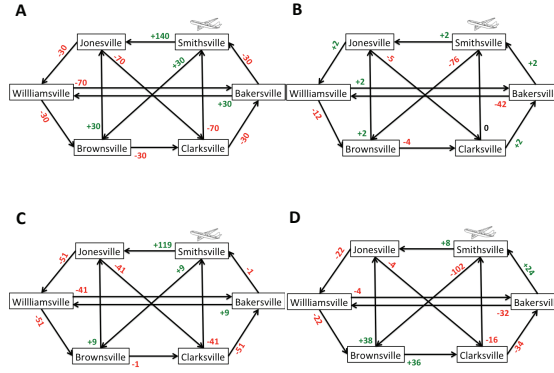
Figure D.1: Conditions of Experiment 1. A: Control condition. B: Embedded pseudo-rewards. C: Separate pseudo-rewards. D: Integrated pseudo-rewards.

$\gamma = 1 - \frac{1}{6}$. The locations correspond to the states of the MDP, the two actions correspond to flying to the first or the second destination available from the current location, the routes correspond to state-transitions, and the points participants received for flying those routes are the rewards. The current state was indicated by the position of the aircraft and was updated according to the flight chosen by the participant.

REWARD TRANSFORMATIONS.   In all three experimental conditions, the pseudo-rewards were mean-centered by subtracting their average to keep the average reward constant; since mean-centering is a linear transformation this retained the guarantees of the shaping theorem (see Eq. D.2). The mean-centered pseudo-rewards were added to the rewards of the control condition (see Figure D.1A) yielding the modified rewards shown in Figure D.1B-D and Table D.1, and the flight map was updated accordingly.

The value function used to compute the approximate, shaping-based pseudo-rewards was

$$\hat{V}_M(s) = \hat{V}_M(s^\star) \cdot \left(1 - \frac{\text{distance}(s, s^\star)}{\max_s \text{distance}(s, s^\star)} \right), \tag{D.3}$$

where the goal state $s^\star$ was *Smithsville*, $\hat{V}_M(s^\star) = 140$ was the highest immediate reward that can be achieved from there, and distance(a, b) is the minimum number of moves required to get from state a to state b.

Experiment 1 was approved by the institutional review board of the University of California,

Berkeley under protocol number 2015-05-7551, study title "Decision Making", and informed consent was obtained from all participants.

D.1.3 Experiment 2

The complete Experiment can be inspected at http://cocosci.berkeley.edu/mturk/falk/PNASExp2/index.html.

Game mechanics.   The character played by the participant could rise from *Trainee* to *ATP senior captain* via 15 intermediate levels. The number of points required to reach the next level increased according to the difficult curve proposed by Bostan and Öğüt (2009). Whenever the player reach the next level a congratulatory message was shown. In addition, participants were told how many stars and dollars were required to reach the next level in the game. To make the levels salient the pilot's shoulder badge was shown in the top right corner of the screen, and a feedback message was shown whenever the character was promoted and earned a badge or was demoted and lost a badge. The player started the game with $+\$50$ so that their balance would remain positive as they learned to play the game.

Experiment 2 was approved by the institutional review board of the University of California, Berkeley under protocol number 2015-05-7551, study title "Decision Making", and all participants gave informed consent.

D.1.4 Experiment 3

Pilot study and task selection.   To select a suitable set of tasks for Experiment 3 we ran a pilot study that acquired subjective ratings of 21 candidate tasks. 100 participants recruited on

Condition	Smiths-		Jones-		Williams-		Browns-		Clarks-		Bakers-	
No PR	140	30	−30	−70	−30	−70	−30	30	−30	−70	−30	−70
Optimal PR	2	−76	2	−5	−12	2	−4	2	2	0	2	−42
Approx. PR	8	−102	−22	−4	−22	−4	36	38	−34	−16	24	−32
Non-Potential-Based PR	119	9	−51	−41	−51	−41	−1	9	−51	41	−1	9

Table D.1: Rewards in Experiment 1. The first entry of each cell is the (modified) reward of the counter-clockwise move and the second one is the (modified) reward of the other move.
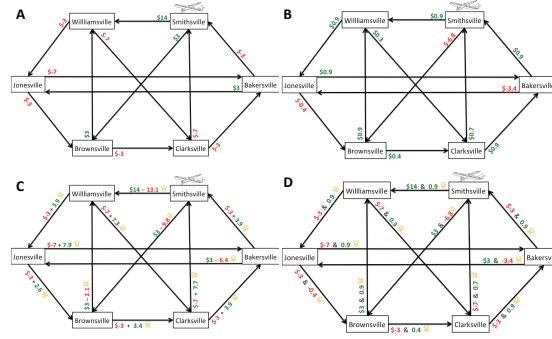
Figure D.2: Conditions of Experiment 2. A: Control condition. B: Embedded pseudo-rewards. C: Separate pseudo-rewards. D: Integrated pseudo-rewards.

Amazon Mechanical Turk evaluated 5 tasks each and were paid \$0.50 in return. For each task, they estimated the fair price that should be paid for the task on Amazon Mechanical Turk and its duration. In addition, they rated the task's difficulty, their willingness to complete it for the price they had indicated, its enjoyableness compared to a typical MTurk HIT, its relative unpleasantness compared to a typical HIT an MTurk, and how likely they would be to postpone it on nine point Likert scales with appropriate anchors. We selected the 4 tasks that participants said they would be most likely to postpone and the task they said they were least likely to postpone. This procedure led to the selection of the following five writing assignments shown in Table D.2. Each assignment required that participants write at least 100 words (assignments 1-4) or at least 50 words (assignment 5).

RECRUITMENT.   The sign-up form was posted on Monday, April 24 2017 and the deadline was at midnight on Wednesday of the following week (i.e., May 3rd 2017). The sign-up form can be inspected at cocosci.berkeley.edu/mturk/falk/ToDoListStudyPart1WritingTasks/. The to-do list website used in this experiment can be inspected at https://todo-list-study.herokuapp.com/.

OPTIMAL GAMIFICATION.   To compute the optimal incentive for completing each of the tasks we first modelled the experiment as a Markov decision process with one action for each task and an additional action for taking a break. The reward function was set up such that each task-action incurred a cost that reflected the task's fair wage as determined in the pilot study described above. Finishing the experiment earns an additional reward of \$20. In the MDP model of the experiment, taking a break earns a reward equivalent to \$0.50 but also comes with a 2.5% chance of forgetting about the tasks. The benefit of finishing the experiment sooner rather than later was captured by a

Writing Assignment	Fair Price	Duration	propensity to postpone	Minimum Length
How has North Korea's economic policy changed since the 1950s? What are the reasons and implications of these changes?	$3	15min	6.6/9	100 words
Please analyze the causes and implications of the British exit referendum in June 2016.	$3.25	25min	6.3/9	100 words
Describe with examples the importance of recognizing and responding to concerns about children and young people's development.	$2.25	20min	6.2/9	100 words
Write an essay about how society should assign value to human life.	$3	27.5min	6.1/9	100 words
What is your favorite TV show and why?	$1	7min	2.8/9	50 words

Table D.2: Writing assignments and their ratings

discount factor of $\gamma = 0.95$.

METHODOLOGICAL DETAILS.   To tempt participants to procrastinate, the to-do list website displayed a series of distracting links to Youtube videos, Reddit articles, news stories, or the game of Tetris.

In addition to measuring participants' motivation and subjective reward of completing a task, the exit survey also recorded age and self-identified gender and inquired if the participant had used any strategies to stay engaged, and which components of the website they found helpful.

We posted a separate reimbursement HIT for participants who decided to quit the experiment was posted on the first day of the experiment, and it also included an exit survey.

Experiment 3 was approved by the institutional review board of the University of California, Berkeley under protocol number 2016-02-8359, study title "To-Do-List Gamification", and all participants gave informed consent.
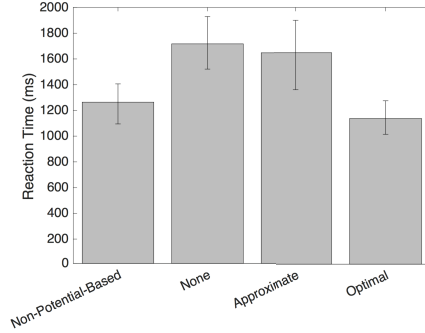
Figure D.3: A: Median reaction times in Experiment 1 with 95% confidence intervals.

D.2 SUPPLEMENTARY RESULTS

EXPERIMENT 1

REACTION TIMES.   A Kruskal-Wallis ANOVA revealed that the type of pseudo-rewards added to the reward function significantly affected people's reaction times ($H(3) = 29.96, p < 10^{-5}$). Given that the pseudo-reward type had a significant effect, we performed pairwise Wilcoxon rank sum tests to compare the medians of the four conditions (see Figure D.3). Optimal pseudo-rewards decreased the median response time from 1.72 to 1.14 sec. per decision ($Z = -4.19, p < 0.0001$), and non-potential-based pseudo-rewards decreased it to 1.12 sec. per decision ($Z = -3.38, p = 0.0007$). People in the condition with approximate potential-based pseudo-rewards took about the same amount of time as people in the control condition (1.65 sec.; $Z = -0.28, p = 0.78$).

EFFECT OF PSEUDO-REWARDS ON CHOICE FREQUENCIES.   The optimal strategy for this experiment was to take the counter-clockwise moves around the circle in all states except *Williamsville* and *Brownsville* (see Figure D.1A). Importantly, at *Williamsville* the optimal policy incurs a large immediate loss, and no other policy achieves a positive reward rate. The optimal pseudo-rewards significantly changed the choice frequencies in each of the six states and successfully nudged participants to follow the optimal cycle *Smithsville* → *Jonesville* → *Williamsville* → *Bakersville* → *Smithsville* (see Figure D.1A). Their strongest effect was to eliminate the problem that most people would avoid the large loss associated with the correct move from *Williamsville* to *Bakersville* ($\chi^2(2) = 1393.8, p < 10^{-15}$). The optimal pseudo-rewards also increased the frequency of all other correct choices along the optimal cycle, that is the decisions to fly from *Bakersville* to
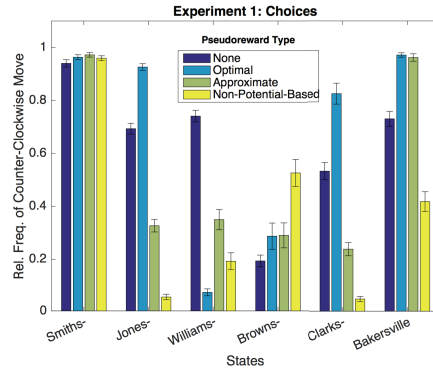
458

Figure D.4: Choice frequencies in each state of Experiment 1 by condition. Error bars enclose 95% confidence intervals.

Smithsville ($\chi^2(2) = 326.5, p < 10^{-15}$), from *Smithsville* to *Jonesville* ($\chi^2(2) = 7.9, p = 0.0191$), and from *Jonesville* to *Williamsville* ($\chi^2(2) = 299.8, p < 10^{-15}$). In addition, the optimal pseudo-rewards increased the frequency of the correct move from *Clarksville* to *Bakersville* ($\chi^2(2) = 92.0, p < 10^{-15}$). The only negative effect of the optimal pseudo-rewards was to slightly increase the frequency of the suboptimal move from *Brownsville* to *Clarksville* ($\chi^2(2) = 13.2, p = 0.0013$). By contrast, the non-potential-based pseudo-rewards misled our participants to follow the unprofitable cycle *Jonesville* → *Clarksville* → *Smithsville* → *Jonesville* by raising the frequency of the reckless moves from *Jonesville* to *Clarksville* ($\chi^2(2) = 1578.6, p < 10^{-15}$) and from *Clarksville* to *Smithsville* ($\chi^2(2) = 813.7, p < 10^{-15}$). The effect of the approximate pseudo-rewards was beneficial in *Smithsville*, *Williamsville*, and *Bakersville*, but negative in *Jonesville*, *Brownsville*, and *Clarksville* (see Figure D.4). This explains why only potential-based pseudo-rewards had a positive net-effect on performance (Figure 1B in the Main Text).

Experiment 2

Effect of presentation format on response times and choice frequencies.

Participants were significantly faster when pseudo-rewards were embedded in the decision environment than when they were presented separately ($Z = -4.06, p < 0.0001$) or in the integrated format ($Z = -2.78, p = 0.0053$). Figure D.5 shows people's choice frequencies for each state depending on the experimental condition. Compared to separately presented pseudo-rewards, embedded pseudo-rewards were significantly more beneficial in all 6 states (all $p \leq 0.0218$) as were integrated
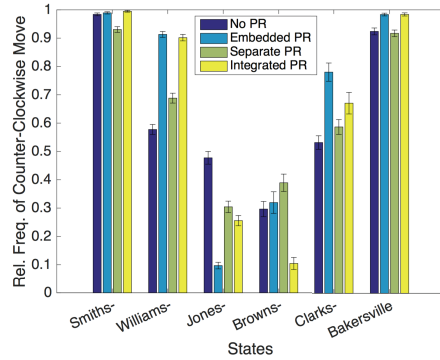
Figure D.5: Choice frequencies in each state of Experiment 2 by condition. Error bars enclose 95% confidence intervals.
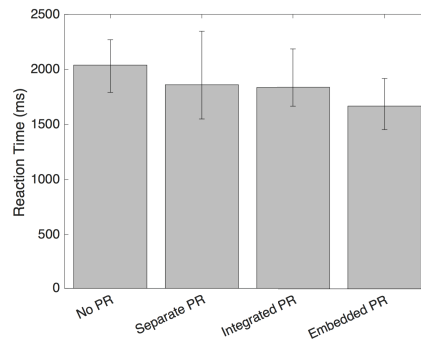

Figure D.6: Median reaction times in Experiment 2 with 95% confidence intervals.

pseudo-rewards (all $p \leq 0.0023$) but separately presented pseudo-rewards were never advantageous to either embedded or integrated pseudo-rewards. Embedded pseudo-rewards were more beneficial than integrated pseudo-rewards in 2 states (all $p \leq 0.0001$); conversely integrated pseudo-rewards were more beneficial than embedded pseudo-rewards in 1 state ($p < 10^{-9}$).

FOLLOW-UP EXPERIMENT.   The integrated pseudo-rewards differ from the separately presented pseudo-rewards in two respects: First, they simplify the decision process by allowing people to base their decision on a single signal. Second, they shift the pseudo-rewards such that the pseudo-reward for the optimal action is always positive. To tease apart the contributions of these two factors, we ran a follow-up experiment in which the separately presented pseudo-rewards were shifted such that the minimum pseudo-reward for an optimal action was the expected return of the optimal policy as it was for the integrated pseudo-rewards.
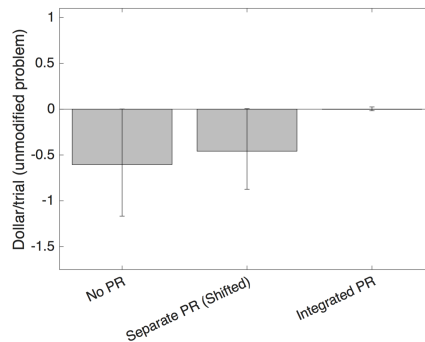
Figure D.7: Performance in Experiment 2b by condition. Error bars enclose 95% confidence intervals.

We recruited 339 participants on Amazon Mechanical Turk. Each participant was randomly assigned them to one of three conditions: no pseudo-rewards, shifted separately presented pseudo-rewards, and integrated pseudo-rewards. Condition 1 and 3 were identical to the equivalent conditions in Experiment 2. In the second condition, the optimal pseudo-rewards were shifted such that the minimum pseudo-reward for taking an optimal action was the expected reward rate of the optimal policy, that is 0.9. In all other regards, this follow-up experiment was identical to Experiment 2.

The median completion time was 23.35 minutes. We excluded 24 participants who had participated in previous flight planning experiments and 15 participants who performed worse or responded faster than 95% of the participants in their condition. Out of the remaining 290 participants 102 were in the condition without pseudo-rewards, 91 were in the condition with shifted separately presented pseudo-rewards, and 97 were in the condition with integrated pseudo-rewards.

We found that the shifted separately presented pseudo-rewards were significantly less effective than the integrated pseudo-rewards ($Z = -2.38, p = 0.0172$) and did not significantly improve people's performance relative to the control condition ($Z = 0.17, p = 0.8617$; median loss: \$11 vs. \$11 in the control condition; see Figures D.7-D.8). By contrast, participants in the condition with integrated pseudo-rewards performed significantly better than participants in the condition without pseudo-rewards ($Z = 2.46, p = 0.0140$; median loss: \$0 vs. \$14.50). Therefore, the primary benefit of the integrated pseudo-rewards appears to be that they simplify the decision process by offloading the computation of adding rewards and pseudo-rewards from the participants.
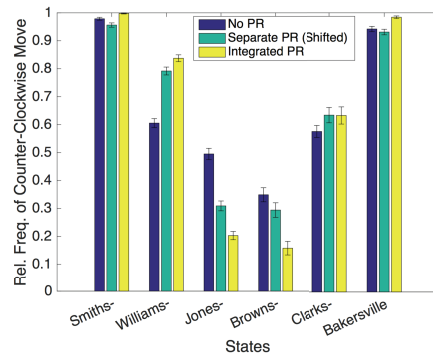
Figure D.8: Choice frequencies in Experiment 2b by condition. Error bars enclose 95% confidence intervals.

Experiment 3

Of the 40 participants who did not complete all tasks (33.9%), only 1 filled out the exit survey. We therefore cannot evaluate the effect of the pseudo-rewards on motivation and perceived reward per se. However, we can analyze its effect on participants who completed all tasks. The following analyses are therefore restricted to this biased subset of participants. Due to this selection bias the results have to be interpreted with caution. For the participants who completed all tasks neither motivation ($\chi^2(1) = 0.04, p = 0.84$) nor experienced reward ($\chi^2(1) = 0.14, p = 0.71$) were significantly affected by optimal gamification. Among these participants, optimal gamification also did not affect how long it took them to complete the tasks ($\chi^2(1) = 0.07, p = 0.79$) or the number of times they aborted a task ($F(76) = 0.27, p = 0.61$). While optimal gamification slightly increased the number of words written per assignment from 155 to 175, this difference was not statistically significant ($F(1, 81) = 1.33, p = 0.25$). Optimal gamification also had no statistically significant effect on the total length of the breaks that these participants took between tasks ($F(1) = 0.42, p = 0.52$) or the number of times that they played Tetris ($F(1, 76) = 0.16, p = 0.69$). Optimal gamification also did not affect how long it took them to submit their first assignment ($\chi^2(1) = 3.26, p = 0.07$), when they started working on it ($\chi^2(1) = 2.35, p = 0.13$), or the delay until the first time they opened an assignment ($\chi^2(1) = 2.18, p = 0.14$).

These negative results suggest that the main effect of optimal gamification was to increase the probability that participants would start working on one of the assignments.

462

THIS THESIS WAS TYPESET using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.