From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Psychology

2022

Cognition as a sequential decision problem

Callaway, Frederick

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01kd17cx080

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Cognition as a sequential decision problem

Frederick Callaway

a dissertation
presented to the faculty
of Princeton University
in candidacy for the degree
of Doctor of Philosophy

recommended for acceptance
by the Department of
Psychology

Advisor: Thomas L. Griffiths

November 2022

© Copyright by Frederick Callaway, 2022. All rights reserved.

Abstract

How should we attempt to understand the mind? Historically, there have been two broad approaches. The *rational* approach focuses on characterizing the problems people have to solve and the optimal solutions to those problems, explaining *why* people behave in the way they do. In contrast, the *mechanistic* approach focuses on identifying the cognitive processes underlying behavior, explaining *how* the mind actually works. Traditionally, these approaches have been viewed as conflicting, but recent years have seen a growing interest in models that synthesize the two approaches.

This dissertation presents a formal framework for deriving models of cognition that are both rational and mechanistic. The key idea to broaden the concept of the "environment" to which cognition adapts: cognitive processes are adapted not only to the external environment (the world), but also to the internal environment (the brain). Formalizing this old idea, I cast cognition as a sequential decision problem in which an agent executes cognitive actions to navigate between mental states and, ultimately, produce effective behavior. In three domains—attention, memory, and planning—I show how the framework can be applied to yield models that explain both how the mind works and why it works that way.

Contents

Abstract **3**

1 Introduction **9**
- 1.1 Sequential decisions in the world and the mind . . . 10
- 1.2 Optimal models of cognition . . . 12
- 1.3 Accounting for constraints . . . 14
- 1.4 The cost and value of mental action . . . 16
- 1.5 Representation, action, and the value of information . . . 20
- 1.6 The sequential nature of thought . . . 22
- 1.7 Optimal sequential models of resource-bounded cognition . . . 25

2 Formalism **28**
- 2.1 Markov decision processes . . . 28
- 2.2 Metalevel Markov decision processes . . . 32
- 2.3 Metalevel policies . . . 36
- 2.4 Two views of computation . . . 37
- 2.5 Bayesian metalevel MDPs . . . 38
- 2.6 Marginalized metalevel MDPs . . . 42
- 2.7 Identifying good metalevel policies . . . 46
- 2.8 Bayesian metalevel policy search . . . 51
- 2.9 Summary . . . 54

3 Attention **56**
- 3.1 Model . . . 59
- 3.2 Results . . . 65
- 3.3 Discussion . . . 75
- 3.4 Methods . . . 79

4 Memory **86**
- 4.1 Model . . . 91
- 4.2 Results . . . 96
- 4.3 Discussion . . . 106
- 4.4 Methods . . . 114

5 Planning **126**
5.1 Model . . . 128
5.2 Results . . . 133
5.3 Discussion . . . 144
5.4 Methods . . . 148

6 Conclusion **158**
6.1 Kindred efforts . . . 159
6.2 Points of weakness and avenues for growth . . . 162
6.3 Parting words: On frameworks . . . 167

Appendix A Supplementary information for Chapter 3 **169**
A.1 Task descriptions . . . 169
A.2 UCB policy optimization . . . 171
A.3 Individual fits . . . 171
A.4 Parameter recovery . . . 172
A.5 Implementation and validation of the aDDM . . . 173
A.6 Derivations for VOI features . . . 177
A.7 Quality of the approximation method in Bernoulli model . . . 182

Appendix B Supplementary information for Chapter 4 **184**
B.1 Deviations from pre-registration . . . 184
B.2 Previously pre-registered experiments . . . 187
B.3 Optimizing the lesioned model to predict effects . . . 191

Appendix C Supplementary information for Chapter 5 **194**
C.1 Deviations from pre-registration . . . 194
C.2 Experiment 5: Interleaved planning and action . . . 196
C.3 Probability-based stopping rules . . . 197

References **227**

To Nora Harhen and Sam Cheyette,
for hearing about all the versions of this work that didn't work.

And to my mother and father,
for giving me the confidence to keep trying anyway.

Acknowledgments

Being a scientist is, I think, one of the greatest privileges a human being can experience. I am enormously grateful to all the people who have made it possible for me to experience it.

I thank Paul Pease, my highschool biology teacher, for planting in my mind the seed of scientific curiosity.

I thank Shimon Edelman, my undergraduate advisor, for nurturing the budding plant—for taking my naïve and ill-conceived ideas seriously, for instilling in me an appreciation for both function and process, and for demonstrating that one person can be both an engineer and an artist.

And I thank Tom Griffiths, my graduate advisor, for teaching me how to garden—for showing more than telling, for pushing me to do things the right way rather than the easy way, for surrounding me with kind and brilliant collaborators, and, most of all, for showing me not just how to do good science, but how to be a good scientist.

Speaking of those kind and brilliant collaborators, I thank all the co-authors I have had the pleasure of working with in graduate school: Antonio Rangel, Bas van Opheusden, Carlos Correa, Emily Liquin, Erin Grant, Falk Lieder, Gustav Karreskog, James Hillis, Jess Hamrick, Ken Norman, Mark Ho, Matt Hardy, Paul Krueger, Priyam Das, Qiong Zhang, Sayan Gul, Tania Lombrozo, Vael Gates, and Yash Jain.

In particular, I would like to thank the graduate students and postdocs who acted as mentors to me (in chronological order): Jess Hamrick, Falk Lieder, and Bas van Opheusden. In each of them I have found both invaluable technical skills and also a role model for the type of scientist I would like to be: rigorous, passionate, patient, knowledgeable, and kind.

I also thank Antonio Rangel, for the incisiveness he brought to the work in Chapter 3, which I have tried to emulate in all my work since.

I thank the organizers of the SLOAN-Nomis workshop on the cognitive foundations of economic behavior—Andrew Caplin, Ernst Fehr, and Michael Woodford—for organizing the most stimulating academic events I have had the privilege of attending, where I first met Antonio, and where I recently escaped from a prolonged lapse in enthusiasm for science.

I thank the Julia development team for designing an efficient high-level programming language, without which I would not not have been able to do the work in this dissertation.

I thank the Princeton Psychology/Neuroscience IT staff for maintaining the hardware necessary for me to run that Julia code, in particular, John Wiggins and Garrett McGrath.

I thank the administrative staff of the Princeton Psychology department for creating a welcoming and supportive environment, and in particular Jill Ray for guiding me through the graduation process.

I thank the members of my disseration committee—Jon Cohen, Ken Norman, Marcelo Mattar, Nathaniel Daw, and Tom Griffiths—for their comments and guidance as I wrote this dissertation.

I thank my mother and father, for prioritizing my education, for supporting the things that made me a better and happier person (and discouraging the rest), and for fostering in me the sense that my achievements would be bounded only by what I wished to achieve.[1]

[1] Unfortunately, this turns out not to be true, but it's a nice thing to believe for the first twenty years or so.

1 Introduction

How can we build theoretically satisfying and practically useful models of the human mind? Historically, there have been two broad approaches. The *rational* approach, exemplified by the work of David Marr (1982) and John Anderson (1990), focuses on characterizing the problems people have to solve and the optimal solutions to those problems. Under the assumption that the mind is well adapted to its environment, these optimal solutions then serve as models of cognition. Rational models are satisfying because they tell us *why* the mind works the way it does, and they are useful because they allow us to make generalizable predictions about how people will behave in new environments (i.e., rationally). However, by construction, such models don't explain *how* the mind achieves the rational ideal, and a growing list of systematic cognitive biases (Kahneman, 2011) draws their predictive utility into question.

In contrast, the *mechanistic* approach focuses on identifying the cognitive processes underlying behavior, often with an emphasis on explaining the behavioral idiosyncrasies that rational models gloss over. This approach can potentially tell us how the mind actually works, and it can produce extremely accurate models. However, lacking the optimality constraint, there is an enormous space of possible mechanistic models, and they typically have free parameters that are tuned for specific experimental setups. We are thus often left wondering why this specific model fit the data best, and whether it would continue to make good predictions in a slightly different context.

Although the rational and mechanistic approaches have traditionally been viewed as conflicting, the past decade has seen a resurgence of an old idea (Simon, 1955): rationality can be seen as a property of cognitive mechanisms themselves. Specifically, a cognitive mechanism is rational if it makes optimal use of limited cognitive resources. Going under various names—cognitively bounded rational analysis (Howes et al., 2009), computational rationality (Lewis et al., 2014; Gershman et al., 2015), and resource-rational analysis (Griffiths et al., 2015; Lieder & Griffiths, 2020) to name a few—this view suggests that we should not expect people to be rational in the traditional sense of taking actions that maximize expected utility (Von Neumann & Morgenstern, 1944). Instead, we should expect people to select actions using mental strategies that strike a good tradeoff between the utility of the chosen action and the cognitive cost of making the decision.

But what defines a "good" tradeoff between action utility and cognitive cost? And how can we identify mental strategies that achieve such a tradeoff? In this dissertation, I suggest answers to these questions based on a key insight: *a rational mental strategy is one that optimally solves the sequential decision problem posed by one's internal computational environment*. Under this view, cognition is a problem of stringing together a series of basic cognitive operations, or "computations", in the service of choosing what to do in the world. An optimal cognitive process strings those basic operations together in such a way that maximizes the difference between the utility of the ultimate behavior and the total cost of all the cognitive operations that support the behavior.

1.1 Sequential decisions in the world and the mind

To make things concrete, consider the problem facing a delivery robot, illustrated in Figure 1.1A. Completing the delivery will require visiting a sequence of locations before arriving at the final destination. And at each location, the robot will need to decide where to go next. Thus, the robot faces a *sequential decision problem*. Figure 1.1B illustrates how this type of problem is often modeled in artificial intelligence research. At each time step, an agent (here, the robot) takes an *action* (e.g., driving forward). This action causes the environment to enter a new *state* (e.g., one where the robot is in a new location). Additionally, the agent receives a *reward*, a number that captures how good or bad the immediate consequences of the action are. The robot's goal is to maximize the total reward received. For example, the delivery robot might receive a large positive reward for reaching the destination and a small negative reward every time it moves (capturing the desire to conserve battery life). After receiving the reward and new state, the agent selects another action and the cycle continues.

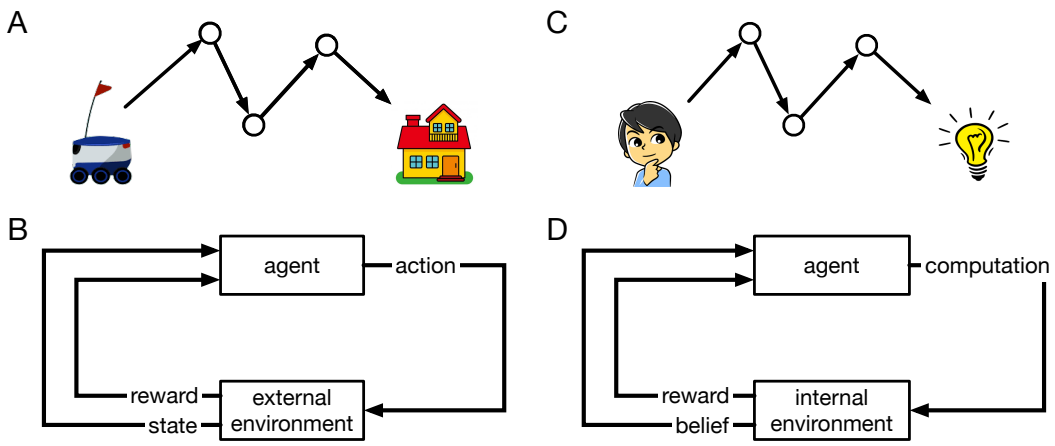

Figure 1.1: Sequential decision problems posed by external and internal environments.

Figure 1.1C illustrates a seemingly very different type of situation: a person trying to come up with a solution to a difficult problem. However, as the diagram suggests, the two cases actually share the same basic structure. Both involve an extended interaction between an agent and an environment; but whereas the robot is interacting with an *external* environment, the thinker is interacting with an *internal environment*: their own mind. Just as the robot makes several moves, and visits several locations before reaching the destination, the thinker has several thoughts, and enters several mental states before discovering the solution. Indeed, as illustrated in Figure 1.1D, this problem can be modeled in precisely the same way as the delivery problem. However, now the actions correspond to computations (thoughts) and the states correspond to beliefs (mental states). Thinking changes one's mental state just as moving changes one's physical state; and it also incurs a cost—at the very least, thinking takes time.

An important property of sequential decision problems is that there is often a dissociation between the short-term reward and the long-term *value* of performing some action. For example, if the robot had the option of simply sitting still, this would incur no cost and would thus be the most rewarding action in a myopic sense. However, the potential for the large reward associated with making a delivery makes paying this cost worthwhile. Thus, moving has value. By the same token, a truly myopic agent (one who only considers immediate rewards) would never do any thinking at all! Thinking only has value insofar as it can inform our future behavior.[1]

[1]Note that, subjectively, thought itself can be rewarding (sometimes intensely so; Gopnik, 1998). However, just as with "secondary reinforcers" like money, this is not because thought is inherently valuable, but because it is associated with value. Nevertheless, this association may be deeply ingrained, perhaps even genetically so. I return to this possibility in Section 6.2.1.

The power of identifying this parallel between external and internal environments is that it allows us to leverage existing knowledge about sequential decision problems (a substantial chunk of AI research) to build rational mechanistic models of cognition. That is, we can apply the same formalisms and algorithms that might help a robot deliver groceries to instead characterize the problem of resource-bounded cognition, and identify cognitive processes that optimally solve that problem.

A long history of work in artificial intelligence and cognitive science has established the foundations upon which this dissertation builds. Formally, our approach draws heavily on concepts and tools developed in *rational metareasoning*, a subfield of artificial intelligence that aims to construct artificial agents that make effective use of their limited computational resources (Russell & Wefald, 1991a). Indeed, in many ways, our approach is simply the application of these ideas to understanding human cognition. At the same time, many of these ideas have parallels in cognitive models, and still more can be traced to the period before the boundary between cognitive modeling and artificial intelligence was well-defined. Below, I review these ideas from a psychological perspective, noting the parallel concepts in AI when relevant. Ultimately, I will synthesize these key insights into a formal definition of rational mechanistic models of cognition.

1.2 Optimal models of cognition

What does it mean, formally, for a cognitive model to be rational? Following Anderson (1990), I formalize rationality in terms of *optimality*. A model of a cognitive process (henceforth just "a cognitive process") is optimal if it performs a cognitive function as well as it possibly could. More precisely, an optimal cognitive process is one that produces the maximum value of an *objective function*, out of a set of possible models:

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}}\, V(\pi). \tag{1.1}$$

Here, V (for value) is the objective function, and each $\pi \in \Pi$ is a possible cognitive model. Defining an optimal cognitive model thus amounts to specifying the function or goal of the cognitive process (through V), and the set of possible processes (through Π).

Perhaps the most fundamental optimal model is expected utility theory (Von Neumann & Morgenstern, 1944), which states that one should select actions that yield the best outcomes in expectation (i.e., on average). Given that the world is in state w, the expected util-

ity of action a is defined

$$\mathrm{EU}(w, a) = \mathop{\mathrm{E}}_{o|w,a}[U(o)] = \sum_{o} U(o)p(o \mid w, a). \tag{1.2}$$

Note that $\mathrm{E}_{o|w,a}$ denotes an expectation over o conditional on w and a. According to expected utility theory, the best action to take in each world state is the one that maximizes expected utility:

$$a^*_w = \operatorname*{argmax}_a \mathrm{EU}(w, a) \tag{1.3}$$

Although it was not originally conceptualized in this way, we can think of expected utility theory as an optimal model where the set of possible cognitive processes Π contains all possible mappings from world states to actions, and the objective function is defined as

$$V(\pi) = \mathop{\mathrm{E}}_{o|\pi}[U(o)] = \mathop{\mathrm{E}}_{w}\left[\mathop{\mathrm{E}}_{a\sim\pi(w)}\left[\mathop{\mathrm{E}}_{o|w,a} U(o)\right]\right]. \tag{1.4}$$

That is, choosing actions according to expected utility theory yields the highest utility outcomes, on average. In rational metareasoning, maximizing V (as defined in Equation 1.4) is known as *perfect rationality* (Russell, 1997).[2]

Expected utility theory is the foundation of modern (neoclassical) economics, allowing analysts to predict aggregate market behavior by assuming that each individual maximizes their own welfare. But it is so abstract that it initially seems to tell us little about cognition itself. Nevertheless, applying the optimization principle in more constrained domains (perhaps implicitly) has yielded important insights about many areas of cognition, including perception (Marr, 1982; Knill & Richards, 1996; Najemnik & Geisler, 2005) categorization (Anderson, 1991; Ashby & Alfonso-Reese, 1995; Tenenbaum & Griffiths, 2001), memory (Anderson & Milson, 1989), and language (Goldwater et al., 2009).

The power of optimization as a tool for cognitive modeling is that it makes decisions for us. That is, it reduces the amount of flexibility we have when specifying a model. This may sound like an impediment—and indeed, it sometimes feels that way—but it can have enormous benefits. Conceptually, optimal models provide a type of explanation that purely mechanistic models cannot: *teleological explanation*. That is, an optimal model tells us not just how a cognitive process works, but why it works that way. Characterizing a cognitive process as an optimal solution to a problem amounts to identifying that problem as the

[2] More precisely, perfect rationality is usually defined in terms of the total utility attained over the agent's lifetime, and it accounts for limits on the information available to the agent (i.e., not knowing w).

function of the cognitive process, and sometimes that function is different from what one might have initially assumed (e.g. Anderson & Milson, 1989).

From a statistical perspective, optimality acts as an *inductive bias*. Any given behavioral phenomenon is consistent with countless cognitive models (Anderson, 1978), but only a small subset of those are optimal (at least, for an intuitively plausible objective function). If people are well-adapted to their environment, then all else being equal, the optimal models are more likely to resemble the truth than an arbitrary alternative model. Of course, people are not perfectly adapted to their environment; thus, one can always achieve a more accurate model of a particular phenomenon by abandoning the assumption of optimality. But when data are limited, or we are exploring a domain we know very little about, or we want to generalize our predictions in non-trivial ways, having a constraint that is mostly true can improve our chances of making predictions that are mostly accurate.

Problems arise, however, when optimality isn't even mostly true. And there are many settings where that seems to be the case. To take just one example, recall that according to expected utility theory, one should take actions that yield the highest utility outcomes in expectation. This seems straightforward enough, right? Not for humans. People systematically violate expected utility theory, making choices that cannot be reconciled with any utility function (Allais, 1953; Ellsberg, 1961; Kahneman & Tversky, 1979). Since then, a major focus of research on human decision-making has been to characterize exactly how and why people violate expected utility theory, for example because they misperceive the probabilities involved (Kahneman & Tversky, 1979) or don't consider the different actions independently (Roe et al., 2001).

1.3 Accounting for constraints

Herbert Simon (1955) famously offered a different type of explanation for people's failure to maximize utility: People don't maximize Equation 1.4 because that's not the problem they're actually faced with. In most real-world problems, there are many actions, and each action could result in many different outcomes. Considering every possible outcome of every possible action each time we are faced with a decision would be paralyzing. Not to mention, the utility of an outcome is itself a complicated thing to evaluate (potentially requiring one to consider the action one would take next given that outcome) and we might not know the exact state of the world (either because the information isn't available or because we simply didn't take note of it). For decisions of the complexity that people face every day, even the largest supercomputers would be completely unable to evaluate expected utilities as in Equa-

tion 1.2. For a person to attempt to do so, using a biological computer that uses less power than an incandescent light bulb, wouldn't just be hopeless—it would be irrational.

This notion, that our choices reflect not only the expected utility of their outcomes, but also our limited ability to compute those utilities is called *bounded rationality* (Simon, 1990a). More generally, Simon suggests that human cognition—and the very definition of rational behavior—is shaped by both "the structure of task environments and the computational capabilities of the actor" (Simon, 1990b). By accounting for those constraints as part of the definition of the problem that people have to solve rather than a flaw in the solution, we can continue to apply the rationality principle in cases where computational constraints render perfect rationality (Equation 1.4) unattainable.

In modern psychological research, Simon's ideas are reflected in many different ways. One of the most prevalent is the notion of *ecological rationality* (Gigerenzer & Todd, 1999; Goldstein & Gigerenzer, 2002; Todd & Gigerenzer, 2012). Ecological rationality is based on the idea that people make decisions using simple, but adaptive heuristics. These heuristics are computationally "frugal", meaning that people can actually apply them in the real world, but they result in good decisions most of the time. This is possible because they take advantage of the structure of the problems people face frequently—they are adapted to our ecology. How can we identify such heuristics? Although Gigerenzer and colleagues emphasize the performance of the heuristics they propose (i.e., that they yield correct decisions; Goldstein & Gigerenzer, 2002), they explicitly reject the idea of optimizing that performance, suggesting that any form of optimization—*especially* optimization that accounts for cognitive constraints—is "demonic" (Gigerenzer & Todd, 1999; Chapter 1).

An alternative interpretation of Simon's ideas, the one adopted here, is that we can account for the role of cognitive constraints in defining the problems people face without abandoning the methodological tools we use to characterize rational solutions to those problems. Importantly, just as Bayesian models of perception do not claim that people are constantly evaluating multi-dimensional integrals in their heads, identifying optimal solutions to cognitively constrained problems does not imply that people actually perform that optimization. But this raises a new question: how exactly can we account for cognitive constraints within the optimality paradigm?

Lewis et al. (2014) propose a very natural answer: we can formalize bounded rationality as *bounded optimality*. Horvitz (1987) defines bounded optimality as "the optimization of computational utility given a set of assumptions about expected problems and constraints on resources" (compare to Simon's "structure of task environments" and "computational capabilities"). More formally, Russell & Subramanian (1995) define bounded optimality as a

property of a program, namely that it yields the maximum expected utility when executed on the agent's computational machine, or "brain" B. We can thus define a bounded optimal cognitive process as

$$\pi_B^* = \underset{\pi \in \Pi_B}{\operatorname{argmax}} \underset{o|\pi,B}{\mathrm{E}}[U(o)], \tag{1.5}$$

where Π_B is the set of all cognitive processes that can be executed by brain B and $\mathrm{E}_{o|\pi,B}[U(o)]$ is the expected utility of the outcome that results from executing process π on brain B.

As a definition of rationality for cognitive processes, bounded optimality is perfect in principle, but prohibitive in practice. In principle, it exactly defines the problem that a resource-constrained agent faces: maximizing utility given computational constraints. In practice, however, it is very difficult to identify bounded optimal cognitive processes. There are two reasons for this. First, it requires optimizing over the set of all cognitive processes a brain could execute. It is not clear how one would even specify this set, let alone find the optimum. Second, bounded optimality is not really a property of cognitive processes—it is a property of *minds*.[3] A mind serves many cognitive functions, and they all draw on the shared resources of one brain. Thus, we cannot directly apply the principle of bounded optimality to identify the optimal cognitive process for any particular function, because doing so would not account for how the resources used by this process might impact other processes. Taken at face value, bounded optimality prevents us from decomposing the monolithic task of characterizing "rational cognition" into a set of more manageable tasks of characterizing specific rational cognitive processes.

Fortunately, these two challenges are not insurmountable, at least if we are willing to make approximations. I address the problem of specifying possible processes in Section 1.6, and optimization in Chapter 2. The challenge of decomposability is addressed in the next section (in particular, Equation 1.8).

1.4 The cost and value of mental action

Traditionally, researchers in psychology have taken a different approach to accounting for cognitive constraints. Intuitively, thinking is *effortful*, and so we avoid it for the same reason we avoid carrying heavy groceries (Shenhav et al., 2017). On the other hand, more thinking generally results in better outcomes. Thus, there is a fundamental trade-off between the util-

[3] Similarly, in bounded optimality, the objective function is evaluated over the agent's entire lifetime, not a single outcome. This can can be addressed by the notion of long-term value (briefly described in Section 1.1 and formalized in Chapter 2). That is, we can define the utility of an outcome in terms of both the immediate and future rewards it yields.

ity of the outcomes we experience and the cost of the mental effort we spend to earn those outcomes (Kool & Botvinick, 2018). Critically, however, the optimal point on this trade-off is not always the same. For important decisions with clear factors to consider, thinking a lot is often worthwhile. For unimportant decisions, or ones that are too complex to effectively reason about, the benefit of thinking a lot is unlikely to outweigh the cost.

The idea that a rational cognitive process should allocate different amount of effort in different situations has been formalized in many different ways (Shenhav et al., 2013; Anderson, 1990; Lieder & Griffiths, 2017). However, all these different approaches draw on a key insight. The choice of how much effort to allocate is analogous to the choice of which action to take in the world. Accordingly, just as we can define the value of an *action* as the expected utility of the outcome it produces (Equation 1.2), we can define the value of *cognition* as the expected utility of the outcome it produces; in the latter case, however, we must also account for cognitive cost. Thus the "value of cognition" can be defined as

$$\mathrm{VOC}(w, c) = \mathop{\mathrm{E}}_{o|w,c}[U(o)] - \mathrm{cost}(c), \tag{1.6}$$

where c corresponds to a "cognitive action". Similarly, just as expected utility theory defines the optimal physical action to take in each state (Equation 1.3), we can define the optimal cognitive action to take in each state as

$$c^*_w = \operatorname*{argmax}_c \mathrm{VOC}(w, c). \tag{1.7}$$

In rational metareasoning, Equation 1.6 is called the *value of computation* (VOC; Russell & Wefald, 1991a).[4]

The content and interpretation of c varies substantially across different instantiations of this general principle. One influential example is the *expected value of control* (EVC; Shenhav et al., 2013). In EVC, c corresponds to a *control signal*, which modulates the behavior of an underlying cognitive process. In the simplest case, the control signal indicates simply how much effort the process exerts. In more complex cases, the control signal can have multiple dimensions, controlling not just the amount but also the nature of cognitive effort exerted (Musslick et al., 2015; Grahek et al., 2020; Ritz et al., 2021).

In another instantiation of this principle, c corresponds to a *strategy* for solving a problem. A classic example of this kind of model considers the decision of whether to use a "model-free" or "model-based" strategy (Daw et al., 2005). Intuitively, a model-free strategy

[4] More precisely, the VOC is typically defined as an *improvement* in expected outcome utility compared to if the computation were not executed.

relies on habits or learned associations, while a model-based strategy involves explicit reasoning about the consequences of one's actions. Generally, the model-based system yields better outcomes, but is more costly to execute; people seem to arbitrate between the strategies accordingly (Keramati et al., 2011; Kool et al., 2017), sometimes using strategies that combine model-free and model-based elements (Keramati et al., 2016). Although this work only makes an intuitive appeal to a cost-benefit tradeoff, Lieder & Griffiths (2017) have proposed a model of strategy selection that explicitly optimizes VOC (Equation 1.6), showing that it can account for the strategies people use to sort lists and make choices between alternatives that vary on many dimensions.

1.4.1 Bounded optimality and metalevel rationality

Treating the choice of cognitive actions as analogous to the choice of physical actions has yielded important insights into the flexibility with which people allocate mental effort. But as researchers, we now face a difficult choice. On the one hand, bounded optimality (Equation 1.5) defines the true problem of resource-constrained cognition. But a cost-benefit tradeoff (Equation 1.7) is something we can actually work with. Must we choose between principle and practice?

Perhaps not. As defined in Equation 1.5, bounded optimality poses a constrained optimization problem whose solution is a single cognitive process that solves all the problems the agent might face using a single shared resource, the brain. This is intimidating, to say the least. But it's not the only way to view the problem. Consider another case where an animal must allocate a finite shared resource among several different activities: foraging. Rather than allocating mental resources across different cognitive functions, the animal must allocate their time across different patches where food might be found. Initially, this problem seems very challenging because the correct amount of time to allocate to each patch depends on the time allocated to every other patch. But as shown by Charnov (1976), the optimal solution is actually quite simple: stay in a patch as long as the rate of food intake is more than what you could expect to find elsewhere (the average rate of intake in the past). More generally, one should allocate resources to an activity as long as the utility gained from that activity exceeds the *opportunity cost* of not allocating those resources to some other activity. As brilliantly observed by Kurzban et al. (2013), this principle applies equally to the allocation of cognitive resources. Thinking about any particular thing has an opportunity cost because it means you aren't thinking about something else.

Leveraging the insight that cognition carries an opportunity cost, Lieder (2018) showed that we can specify the value of a specific cognitive processes (in a bounded optimal sense)

as an additive combination of outcome utility and cognitive cost,

$$V_B(\pi) = \mathop{\mathrm{E}}_{o|\pi,B}[U(o)] - \mathrm{cost}_B(\pi), \tag{1.8}$$

where $\mathrm{cost}_B(\pi)$ captures the opportunity cost of the resources consumed by π on brain B.[5] This suggests that a cost-benefit tradeoff, such as we see in EVC, could actually be a direct expression of bounded optimality. Perhaps we do not need to make a choice after all!

If this seems too good to be true, that's because it is. To see why, we must ask ourselves: what is the optimal cognitive process in EVC, and what objective function is it optimizing? Initially, we might think of c^*_w as the optimal cognitive process. This makes intuitive sense given that c is often implemented as a parameter that modulates the behavior of an underlying process (like a drift diffusion model, described below; Musslick et al., 2015). But the problem with defining c^*_w as the optimal cognitive process is that we would then have a different objective function for each state. This amounts to having a different cognitive model for each condition in an experiment—an undesirable state of affairs. Instead, just as we viewed expected utility theory as an optimal mapping from states to actions, we must view the cognitive process in EVC (and other instantiations of Equation 1.7) as a mapping from states to control signals, with the objective function defined analogously to Equation 1.4:

$$V(\pi) = \mathop{\mathrm{E}}_{w}\left[\mathop{\mathrm{E}}_{c\sim\pi(w)}\left[\mathop{\mathrm{E}}_{o|w,c}[U(o)] - \mathrm{cost}(c)\right]\right]. \tag{1.9}$$

With some rearrangement, we can make this look more like the bounded optimal objective in Equation 1.8,

$$V(\pi) = \mathop{\mathrm{E}}_{o|\pi}[U(o)] - \mathop{\mathrm{E}}_{c|\pi}[\mathrm{cost}(c)], \tag{1.10}$$

but there is one critical difference: $\mathrm{cost}_B(\pi)$ is not equal to $\mathrm{E}_{c|\pi}[\mathrm{cost}(c)]$. The former capture the entire cost incurred by the cognitive process. The latter accounts for the cost of the cognitive actions that are selected, but not *the cost of selecting those actions*.

This is the difference between *metalevel rationality* (Equation 1.9) and bounded optimality (Equation 1.5). Metalevel rationality is defined as selecting computations (cognitive actions) that optimally trade-off utility with computational cost (Russell, 1997). But, unlike bounded optimality, metalevel rationality is not something that a physically implemented

[5] Lieder (2018) presents a derivation of the opportunity cost in terms of the amount of time that different resources are allocated (each of which is assumed to have some fixed cost). However, more nuanced factors are likely at play, such as the need to be able to quickly re-allocate resources when necessary (Musslick & Cohen, 2021), and the changing value of other cognitive activities (Agrawal et al., 2022).

agent can actually achieve; it assumes that computations are selected by a metalevel process, which is separate from the "object-level" process that actually performs the computations, and which is not itself subject to any cost or constraints. This assumption is particularly troubling in light of the parallel between selecting computations and selecting actions (between VOC and expected utility). The observation that people cannot maximize expected utility is what set us on our quest to characterize cognitive constraints, but it has led us right back to the assumption that people are maximizing expected utility! This circularity, sometimes called "the problem of infinite regress", is the reason many strategy-selection researchers have abandoned the metacognitive cost-benefit analysis of VOC in favor of models that select strategies through simpler, model-free mechanisms (Shrager & Siegler, 1998; Erev & Barron, 2005; Rieskamp & Otto, 2006).

Is all hope lost? Perhaps not. As suggested by the strategy-selection models just mentioned, one does not have to actually perform a cost-benefit analysis in order to select strategies that effectively balance costs and benefits. Building on this idea, Lieder & Griffiths (2017) present a model that explicitly approximates the VOC using learned predictive models of the computational cost and outcome utility that will result from applying a strategy to a particular problem. They call this strategy *metacognitive reinforcement learning*. Although a human being with finite experience may not be able to learn VOC exactly, they may be able to learn a relatively good approximation to VOC, especially for the kinds of metacognitive decisions they encounter frequently. In these cases, their behavior may come close to true metalevel rationality. By the same token bounded optimality and metalevel rationality will actually not be so different in these cases. Thus, going forward, we will use metalevel rationality as an approximation to bounded optimality. I return to this point in Section 6.2.1.

1.5 Representation, action, and the value of information

Our discussion so far has remained agnostic about an important question: How exactly does cognition influence outcomes? To answer this question, I will draw on another class of models that are used to characterize rational cognition under computational constraints, those based on *information theory*. These models vary widely in their structure and go under different names, including efficient coding (Barlow, 1961; Stocker & Simoncelli, 2006), rate-distortion theory (Sims, 2016), the free-energy principle (Friston, 2010), and rational inattention (Sims, 1998; Caplin & Dean, 2013).

The key idea in information-theoretic approaches to bounded rationality is to view cognitive processes as information channels. Specifically, the process consists of two components:

an encoder π_{enc} that transforms the world state w into a mental representation m, and a decoder π_{act} that chooses an action a based on that representation:

$$w \xrightarrow[\pi_{\text{enc}}]{} m \xrightarrow[\pi_{\text{act}}]{} a \tag{1.11}$$

The agent then receives utility $U(w, a)$, which we can interpret as the expected outcome utility given the state and selected action.

The agent's goal is to maximize utility minus[6] a cost on the "fidelity" of the encoder. Many forms of this cost are possible; the most common is *mutual information*, which specifies how much information the mental state has about the world state. Formally, the objective function is defined as

$$V(\pi) = \mathop{\mathrm{E}}_{w}\left[\mathop{\mathrm{E}}_{m\sim\pi_{\text{enc}}(w)}\left[U\big(w, \pi_{\text{act}}(m)\big)\right]\right] - \text{cost}(\pi_{\text{enc}}). \tag{1.12}$$

There are two key ideas to take away from this equation. First, cognitive processes can produce intermediate mental states, which are used to select actions. Second, both the action utility and cost depend on the mental state, and these two forces are in a fundamental conflict. A precise representation of the world allows one to select good actions, but such representations are costly to form.

Rational metareasoning provides another way to think about the mental state. Specifically, we can think about m as a *belief* about the state of the world, $p(w \mid m)$. Assuming that π_{act} is optimal (that is, that it always selects the action with highest expected utility given the representation), we can define the *value of information* (Matheson, 1968) in the representation as the expected utility of the action that would be taken based on that representation:

$$\text{VOI}(m) = \mathop{\mathrm{E}}_{w|m}\left[U\big(w, \pi_{\text{act}}(m)\big)\right] = \max_{a} \mathop{\mathrm{E}}_{w|m}[U(w, a)] \tag{1.13}$$

A derivation for the second equality is given in Equation 2.28 (see Section 2.8.1 for more discussion of VOI). Using VOI, we can rewrite the information-theoretic objective function (Equation 1.12) as

$$V(\pi) = \mathop{\mathrm{E}}_{m|\pi}[\text{VOI}(m)] - \text{cost}(\pi_{\text{enc}}) \tag{1.14}$$

[6] Rate-distortion models are often framed in terms of a fixed capacity limit rather than a continuous cost. However, because the commonly used mutual information cost is defined as an expectation over values of m and w, one can find a cost-multiplier λ that satisfies any given constraint C (the Lagrange multiplier; see Ortega & Braun, 2013). Note, however, that the λ-to-C conversion depends on the utility function and distribution of w. Thus, assuming a fixed λ (rather than fixed C) predicts that the information rate can change across experimental conditions (and indeed it appears to do so; van den Berg & Ma, 2018)

Thus, from an information-theoretic standpoint, the value of a cognitive process is the difference between the value of the information it produces and the cost of that information.

Information-theoretic models have yielded important insights about how we represent the world, in particular in visual perception (Attneave, 1954; Barlow, 1961; Simoncelli & Olshausen, 2001) and working memory (Sims et al., 2012; van den Berg & Ma, 2018). Typically, however, these models do not emphasize the down-stream function of mental states for directing action (but see Yoo et al., 2018; Bates et al., 2019).

In contrast, information-theoretic models of decision making often focus exclusively on actions. In fact, with the commonly used mutual information cost, the optimal encoder assigns a single representation to each action; that is, there is a one-to-one correspondence between a and m (Matêjka & McKay, 2015). This allows one to eliminate m entirely, yielding a reduced-form model in which the cost is simply the mutual information between state and action, as proposed in free-energy models (Friston, 2010; Ortega & Braun, 2013). This reduction is mathematically fascinating, and it can explain some interesting cognitive phenomena such as perseveration (Gershman, 2020). But if our goal is to define rational *mechanistic* models of cognition, we seem to be going backwards. By eliminating the mental representation from the model, we end up with a more sophisticated form of behaviorism, where we attempt to understand a cognitive process simply as a mapping from stimulus to response.[7]

1.6 The sequential nature of thought

Thus far, I have described models that rationally select mental actions, and models that form mental representations to inform the actions they take in the world. These are key ingredients for a rational mechanistic model of cognition, but something is still missing. All the models we've discussed so far treat the problem of rational metacognitive control as a *static* problem (or a sequence of independent static problems). That is, they consider the selection of a single mental action or a single mental state.[8] But cognition is not static; it is *dynamic*. Cognitive processes are not defined by solitary representations or operations; they consist of sequences of mental states, and the mental actions that produce and transform them.

[7] Importantly, this reduction is only possible when making very weak assumptions about the constraints on the encoder. Specifying cognitively-motivated constraints can yield models that give greater insight into the mental representations and processes underlying decision-making (e.g., Bhui & Gershman, 2018).

[8] It is perhaps unfair to call EVC or strategy selection "static", given that the control signals often parameterizes a dynamic evidence accumulation process and the strategies are correspond to multi-step algorithms. More precisely, it is the rational metacognitive component of these models that is static. I discuss this further at the end of this section.

The fact that cognitive processes are sequential, in the sense that they occur over time rather than all at once, is self-evident. Thus, sequentiality is a common feature of mechanistic models of cognition. One widely used class of sequential models are *evidence accumulation* models (also called "sequential sampling models"), such as the drift diffusion model (DDM; Ratcliff, 1978), leaky competing accumulators (Usher & McClelland, 2001), and decision by sampling (Stewart et al., 2006). According to these models, decision making involves accumulating noisy evidence in favor of each possible choice until the evidence for one choice is sufficiently greater than the evidence for the other(s). That is, the cognitive process can be understood as a sequence of operations, each of which accumulates a small amount of evidence. In their simplest form, there is only one type of operation; these models can explain not only the choices we make (including when we make mistakes), but also how long it takes to make those choices. More complex evidence accumulation models take into account the possibility of attending to different sources of information, such as different options (Krajbich et al., 2010) or attributes (Russo & Dosher, 1983); these models can account for (if not predict) additional data such as what we look at when making a decision, and they can account for systematic deviations from expected utility (Busemeyer et al., 2019).

Another important class of dynamic models, *cognitive architectures*, aims to capture a more diverse range of mental activities, beyond simply accumulating more evidence. Cognitive architectures, most notably ACT-R (Anderson, 1996) and SOAR (Laird et al., 1987), explicitly model individual cognitive operations such as perceptually encoding a stimulus, recalling information from memory, and transforming symbolic representations the world. These models can trace their intellectual roots to the infancy of artificial intelligence research (Newell & Simon, 1956), where the discovery that digital computers could solve complex problems by breaking them down into a sequence of very simple operations led to the hypothesis that a similar principle might underlie human intelligence (Newell et al., 1958, 1972). The core assumption of these models is that all cognitive processes can be broken down into *elementary information processes* (Simon, 1979; Posner & McLeod, 1982; Chase, 1978), or simply, "cognitive operations".

How can we formally define sequentiality? Intuitively, a sequential cognitive process can be decomposed into two components: the cognitive architecture and the strategy for using that architecture. The cognitive architecture consists of the set of operations that can be performed $\mathcal{C}$, the set of (mental) states the system can be in $\mathcal{M}$, and the relationship between the two: the way that cognitive operations affect mental state. We can define this relation-

ship with a *transition function* T, such that

$$m_{t+1} \sim T(m_t, c_t, w). \tag{1.15}$$

That is, at each time point, the next mental state is (stochastically) determined by the previous mental state, the operation that was executed, and the state of the world. Together, $\mathcal{C}$, $\mathcal{M}$, and T can be seen as a formalization of the agent's "computational capabilities" (Simon, 1955), or the architecture on which a program is executed (Russell & Subramanian, 1995).

The second component of a sequential cognitive process is the strategy for using the architecture. Drawing on the models discussed in Section 1.4, we will formalize this strategy in terms of a policy for choosing mental actions. However, there, we assumed that the agent selected cognitive operations based on the state of the world. To capture the fact that the right operation to execute depends on one's mental state (and also the fact that the world state is often not directly accessible), we can instead assume that the operation is selected based on the current mental state, that is,

$$c_t \sim \pi(m_t). \tag{1.16}$$

Note that I use the notation π to refer specifically to the strategic component of the cognitive process. This reflects the assumption that the agent has control over their strategy, but not their cognitive architecture—that is, only the strategy π can be optimized. On developmental and evolutionary timescales, however, the architecture itself may itself be optimized (subject, of course, to some form of constraint). I return to this point in Section 6.2.4.

Formalizing cognitive processes as mappings from mental states to cognitive operations provides a very general framework, one in which nearly any existing cognitive process model can be formalized. However, this generality has a price. Many models have a very large space of possible mental states, and we need to specify which cognitive operation is executed in each one. If there are M possible mental states and C possible operations, there are C^M possible (deterministic) mappings. If M is large—and it often is—this results in a huge space of possible cognitive processes. Searching this entire space for the one that best accounts for human data is effectively impossible. Payne et al. (1988) circumvented this problem by manually specifying a small set of candidate strategies (mappings). But how can we be sure that this set includes the strategies people use, or that achieve the best cost-benefit trade-offs? Indeed, as shown by Howes et al. (2009), failing to consider all the possible strategies can yield incorrect conclusions about which cognitive architectures are consistent with a given pattern of behavioral data. But as the set of strategies grows, more

and more architectures become consistent with the data. As Howes et al. convincingly argued, adopting additional constraints—specifically, the assumption of optimality—may be necessary to distinguish between candidate architectures (see also, Lewis et al., 2014).

Before we continue, it is important to clarify how the type of sequentiality described in this section (and the proposed framework) differs from the models described in Section 1.4. Those models were actually not static in a strict sense, as the mental actions under consideration themselves corresponded to dynamic cognitive processes (e.g. decision-making strategies). Critically, however, those processes were not themselves subject to rational metacognitive control. In contrast, in the proposed framework, we will assume that the entire cognitive process is optimized; that is, we will consider cognitive processes that are not just sequential and optimal, but *sequentially optimal*.[9]

1.7 Optimal sequential models of resource-bounded cognition

The literature reviewed above highlights just a sliver of the progress we have made in the project of building rational mechanistic models of cognition. Each of these research programs has yielded important insights into the structure that rational mechanistic models should have. In this dissertation, I attempt to synthesize key insights from each of these programs into a framework for constructing rational mechanistic models that are (approximately) bounded optimal and sequential:

1. An optimal cognitive process is one that maximizes an objective function (Section 1.2).
2. Cognitive processes are shaped by the structure of both the external environment and the agents computational capabilities (Section 1.3).
3. The value of a mental action depends on both the utility of the outcome it produces and the opportunity cost of performing that action (Section 1.4).
4. The behavior produced by a cognitive process is mediated by the mental state it produces (Section 1.5).
5. A cognitive process can be broken down into a sequence of elementary cognitive operations (Section 1.6).

[9] The original definition of EVC (Shenhav et al., 2013) defined outcome utility in terms of EVC at the next time step, making EVC a sequentially optimal framework in principle. In practice, however, most applications of EVC do not actually use this feature, instead assuming that the utility of the outcome depends only on the behavior in the current trial. One notable exception is the visual attention model in Lieder et al. (2018), which employs an early version of the framework proposed here.

Combining these ideas, yields the following definition of an optimal sequential cognitive process:

$$\pi^* = \underset{\pi}{\text{argmax}}\ \mathrm{E}\left[U(w, \pi_{\text{act}}(m_N)) \ - \sum_{t=1}^{N} \text{cost}(c_t) \ \middle|\ c_t \sim \pi(m_t),\ m_t \sim T(m_{t-1}, c_{t-1}, w) \right] \tag{1.17}$$

That is, the optimal cognitive process is a way of selecting cognitive operations that yields a mental state from which high utility actions can be taken, while at the same time minimizing the total cost of those operations. The equation is certainly complex; but as the coloring reveals, none of these ideas are new.

The idea combine these ideas is not entirely new either. For example, there is a long history of modeling optimal speed-accuracy tradeoffs using evidence accumulation models, like the ones described in the previous section. Bogacz et al. (2006) showed that the drift diffusion model is the continuous limit of the *sequential probability ratio test* (SPRT; Wald, 1945), an optimal stopping rule for collecting evidence about binary hypotheses (c.f. Gold & Shadlen, 2002). For any given level of evidence coherence (problem difficulty), there is some fixed decision threshold that maximizes the reward rate; this is equivalent to maximizing a linear combination of choice utility and cost under certain assumptions.[10]. However, this analysis relies on a number of assumptions that rarely hold in the real world.

Explicitly modeling evidence accumulation as a sequential decision problem has allowed researchers to characterize optimal stopping rules in more complex cases (Drugowitsch et al., 2012; Fudenberg et al., 2018; Tajima et al., 2019). Going beyond the decision of *when* to accumulate evidence, the question of *where* to accumulate evidence has been addressed in optimal models of visual search, which characterize the optimal sequence of eye movements one should make when searching for a target image among distractors (Butko & Movellan, 2008; Acharya et al., 2017; Hoppe & Rothkopf, 2019). Optimal sequential models have also been applied to characterize rational strategies for using working memory (O'Reilly & Frank, 2006; Todd et al., 2008; Suchow & Griffiths, 2016), episodic memory (Lu et al., 2022), and long-term memory (Zhang et al., 2022). I further discuss these related approaches in Section 6.1.

This dissertation helps to unify the work reviewed above by providing a common frame-

[10] For a given decision problem, there is some time cost such that maximizing the reward rate will also maximize the reward minus cost (Wald & Wolfowitz, 1948; Drugowitsch et al., 2012). However, this mapping will vary for problems of different difficulty or stakes. For example, increasing the payoffs for correct and incorrect decisions by a multiplicative factor does not change the optimal reward-rate strategy, but it does change the optimal utility-minus-cost strategy.

work for modeling many different kinds of cognitive processes. Furthermore, by explicitly connecting to both the rational metareasoning literature and the more general reinforcement learning literature, we build bridges with which we can import both mathematical tools and psychological concepts. The former allows us to identify optimal strategies for architectures that are too complex for standard methods (e.g., Section 2.8). The latter allows us to apply what we know about how people learn to act adaptively in the world to understand how they learn to think adaptively using their own cognitive architectures (through "metacognitive reinforcement learning", discussed further in Section 6.2.1).

In the next chapter, I present the formal framework that makes this possible: *metalevel Markov decision processes*. This framework formalizes the problem of sequential interaction with a cognitive architecture, the problem for which Equation 1.17 is the optimal solution.

We must be prepared to accept the possibility that what we call "the environment" may lie, in part, within the skin of the biological organism

Herbert Simon (1955)

2

Formalism

Meta-level Markov decision processes

The key insight behind the proposed framework is that cognitive processes are solutions to sequential decision problems. Drawing on a subfield of artificial intelligence known as *rational metareasoning* (Matheson, 1968; Russell & Wefald, 1991a), we formalize this insight using the framework of *metalevel Markov decision processes* (metalevel MDPs; Hay et al., 2012). In this framework, a cognitive process is formalized as a sequential process of executing computational actions that update an agent's mental state. At each moment, the agent must choose whether to continue thinking, refining their mental state but accruing computational cost, or to instead stop thinking and take action. In the former case, they must additionally decide which computation to execute next (i.e., what to think about). In the latter case, they select the action that seems best given their current mental state and receive a reward associated with the external utility of that action.

In this chapter, I provide a formal description of the framework. The formalization presented here is an adaptation of the framework proposed in Hay (2016). I have modified the framework to better facilitate the specification of cognitive models.

2.1 Markov decision processes

A metalevel Markov decision process is an extension of a standard Markov decision process (MDP), illustrated in Figure 2.1. Thus, I begin with a brief overview of standard MDPs. See Puterman (2014) and Sutton & Barto (2018) more thorough overviews.

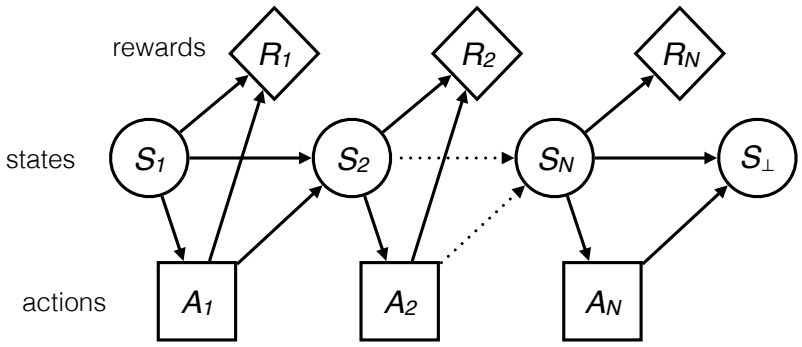

Figure 2.1: Markov decision processes. A Markov decision process (MDP) formalizes the problem of acting adaptively in a dynamic environment. The agent executes actions (squares) that change the state of the world (circles) and generate rewards (diamonds), which the agent seeks to maximize. The arrows indicate direct causality; thus, the reward and state at each time step depend only on the state and action at the previous time step, and the agent selects an action based only on the current state. The dotted arrow indicates the elided sequence of states between the first and last two.

MDPs are the standard formalism for modeling the sequential interaction between an agent and a stochastic environment. An MDP is defined by a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition function T, and a reward function R. A state $s \in \mathcal{S}$ specifies the relevant state of the world. An action $a \in \mathcal{A}$ is an action the agent can perform. The transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$[1] encodes the dynamics of the world as a distribution of possible future states for each possible previous state and action. Finally, the reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the expected[2] reward or utility for executing a given action in a given state. We additionally assume an initial state s_1 that the environment is initialized in and a set of terminal states $\mathcal{S}_\perp$ such that the episode ends when the agent reaches one of those states. An *episode* describes one interaction between the agent and the environment (beginning in the initial state and ending in a terminal state).

2.1.1 Optimal policies and value functions

The solution to an MDP is a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that selects which action to perform next given the current state. That is, $a_t \sim \pi(s_t)$. The goal is to find a policy that maximizes the

[2]One could equally well specify a stochastic reward function $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$. Because the optimal policy only depends on the expected reward, we do not consider this case here. However, a stochastic reward function would affect the performance of a learning algorithm; this could easily be integrated into the framework.

[2]$\Delta(\mathcal{S})$ denotes the set of all distributions over the set $\mathcal{S}$. Note that this definition is equivalent to defining the transition function as a probability mass function (i.e., $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$). I will use $T(s' \mid s, a)$ to denote the probability of transitioning to state s' when executing action a in state s.

expected cumulative reward attained, that is, the *return*. The optimal policy is thus defined,

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathrm{E}\left[\sum_{t=1}^{N} R(s_t, a_t) \,\middle|\, a_t \sim \pi(s_t)\right], \tag{2.1}$$

where N is the timepoint at which the episode terminates (when $s_{t+1} \in \mathcal{S}_\perp$). Note that the expectation implicitly conditions on the transition function, i.e., $s_{t+1} \sim T(s_t, a_t)$.

How can we identify such a policy? This question is the subject of a huge field of research in artificial intelligence, and countless methods have been developed. Many of these methods draw on the concept of a *value function*. The *state* value function (or just "value function") is defined as

$$V^\pi(s) = \mathrm{E}\left[\sum_{t=1}^{N} R(s_t, a_t) \,\middle|\, s_1 = s,\ a_t \sim \pi(s_t)\right]. \tag{2.2}$$

It specifies the expected total reward one will receive if one begins in state s and selects actions according to the policy π. Similarly, the *action* value function (or "state-action value function") is defined as

$$Q^\pi(s, a) = \mathrm{E}\left[\sum_{t=1}^{N} R(s_t, a_t) \,\middle|\, s_1 = s,\ a_1 = a,\ a_{t\neq 1} \sim \pi(s_t)\right]. \tag{2.3}$$

The action value function just like the state value function except that it also specifies the first action to be taken.

The value functions for the optimal policy are called the optimal value functions. They can be defined simply $V^* = V^{\pi^*}$ and $Q^* = Q^{\pi^*}$. By combining Equations 2.2 and 2.3 with Equation 2.1, we can see that the optimal value functions specify the maximal expected reward one could expect to gain beginning with a given state (and action) under any policy,

$$\begin{aligned} V^*(s) &= \max_{\pi} \mathrm{E}\left[\sum_{t=1}^{N} R(s_t, a_t) \,\middle|\, s_1 = s,\ a_t \sim \pi(s_t)\right] \\ Q^*(s, a) &= \max_{\pi} \mathrm{E}\left[\sum_{t=1}^{N} R(s_t, a_t) \,\middle|\, s_1 = s,\ a_1 = a,\ a_{t\neq 1} \sim \pi(s_t)\right] \end{aligned} \tag{2.4}$$

Putting aside for now the problem of identifying these functions, the optimal policy can

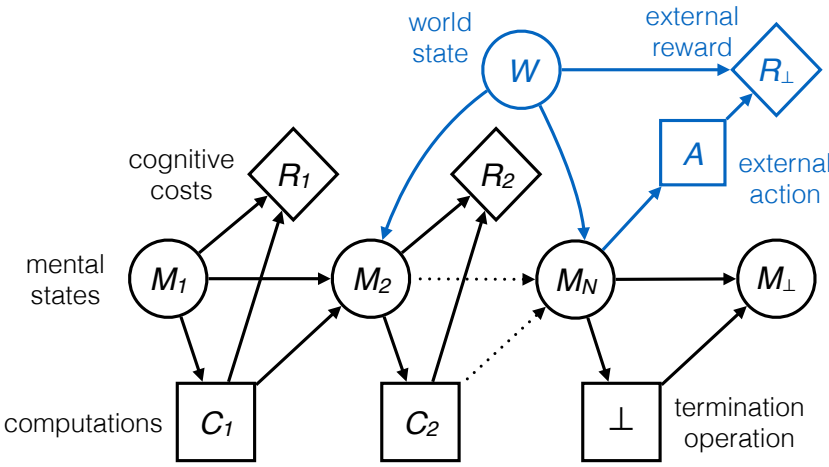

Figure 2.2: Metalevel Markov decision processes. A metalevel MDP formalizes the problem of thinking efficiently within one's own internal mental environment. The agent executes computations that update their mental state and incur cognitive cost. When the agent executes the termination operation $\perp$, they take an action in the external environment and receive an external reward. The elements that capture the external environment are indicated in blue. These formally distinguish metalevel MDPs from standard MDPs.

be defined as simply

$$\pi^*(s) = \text{Uniform}\left(\underset{a}{\text{argmax}}\, Q^*(s,a)\right). \tag{2.5}$$

That is, the optimal policy selects the action that produces the greatest expected long-term reward, breaking ties randomly.[3] When modeling human data, one typically assumes that this maximization is performed imperfectly. In particular, we will assume that actions are drawn according to a softmax (or Boltzmann distribution),

$$\pi(a \mid s) \propto \exp\{\beta \cdot Q^*(s,a)\} \tag{2.6}$$

where the inverse-temperature parameter β controls how well the policy maximizes, behaving completely randomly when $\beta = 0$ and approaching the optimal policy as $\beta \to \infty$.

This concludes our brief overview of MDPs. We are now ready to describe metalevel MDPs.

[3] To be more precise, Equation 2.5 defines the *maximum entropy* optimal policy, that is, the optimal policy whose action distributions are maximally random. For simplicity, I will refer to the maximum entropy optimal policy as simply "the" optimal policy.

Meta-level Markov decision processes (metalevel MDPs) extend the standard MDP formalism to model the sequential decision problem posed by resource-bounded computation (Hay et al., 2012). Like a standard MDP, a metalevel MDP is defined by sets of states and actions, and transition and reward functions. However, as illustrated in Figure 2.2, the states corresponded to mental states, and the actions correspond to computations (cognitive operations). The metalevel transition function describes how computations update mental states, and the metalevel reward function captures both the internal costs (e.g., time) and external benefits (e.g., better decisions) of computation.

Formally, we define a metalevel MDP by a set of mental states $\mathcal{M}$, a set of computations $\mathcal{C}$, a transition function T, and a reward function R. These four components are analogous to the states, actions, transition function, and reward function in a standard MDP. We additionally define a set of world states $\mathcal{W}$, upon which the transition and reward functions depend. This is the only formal distinction between a metalevel MDP and a standard MDP. However, as I discuss in Section 2.6, we can sometimes convert metalevel MDPs into equivalent MDPs by marginalizing over the world state.[4] I now describe the five components of a metalevel MDP in more detail.

To make things concrete, we will use a simple running example based on the tallying heuristic for deciding between two possible actions (Gigerenzer & Gaissmaier, 2011). We will consider applying this heuristic to the decision of which car to purchase. To follow this heuristic, we consider a sequence of "cues" that might discriminate between the options, marking which car each cue favors (which has better gas mileage? which is more comfortable? etc...). As we go, we simply count up how many cues have favored each car. Then, after considering some number of cues, we pick the car that is favored by more cues. But how many cues should we consider? Gigerenzer & Gaissmaier propose considering a fixed number of cues (and, in the case of a tie, considering more cues until one car has more in its favor). Is this the best version of tallying we could use? To answer this question, we can define tallying as a metalevel MDP.

A well-documented and (hopefully) beginner-friendly Julia implementation of the tallying metalevel MDP can be found at `https://github.com/fredcallaway/metamdp-example`.

[4] This is similar to the *belief MDP* transformation of a partially observable MDP (POMDP; Kaelbling et al., 1998). see Section 2.5.4 for additional discussion on the relation between metalevel MDPs and POMDPs.

2.2.1 World states

A world state $w \in \mathcal{W}$ captures the state of the world that is relevant to the agent's current task. Importantly, the agent does not have direct access to the world state, but must infer it from the outcome of computations they perform. Formally, the world state includes any information that is not known to the agent, but affects either the reward or transition functions. In the tallying example, the world state specifies the ratio of cues that favors one car vs. the other; thus, $\mathcal{W}_{\text{tally}} = [0, 1]$.

2.2.2 Mental states

A mental state $m \in \mathcal{M}$ captures the agent's internal state, as relevant to the task at hand. The interpretation of a mental state can vary from model to model. In some cases, it will correspond to a belief, or a representation of the world, but it can also capture arbitrary variables in a cognitive model. In the tallying example, the mental state specifies the number of cues one has considered that favor each option. Formally, we define each $m \in \mathcal{M}_{\text{tally}}$ as a tuple (x, y), where x is the number of cues favoring car X and y is the number of cues specifying car Y.

Analogously to standard MDPs, we additionally specify an initial mental state, m_1. This is the mental state the agent has at the beginning of each task instance. In the tallying example, the initial mental state is defined $m_1 = (0, 0)$. Additionally, all metalevel MDPs have a single terminal state $m_\perp$, which is only entered when computation is terminated (as described below).

2.2.3 Computations

A computational operation $c \in \mathcal{C}$ is a primitive operation afforded by the agent's cognitive architecture. Formally, it is a metalevel action that changes the mental state in much the same way as an external action might change the world state. In a metalevel MDP model, all cognition can be broken down into a sequence of these computations, but the model makes no attempt to explain how those basic operations are themselves implemented. The concept is thus very similar to *elementary information processes* (Chase, 1978; Simon, 1979; Posner & McLeod, 1982; Payne et al., 1988). In the tallying example, there is a single computation, which corresponds to considering another cue.

All metalevel MDPs include a special computation, the termination operation $\perp$, which indicates that computation should be terminated. Upon termination, the agent performs an external action, specifically the action that has maximal expected utility given the current

mental state. Thus, the most fundamental metalevel problem—how long to compute—is captured by the decision about when to execute $\perp$. In the tallying example, executing $\perp$ corresponds to purchasing the car that is favored by more cues (choosing randomly in the case of a tie.)

2.2.4 Transition function

The transition function $T : \mathcal{M} \times \mathcal{C} \times \mathcal{W} \to \Delta(\mathcal{M})$ describes how computation updates mental states. Formally, $T(m, c, w)$ is a distribution of possible new mental states that would result from performing a computation c in mental state m when the true state of the world is w. At each time step, the next mental state is sampled from this distribution:

$$m_{t+1} \sim T(m_t, c_t, w). \tag{2.7}$$

Terminating computation (executing $\perp$) always transitions to the terminal state, $m_\perp$:

$$T(m, \perp, w) = \text{Uniform}(\{m_\perp\}) \tag{2.8}$$

In the tallying example, the transition function specifies the probability that each cue count will be incremented when one considers another cue:

$$T_{\text{tally}}(m_t, c_t, w) = \begin{cases} (x_t + 1, y_t) & \text{with probability } w \\ (x_t, y_t + 1) & \text{with probability } 1 - w, \end{cases} \tag{2.9}$$

where $m_t = (x_t, y_t)$.

2.2.5 Reward function

The metalevel reward function $R : \mathcal{M} \times \mathcal{C} \times \mathcal{W} \to \mathbb{R}$ describes both the costs and benefits of computation. It is defined in terms of three components: a cost function that specifies the cost of executing each computation, an action policy π_{act} that selects an action to take given a mental state, and a utility function U that specifies the utility of each possible action in each world state.

The cost of computation is captured in the reward for non-terminal operations:

$$R(m, c, w) = -\text{cost}(m, c) \text{ for } c \neq \perp. \tag{2.10}$$

We assume that the cost of a computation can depend on the current mental state but not

on the state of the world. The cost of computation may include multiple factors. At a minimum, it captures the opportunity cost of the time spent executing the computation (rather than taking actions in the world). The simplest choice is to assume a constant cost for each computation executed. This is a natural choice for the tallying example, with total cost proportional to the number of cues considered.

The benefits of computation are captured by the reward for the termination operation, $R(m, \perp, w)$. Intuitively, the benefit of computation is that it allows one to take better actions in the world. Formally, the reward for termination is defined as the utility of the external action that the agent would execute given the current mental state,

$$R(m, \perp, w) = U\big(w, \pi_{\text{act}}(m)\big) \tag{2.11}$$

where $U(w, a)$ specifies the utility of executing action a in the world state w and π_{act} is the *action selection policy*, which chooses an action to take based on the current mental state. To simplify notation, I will assume that π_{act} selects an action deterministically, but it can also return a distribution over actions. This policy should be very simple. That is, the mental state should contain sufficient information to choose an action without much additional computation.

In the tallying example, π_{act} deterministically selects the car with more favorable cues, selecting randomly otherwise. Specifying U is less straightforward. For simplicity, we assume that the utility derived from car X is proportional to the ratio of cues favoring X. This results in

$$U_{\text{tally}}(w, a) = \begin{cases} w & \text{if } a = \text{X} \\ 1 - w & \text{if } a = \text{Y}. \end{cases} \tag{2.12}$$

Substituting our definitions of U and π_{act} into Equation 2.11 yields:

$$R_{\text{tally}}(m, \perp, w) = \begin{cases} w & \text{if } x_t > y_t \\ 1 - w & \text{if } x_t < y_t \\ 1/2 & \text{if } x_t = y_t \end{cases} \tag{2.13}$$

where the $1/2$ comes from taking the average of w and $1 - w$.

Equation 2.11 may appear to constrain us to simple one-shot decision-making tasks, such as the ones tallying is intended for. However, it is not as restrictive as it first appears. For example, in Chapter 4, we model a memory recall task by assuming that π_{act} can only perform the "recall" action when the activation of a memory exceeds a threshold. In Chapter 5, we

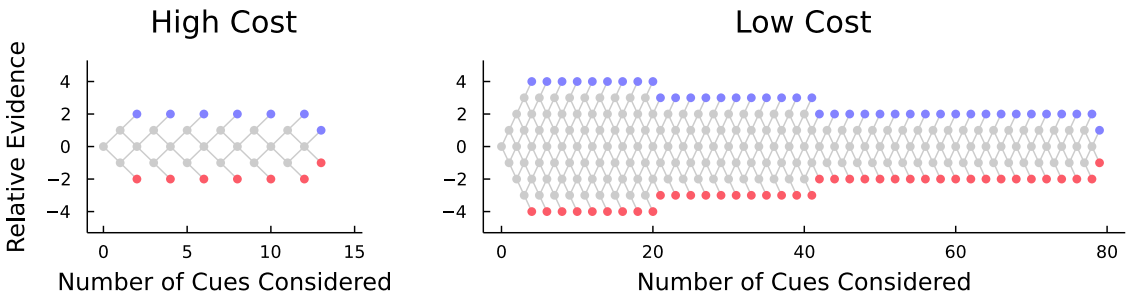

Figure 2.3: Optimal policies for the tallying metalevel MDP. The policy is illustrated as a graph, where each node represents a mental state. States in which the optimal policy continues computing are gray, and states in which the optimal policy terminates are in blue and red, corresponding to the two possible external actions. The two panels show the optimal policy under different levels of computational cost (0.01 and 0.002, in units defined by the maximal action utility of 1). The code that generated this figure can be found at `https://github.com/fredcallaway/metamdp-example`.

model a planning task by defining a abstractly, as a sequence of concrete actions (or more generally, an *option*; Sutton et al., 1999).[5]

2.3 Metalevel policies

If a metalevel MDP defines the problem a cognitive process must solve, a metalevel policy defines the solution. It is a strategy for selecting which cognitive operation to execute next given the current mental state. Formally, the metalevel policy, $\pi : \mathcal{M} \to \Delta(\mathcal{C})$, is a mapping from beliefs to distributions over computations. At each time step, the next computation is drawn from this distribution, $c_t \sim \pi(m_t)$.

How should we determine this policy? The classical cognitive modeling approach is to specify a plausible strategy, perhaps motivated by aspects of human behavior. In Chapter 5, we show how classical heuristics for decision-tree search can be naturally modeled as policies in a metalevel MDP. However, in the resource-rational approach pursued here, we take a different approach. Specifically, we are interested in the *optimal policy* for the metalevel MDP. Paralleling Equation 2.1, the optimal metalevel policy is defined as

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \, \mathrm{E}\left[\sum_{t=1}^{N} R(m_t, c_t, w) \;\middle|\; c_t \sim \pi(m_t)\right]. \tag{2.14}$$

That is, it maximizes the expected return. In a metalevel MDP the return can be broken

[5] This does require that all computation is executed before any external action is performed, but this constraint does not reduce performance in deterministic environments. See Section 6.2.3 for a discussion of the problem of interleaved computation and action.

down into two components capturing the costs and benefits of computation,

$$\pi^* = \operatorname*{argmax}_{\pi} \mathrm{E}\left[U\big(w, \pi_{\text{act}}(m_N)\big) - \sum_{t=1}^{N-1} \text{cost}(m_t, c_t) \;\middle|\; c_t \sim \pi(m_t) \right]. \tag{2.15}$$

This is the definition of optimal sequential cognitive processes given in Equation 1.17.[6] It emphasizes that the optimal policy is the cognitive strategy that best trades off between the costs and benefits of computation.

The optimal policy for the tallying example is illustrated in Figure 2.3. When the cost of computation is high relative to the stakes of the decision, we see that the optimal tallying policy consider cues until either one option has a lead of two cues in its favor, or thirteen cues have been considered in total. This is interestingly different from the suggestion of Gigerenzer & Gaissmaier (2011), to first consider a fixed number of cues and then make a choice as soon as one leads by any amount. When deliberation is less costly (or the decision is more important), one option has to lead by four cues to be chosen; but over time, this requirement becomes less strict, with a decision being made after considering at most 79 cues. This resembles the "collapsing boundaries" that are often found in optimal evidence accumulation models (Drugowitsch et al., 2012).

Unfortunately, identifying optimal metalevel policies is substantially more challenging than writing down their definition. In Section 2.7, I discuss various strategies for tackling this problem.

2.4 Two views of computation

So far, we have characterized computation as a process of executing mental actions that update an agent's internal state. These mental actions can be contrasted with external actions that update the state of the environment—but they are fundamentally the same type of thing. In the language of Nicholas Hay (2016; Chapter 7), this is the *mechanical view* of computation. This view is non-restrictive; it is relatively easy to see how any algorithm or cognitive model one might concoct could be formalized in this way.

An alternative view, the *Bayesian view*, takes a stronger stance on the nature of computation. It formalizes computations as experiments that generate information about the world (Matheson, 1968). The results of these experiments are synthesized, by Bayesian inference,

[6] Note that the expectation implicitly depends on the transition function, which was explicit in Equation 1.17, as well as the initial mental state m_1 and the set of possible computations.

into posterior distributions over the utility of different possible external actions, thus informing the agent's choice of which action to take.

It is worth emphasizing that this view is quite different from the standard Bayesian approach in cognitive science (e.g., Tenenbaum et al., 2011). While the standard approach treats cognition as a problem of drawing inferences given data, this view treats cognition as a problem of generating the data that drives inference. A closer analogue in cognitive science is found in models of active learning (Gureckis & Markant, 2012; Gottlieb et al., 2013), but in that case, the data is assumed to be found in the world, rather than generated in the mind.

Importantly, the two views are not mutually exclusive. They are often compatible interpretations of a single system. The mechanical model emphasizes the process of computation, while the Bayesian model emphasizes its function. In this way, the mechanical and Bayesian models are analogous to Marr's algorithmic and computational levels. Unlike in Marr's levels, however, adopting the Bayesian model is more than just an interpretation; it has practical consequences for what one can do with the model. Specifically, it puts constraints on the types of computation one can consider. All Bayesian metalevel MDPs have a mechanical interpretation, but not vice versa. On the other hand, adopting the Bayesian view has conceptual and technical advantages. Specifically, it provides a formal link between mental states and world states, and it allows us to convert metalevel MDPs into standard MDPs.

In the following section, I define Bayesian metalevel MDPs, as a special case of the general metalevel MDP framework outlined above.

2.5 Bayesian metalevel MDPs

Taking the Bayesian view amounts to putting a restriction on the set of mental states and the transition function. Specifically, the mental states must correspond to posterior distributions over the state, and the transition function must encode describe a Bayesian updating procedure. I detail these two requirements below.

2.5.1 Belief states

When adopting the Bayesian view, mental states correspond to *beliefs*, formally expressed as distributions over the world state. I will use the notation b_m to denote the Bayesian belief associated with mental state m. When it is clear from context, I will drop the m subscript, e.g., using b_t in place of b_{m_t}.

Mapping mental states to Bayesian beliefs is powerful because it provides a formal link between mental states and world states. For example, we can denote the probability of a world state under a belief state as $b_m(w) = \Pr(W = w \mid M = m)$.[7] Similarly, we can express expectations about functions of world state given the mental state as $\mathrm{E}_{w \sim b_m}[f(w)]$. This allows us to, for example, define an action policy that takes the optimal action given the current mental state,

$$\pi_{\text{act}}(m) = \text{Uniform}\left(\underset{a}{\text{argmax}} \underset{w \sim b_m}{\mathrm{E}}[U(w, a)]\right). \tag{2.16}$$

To be more precise, this policy randomly selects one of the actions that has maximal expected utility given the mental state. This is the default action policy in a Bayesian metalevel MDP, meaning that it is one less choice we have to make as cognitive modelers.

In the Bayesian view, the initial mental state m_1 has a special interpretation. It is the *prior*, the distribution over the world state the agent assumes before performing any computation. By default, we will assume that this prior is accurate, i.e., $b_1(w) = \Pr(W = w)$. However, in Chapter 3, we will need to assume some bias in the prior to fully capture human behavior.

2.5.2 Bayesian updating

In the Bayesian view, computations correspond to experiments that generate information about the world state; this information is then integrated into a belief by Bayesian inference. The transition function describes this process. Formally, each computation defines a state-dependent distribution of outcomes $p_c(o \mid w)$. Given the previous mental state m_t and the outcome o_t, the new mental state m_{t+1} is defined such that

$$b_{t+1}(w) = p(w \mid m_t, o_t) = \frac{p_c(o_t \mid w) b_t(w)}{p(o_t \mid m_t, c_t)}, \tag{2.17}$$

where the second equality is the application of Bayes rule, updating the prior b_t given the likelihood $p_c(o \mid w)$. The transition function describes the full process of sampling an observation and updating the belief accordingly. Denoting the update in Equation 2.17 as "bayes-update", the full transition function is defined as

$$\begin{aligned} o_t &\sim p_c(\cdot \mid w) \\ m_{t+1} &= \text{bayes-update}(m_t, o_t, p_c) \end{aligned} \tag{2.18}$$

[7] For notational convenience, I assume in this chapter that $\mathcal{W}$ is countable. The definitions can easily be extended to the continuous case.

2.5.3 Tallying as a Bayesian metalevel MDP

As emphasized earlier, the Bayesian view is not incompatible with the mechanical view. All Bayesian metalevel MDPs have a mechanical view (or at least, all those that can be implemented on a physical machine). And in some, especially fortuitous cases, one may even find that a model one initially specified in mechanical terms has a Bayesian interpretation as well. As luck would have it, our tallying example is just such a case! Indeed, it is a minor modification of one of the first explicitly formalized metalevel MDPs (Hay et al., 2012).[8]

Viewing tallying from a Bayesian perspective, the mental state corresponds to a distribution over the ratio of cues in favor of each car, w. The standard choice for a distributions over ratios and probabilities is the Beta; thus, we define $b_t = \text{Beta}(\alpha_t, \beta_t)$. The initial values, α_1 and β_1 specify a prior over w. A natural choice is $\alpha_1 = \beta_1 = 1$, which results in a Uniform distribution.

Turning to the transition function, recall that each computation corresponds to considering an additional cue, which may be in favor of one car or the other. We can think of which car the cue favors as the outcome of the computation o. By definition, the probability that each cue favors car X is w, and so we have $o_t \sim \text{Bernoulli}(w)$. Finally, we must specify how these observations are integrated into the belief state ("bayes-update" in Equation 2.18). Because the Beta distribution is the conjugate prior for the Bernoulli distribution, this update has a very simple form:

$$\text{bayes-update}(m_t, o_t, p_c) = \begin{cases} \text{Beta}(\alpha_t + 1, \beta_t) & \text{if } o_t = 1 \\ \text{Beta}(\alpha_t, \beta_t + 1) & \text{if } o_t = 0 \end{cases} \tag{2.19}$$

Not only is this form simple, it is remarkably similar to Equation 2.9. Indeed, given that $p(o_t = 1) = w$, they are equivalent. The only difference between α_t and x_t (and between β_t and y_t) is that the x and y are assumed to be initialized to 0, while α and β are initialized to 1 (or some other values to capture a non-uniform prior). That is, $b_t = \text{Beta}(\alpha_1 + x_t, \beta_1 + y_t)$, providing a direct link between the Bayesian belief state and the mechanical mental state.

2.5.4 Relation to partially observable Markov decision processes

For those who are familiar with the reinforcement learning and planning literatures, the equations above will likely look familiar. Specifically, they closely resemble the equations as-

[8] Specifically, it corresponds to the one-armed Bernoulli metalevel probability model. The modification is in the termination reward. We assume that the agent can choose between two options with value w and $1 - w$, whereas Hay et al. assume a choice between w and a known constant value.

sociated with belief updating in a partially observable Markov decision process (POMDP; Kaelbling et al., 1998). POMDPs are generalizations of MDPs where the agent does not know the state, but instead receives an observation conditional on the state and action at each time step: $o_t \sim O(s_{t+1}, a_t)$. Given these observations, the agent maintains a belief about the current state using a Bayesian update similar to Equation 2.17, but additionally accounting for the possibility that the state changes,

$$b_{t+1}(s_{t+1}) = p(s_{t+1} \mid b_t, o_t) = \frac{O(o_t \mid s_{t+1}, a_t) \sum_{s_t \in \mathcal{S}} T(s_{t+1} \mid s_t, a_t) b_t(s_t)}{p(o_t \mid b_t, a_t).} \tag{2.20}$$

Bayesian metalevel MDPs can thus be understood as a subset of POMDPs in which the state never changes. In this case, the belief update reduces to

$$b_{t+1}(s) = \frac{O(o_t \mid s, a_t) b_t(s)}{p(o_t \mid b_t, a_t)}, \tag{2.21}$$

which is exactly analogous to Equation 2.17. Metalevel MDPs additionally require that all but one action yield strictly negative reward and that the remaining action, $\perp$, leads to a terminal state.

Given that Bayesian metalevel MDPs are a special case of POMDPs, one could reasonably ask why we should bother with metalevel MDPs at all. Indeed, as discussed in Section 6.1.3, other researchers have proposed using POMDPs to model an agent's internal environment. However, there are conceptual and technical advantages to using the more restrictive formalism of metalevel MDPs. Conceptually, it is useful to formally distinguish between mental states and world states, and between computational actions and external actions. Although it is sometimes natural to map the belief state in a POMDP to a mental state, a mental state cannot always be reduced to a belief (i.e., when we are not taking a strictly Bayesian view of computation). From a technical perspective, POMDPs describe a very general and challenging class of problems; as a result, general-purpose POMDP solvers are notoriously inefficient.[9] Focusing on the specific case defined by metalevel MDPs allows us to develop more targeted and efficient solution strategies (e.g., Section 2.8).

Nevertheless, noting the formal similarities between Bayesian metalevel MDPs and POMDPs allows us to take advantage of a powerful tool from the POMDP literature. Specifically, we can draw on the the concept of *belief MDPs* (Kaelbling et al., 1998) to convert a Bayesian

[9] "The best thing about POMDPs is that everything's a POMDP. But the worst thing about POMDPs is that everything's a POMDP." (Michael Littman, personal communication).

metalevel MDP into an equivalent MDP, where the world state has been marginalized out. I refer to this as the "marginalized" metalevel MDP.[10]

2.6 Marginalized metalevel MDPs

As observed by Kaelbling et al. (1998), it is possible to convert a POMDP into an MDP by defining modified transition and reward functions that take beliefs (rather than states) as input and return expected transition probabilities and rewards, marginalizing over the state of the world. This is desirable because it allows one to identify the optimal policy for a POMDP using tools developed to solve MDPs. As outlined below, we can apply a similar strategy to metalevel MDPs. Specifically, we define *marginal* versions of the reward and transition functions that do not depend on the world state. Together with the set of mental states $\mathcal{M}$ and computational "actions" $\mathcal{C}$, this yields a standard MDP. I define the marginalized transition and reward functions below.

2.6.1 Marginal transition function

The marginal transition function $T : \mathcal{M} \times \mathcal{C} \to \Delta(\mathcal{M})$ is most easily defined in generative form:

$$\begin{aligned} w &\sim b_t \\ m_{t+1} &\sim T(m_t, c_t, w). \end{aligned} \tag{2.22}$$

One simply samples the world state from the belief before applying the transition dynamics. This is sufficient to sample from $T(m, c)$. However, we will often need an explicit probability mass function. This is defined as

$$T(m_{t+1} \mid m_t, c_t) = \mathop{\mathrm{E}}_{w \sim b_t} \left[T(m_{t+1} \mid m_t, c_t, w) \right]. \tag{2.23}$$

This expression cannot be simplified in the general case. In practice, we will work with conjugate or discrete beliefs that make this integration tractable. The general strategy is to derive a posterior predictive distribution over the observation, $p(o_t \mid c, m)$, which can be transformed into the transition function by applying bayes-update to the support of this distribution.

In the tallying example, the relevant posterior predictive is given by a Beta-Bernoulli pro-

[10] I don't just call them "belief metalevel MDPs" because the state still corresponds to a mental state, which can function as more than just a belief state (see Section 2.6.3).

cess; the marginal transition is thus defined as

$$T_{\text{tally}}(m_{t+1} \mid m_t, c_t) = \begin{cases} \frac{\alpha_t}{\alpha_t+\beta_t} & \text{if } m_{t+1} = (x_t + 1, y_t) \\ \frac{\beta_t}{\alpha_t+\beta_t} & \text{if } m_{t+1} = (x_t, y_t + 1). \end{cases} \tag{2.24}$$

2.6.2 Marginal reward function

The marginal reward function $R : \mathcal{M} \times \mathcal{C} \to \mathbb{R}$ is defined as

$$R(m, c) = \mathop{\mathrm{E}}_{w \sim b_m} [R(m, c, w)] . \tag{2.25}$$

For $c \neq \perp$, $R(m, c, w)$ does not depend on w; we thus have:

$$R(m, c) = -\text{cost}(m, c). \tag{2.26}$$

The reward for terminating, however, may depend on the state of the world; we must marginalize it out. This results in:

$$R(m, \perp) = \max_a \mathop{\mathrm{E}}_{w \sim b_m} [U(w, a)] . \tag{2.27}$$

That is, the marginal termination reward is simply the maximal expected utility of any action. Although this may seem counter-intuitive, there is a simple intuition: if you select an action that has maximal expected utility, the expected utility of the chosen action will be maximal. A skeptical reader might ask whether this is an equivocation: the first use of "expected" refers to our subjective predictions, while the latter refers to the actual, objective outcome. However, because we assume beliefs to be accurate (Equation 2.30), the subjective and objective expectations are one and the same. This can be seen in the following derivation:

$$\begin{aligned} R(m, \perp) &= \mathop{\mathrm{E}}_{w \sim b_m} [R(m, \perp, w)] \\ &= \mathop{\mathrm{E}}_{w \sim b_m} [U(w, \pi_{\text{act}}(m))] \\ &= \mathop{\mathrm{E}}_{w \sim b_m} \left[U\left(w, \operatorname*{argmax}_a \mathop{\mathrm{E}}_{w' \sim b_m} [U(w', a)] \right) \right] \\ &= \max_a \mathop{\mathrm{E}}_{w \sim b_m} [U(w, a)] . \end{aligned} \tag{2.28}$$

The final line follows from $f\left(\operatorname{argmax}_a f(a)\right) = \max_a f(a)$, where $f(a) = \mathrm{E}_{w \sim b_m} [U(w, a)]$. Note that the logic generalizes to the case where π_{act} samples from the set of optimal actions.

In the tallying example, the marginal termination reward can be derived as

$$\begin{aligned} R_{\text{tally}}(m, \perp) &= \max_a \mathop{\mathrm{E}}_{w \sim b_m} [U(w, a)] \\ &= \max \left\{ \mathop{\mathrm{E}}_{w \sim b_m} [U(w, X)] , \mathop{\mathrm{E}}_{w \sim b_m} [U(w, Y)] \right\} \\ &= \max \left\{ \mathop{\mathrm{E}}_{w \sim b_m} [w] , \mathop{\mathrm{E}}_{w \sim b_m} [1 - w] \right\} \\ &= \max \left\{ \frac{\alpha}{\alpha + \beta}, \frac{\beta}{\alpha + \beta} \right\}, \end{aligned} \tag{2.29}$$

where $\frac{\alpha}{\alpha+\beta}$ is the posterior mean of the Beta distribution b_m.

2.6.3 Marginalizing mechanical metalevel MDPs

Being able to marginalize a metalevel MDP model is very desirable, as it allows us to use general MDP-solving techniques to identify optimal metalevel policies (Section 2.7). But if marginalization is only possible for Bayesian metalevel MDPs, this would limit our ability to use the more general mechanical definition (Section 2.2). Fortunately, marginalization is sometimes still possible when we are not taking the Bayesian view. Inspecting Equations 2.23 and 2.27, we see that computing the marginalized transition and reward functions simply requires taking an expectation over the world state with respect to the belief state. Thus, as long as we can define a belief state associated with any mental state, we can apply the marginalization. However, for the marginalization to result in a valid MDP that is truly equivalent to the original metalevel MDP, the belief state must satisfy two conditions: it must be *accurate* and *complete*.

The accuracy requirement is relatively straightforward. It simply states that the probability the belief state b_m assigns to a world state w is the actual probability that the world is in state w given that you arrived in mental state m. Formally, accuracy requires that

$$b_t(w) = p(w \mid m_t) = \Pr(W = w \mid M_t = m_t). \tag{2.30}$$

It is easy to see why this property is necessary. If one computes the marginal transition and reward functions given incorrect assumptions about the world state, those functions will also be incorrect. In principle, the accuracy requirement does not impose any restrictions on the metalevel MDP. It can always be satisfied by defining the belief state using Bayes'

rule,

$$b_t(w) = \frac{p(m_t \mid w)p(w)}{p(m_t)}. \tag{2.31}$$

In practice, however, one must specify the mental state in such a way that these probabilities can be computed very quickly.

The completeness requirement is more nuanced, and restrictive. It requires that the belief state contain all the information about the world state that it could possibly have, given the full history of the episode up to that time point. Formally, completeness requires that

$$b_t(w) = p(w \mid \mathbf{m}_{1:t}, \mathbf{c}_{1:t}). \tag{2.32}$$

Combining Equations 2.30 and 2.32, we can restate the completeness requirement purely in terms of the mental state:

$$p(w \mid m_t) = p(w \mid \mathbf{m}_{1:t}, \mathbf{c}_{1:t}). \tag{2.33}$$

This imposes a true constraint on the metalevel MDPs we can marginalize. Specifically, Equation 2.32 says that the mental state must be a *sufficient statistic* for the full history of mental states and computations with respect to the world state. To see why this property is necessary, note that Markov decision processes must satisfy the Markov property. That is, the probability of the next state must depend only on the current state and action; formally,

$$p(m_{t+1} \mid m_t, c_t) = p(m_t \mid \mathbf{m}_{1:t}, \mathbf{c}_{1:t}). \tag{2.34}$$

We can then show that this property implies Equation 2.33. Making explicit the marginalization over w on each side of the equation, we have

$$\sum_w p(m_{t+1} \mid m_t, c_t, w)p(w \mid m_t, c_t) = \sum_w p(m_{t+1} \mid \mathbf{m}_{1:t}, \mathbf{c}_{1:t}, w)p(w \mid \mathbf{m}_{1:t}, \mathbf{c}_{1:t}). \tag{2.35}$$

Next we note that $p(m_{t+1} \mid \mathbf{m}_{1:t}, \mathbf{c}_{1:t}, w) = p(m_{t+1} \mid m_t, c_t, w) = T(m_{t+1} \mid m_t, c_t, w)$ by Equation 2.7. That is, the full (non-marginalized) transition function satisfies the Markov property. This gives us

$$\sum_w T(m_{t+1} \mid m_t, c_t, w)p(w \mid m_t, c_t) = \sum_w T(m_{t+1} \mid m_t, c_t, w)p(w \mid \mathbf{m}_{1:t}, \mathbf{c}_{1:t}). \tag{2.36}$$

And from this it is clear that

$$p(w \mid m_t, c_t) = p(w \mid \mathbf{m}_{1:t}, \mathbf{c}_{1:t}) \tag{2.37}$$

Finally, noting that $p(w \mid m_t, c_t) = p(w \mid m_t)$, we arrive at Equation 2.33. Thus, the Markov property (Equation 2.34) implies the sufficiency of the mental state (Equation 2.33), meaning that we cannot have the Markov property without ensuring that the mental state is a sufficient statistic. This in turn means that the marginalized metalevel MDP can only be a *Markov* decision process if the mental state is a sufficient statistic.

2.7 Identifying good metalevel policies

Here I discuss a few general methods for identifying optimal (or at least reasonable) policies for metalevel MDPs. Note that understanding the material in this section is not important for understanding the results presented in the following chapters, and so the reader is encouraged to skip this section if they are not interested in the solution strategies themselves.

Following Equation 2.5, the optimal metalevel policy can be expressed as

$$\pi^*(m) = \text{Uniform}\left(\operatorname*{argmax}_b Q^*(m, c)\right). \tag{2.38}$$

Each of the methods below provides a different way to compute or approximate Q^*.

Historical note: The value of computation

Historically (e.g. Russell & Wefald, 1991a), rational metareasoning has been defined in terms of a slightly different quantity, the *value of computation* (VOC). The VOC is exactly what it sounds like; it specifies the value of performing a computation, where "value" refers to the long-term value in the same sense as the action value function, Q^*. However, the VOC specifically refers to the *increase* in reward one would gain by computing instead of deciding immediately. That is,

$$\text{VOC}(m, c) = Q^*(m, c) - R(m, \bot). \tag{2.39}$$

One advantage of this formulation is that we can define the optimal termination rule as executing $\bot$ whenever no computation has positive VOC. However, the Q function will be more familiar to most researchers, and is easier to work with in practice. For this reason, I will use the Q function throughout. Note, however, that the only difference between the two functions is constant with respect to c (i.e., $-R(m, \bot)$). Thus, maximizing either will yield the optimal policy (we can replace Q^* with VOC in Equation 2.38).

2.7.1 Backward induction

For metalevel MDPs with sufficiently small state spaces, the most robust and accurate method for identifying an optimal policy is backward induction, a form of dynamic programming. See Puterman (2014) for a general overview of these methods. Here, I provide a brief introduction and a few practical suggestions for applying this approach to metalevel MDPs.

Backward induction is a method for computing the optimal value functions, Q^* and V^*, of an MDP. It is based on recursive definitions of the optimal value functions:

$$\begin{aligned} Q^*(s,a) &= R(s,a) + \mathop{\mathrm{E}}_{s' \sim T(s,a)} \left[V^*(s') \right]. \\ V^*(s) &= \max_a Q^*(s,a) \end{aligned} \tag{2.40}$$

These are referred to as *Bellman equations*. Backward induction is an especially simple and efficient application of the Bellman equations for MDPs that are finite and acyclic—that is, there are a finite number of states and one cannot visit the same state twice within a single episode. In such cases, we can assume (without loss of generality) that there is a single absorbing terminal state $s_\perp$ whose value is $V^*(s_\perp) = 0$, by definition. Because there are a finite number of states and no state can be reached from itself, any invocation of Q^* or V^* must eventually hit $V^*(s_\perp) = 0$, the base case.

In a metalevel MDP, it is more natural to define the base case with $Q^*(m, \perp) = R(m, \perp)$.[11] The value functions can then be defined

$$\begin{aligned} Q^*(m,c) &= \begin{cases} R(m,\perp) & \text{if } c = \perp \\ \mathrm{E}_{m' \sim T(m,c)} \left[V^*(m') \right] - \mathrm{cost}(m,c) & \text{otherwise} \end{cases} \\ V^*(m) &= \max_a Q^*(m,c) \end{aligned} \tag{2.41}$$

These equations can be directly implemented as a recursive program, as illustrated in Listing 1. Note, however, that a naive implementation may compute the value of single mental state many times if it can be reached in multiple ways. To prevent this, we *memoize* the value function: this means that if the function is called twice with the same argument, it will return the result from the first call, rather than recomputing it. The combination of memoization and recursion is the defining feature of a "top-down" implementation of backward induction. It ensures that V^* is called on each mental state exactly once (assuming that all

[11] This simply brings the base case up one level in the recursion, as executing $\perp$ always results in the terminal state.

```
function Q(m::MentalState, c::Computation)
    if c == TERM
        term_reward(m) # R(m, ⊥)
    else
        # E_{m'∼T(m,c)} [V*(m')] − cost(m, c)
        sum(p * V(m') for (p, m') in transition(m, c)) - cost(m, c)
    end
end

@memoize function V(m::MentalState)
    # max_{c∈C(m)} Q(m, c)
    maximum(Q(m, c) for c in computations(m))
end
```

Listing 1: Recursive implementation of backward induction in Julia.

mental states are reachable from the initial mental state).

Implementation concerns

Here I address a few practical concerns that arise when applying backward induction to metalevel MDPs.

- Backward induction only applies for discrete state spaces. If the mental state space is continuous (and low-dimensional), one can discretize it. That is, we divide the space into evenly sized bins and create one state for the center of each bin. The discretized transition function is identical to the original transition function except that it "rounds" the generated mental state to the nearest discretized state. To compute an explicit probability mass function, one must integrate over each bin to compute the probability of transitioning to the corresponding state.

- The state space must be finite in addition to discrete. In many cases, however, the natural state space will be unbounded. To address this, we can impose a bound on the number of computations that can be performed (ideally, this bound will be also imposed on the experimental task one is modeling, e.g. by a time limit). The bound is implemented by adding a timestep to the mental state and removing all computations except $\bot$ from the set of possible computations in mental states with maximal timestep. Note that adding the timestep counter also ensures that the MDP is acyclic.

- The state spaces of metalevel MDPs often have symmetry structure that reduces the effective size of the state space. For example, in choice tasks, the order of the items is irrelevant. One way to implement this is to define a hash function that returns the

same value for mental states that are functionally identical (specifically, that must have the same $V^*(s)$). This hash function can be used for memoization such that the value of a set of functionally identical states is only computed once (with the memoized value returned for all others). In Chapter 5, we use a hash function that processes a decision tree recursively, using a commutative operation (summation) to combine the keys for subtrees whose roots are siblings.

- Although the "top-down" implementation of backward induction (Listing 1) is simple, perhaps even beautiful, it is not the most efficient. When speed is a concern—and it almost always is—a "bottom-up" implementation is usually preferred. Such an implementation explicitly iterates over the state space, beginning with all states at the maximal time step and proceeding backwards. See Puterman (2014) for further details on the algorithm.

While backward induction can identify an exact optimal policy (or approximate it to arbitrary precision), it is only tractable when the (discretized) mental state space is small enough that one can iterate over every state in a reasonable amount of time. When the state space is too large, one must turn to approximate solutions. I discuss several possibilities in the following sections.

2.7.2 The myopic policy

In their pioneering work on metareasoning Russell & Wefald (1991a) suggested an approximation to rational metareasoning by one-step look-ahead, what they called the metalevel greedy approximation. This myopic (or "meta-greedy") policy can be defined

$$\pi^{\text{myopic}}(m) = \text{Uniform}\left(\underset{b}{\text{argmax}}\, Q^{\text{myopic}}(m, c)\right), \tag{2.42}$$

where

$$Q^{\text{myopic}}(m, c) = \underset{m' \sim T(m,c)}{\text{E}}\left[R(m', \bot)\right] - \text{cost}(m, c) \tag{2.43}$$

with $Q^{\text{myopic}}(m, \bot) = R(m', \bot)$. The myopic action value function Q^{myopic} gives the expected termination reward after performing one more computation, less the cost of that computation. Thus, the myopic policy selects each computation as if it will be the last one executed.

The myopic policy is often a decent approximation, and displays reasonable behavior in many cases. The problem with it is that it systematically underestimates the value of computation, which leads it to stop computing too early. We can see this easily in the tallying example. If one has the mental state $(3, 1)$, then considering one additional cue cannot change

the decision one would make. The next mental state will be either $(4, 1)$ or $(3, 2)$, and car X will be in the lead in either case. Clearly there is no benefit to considering a single additional cue, so the myopic policy will terminate. However, given the opportunity to consider several more cues, the tide could easily be swayed in favor of the other car. If computation were not very costly, it would likely be worth computing more.

2.7.3 Multi-step lookahead

In their analysis of economic information seeking, Gabaix & Laibson (2005) took note of the myopic policy's struggle with premature termination. They proposed a possible solution in their *directed cognition* model. This model can be understood as a metalevel policy that selects *sequences* of computations rather than computations (exactly like options in hierarchical reinforcement learning; Sutton et al., 1999). Thus, if any sequence of computations has a better expected value than terminating, the policy will continue to compute. In the example above, a sequence would be simply multiple executions of the single computation. A sequence of three computations could result in $m = (3, 4)$, a mental state in which the agent would switch to selecting a_2. Thus, this sequence of computations could have higher expected value than terminating immediately.

When applying this strategy, one need not (and generally should not) actually commit to taking the full sequence of computations that previously had maximal value. This is because the outcome of the first computation may make a different computation more valuable. Note, however, that this results in a dissociation between the agent's implicit assumptions when selecting computations (that they will execute all the computations in the sequence) and reality (that they can choose not to complete the sequence). More precisely, the agent is assuming that they will have less control over future computations than they actually will. This is precisely the same situation as the myopic policy was in (assuming that it would always terminate on the next step), but less severe.

Hay et al. (2012) proposed another solution to the early stopping problem, the *blinkered approximation*. Like directed cognition, the blinkered approximation engages in a sort of lookahead that reduces the flexibility in choosing future computations. Unlike directed cognition, however, the blinkered approximation takes into account the fact that it can choose to adjust its plan based on the outcome of each computation. Glossing over details, the value of a computation is approximated by its value in a smaller metalevel MDP that includes only the computations that reason about the expected return of the corresponding external action. In this way, the solution to a large metalevel MDP is approximated by the composition of solutions to many smaller metalevel MDPs.

Unlike directed cognition, the blinkered approximation is strictly better than the myopic policy. However it can only be applied in cases where a small number of computations are relevant to each action. And even then, it cannot account for the synergistic value of learning about multiple external actions.

All of the strategies I have discussed so far are *model-based*. That is, they estimate the value of computations by simulating their possible outcomes. This can be contrasted with *model-free* strategies that learn policies by trial and error. In the following section, I describe an algorithm that my colleagues and I developed, which combines model-based and model-free reasoning to quickly identify high-performing metalevel policies (Callaway et al., 2018).

2.8 Bayesian metalevel policy search

Bayesian metalevel policy search, or BMPS, is an algorithm that learns policies for metalevel MDPs. At its core, BMPS is a reinforcement learning (RL) algorithm. And in principle, any reinforcement-learning method could be applied to learn metalevel policies. However, metalevel MDPs pose an especially challenging type of problem for typical RL algorithms. In particular, metalevel MDPs present an extreme form of the *credit assignment problem*. In each episode, the agent takes many computations, but receives only a single external reward. If the agent receives a large termination reward, it is unclear which of the many executed computations were important for producing that good outcome and which could have been skipped. This makes it hard to learn which computations are worth performing. To make matters worse, the termination reward often depends greatly on factors outside of the agent's control. If one is choosing between many bad options, the agent cannot get a good termination reward, not matter how well they compute. Together, these factors make metalevel reinforcement learning very challenging.

The BMPS algorithm attempts to make the learning problem easier by endowing the agent with rich knowledge about the structure of the problem. To do so, it draws on work in rational metareasoning aiming to quantify and understand the value of computation (Matheson, 1968; Horvitz, 1987; Russell & Wefald, 1991a). Specially, BMPS draws on work quantifying the *value of information* generated by computation.

2.8.1 The value of information

The value of information (VOI) is defined as the expected utility of a decision you could make based on that information (versus making the decision without that information). For

information we already have, this is simply the termination reward. Thus, we can write

$$\text{VOI}(b) = R(b, \perp) = \max_a \mathop{\text{E}}_{w \sim b} [U(w, a)] , \tag{2.44}$$

using the termination reward defined in Equation 2.27. Note that VOI can only be used when taking the Bayesian view of computation; therefore, I will not distinguish between mental states and beliefs, using b for both.

VOI is most useful for quantifying the value of information that *we don't have yet*. More precisely, VOI quantifies the expected value of a belief state we will have after gathering more information. Given a distribution of possible future belief states, B', the VOI is defined as

$$\text{VOI}(B') = \mathop{\text{E}}_{b' \sim B'} [R(b', \perp)] \tag{2.45}$$

Using this notation, we can express the Q^* function as

$$Q^*(b_t, c_t) = \text{VOI}(B_N) - \text{E}\left[\sum_{i=t}^{N} \text{cost}(b_i, c_i)\right] , \tag{2.46}$$

where B_N is the distribution of terminal beliefs (when following the optimal policy). The challenge of course is that we do not know what this distribution is. However, we can put bounds on it.

2.8.2 Bounding the value of information

As illustrated in Figure 2.4, the value of information acquired by an optimal metalevel policy increases monotonically with the number of computations it performs. Moreover, it is bounded by two values, each giving the VOI associated with a distribution of possible beliefs.

The minimum VOI is the value of information produced by the first computation. In this case, the distribution of beliefs is simply $B' = T(b, c)$. Thus, the myopic value of information is defined as

$$\text{VOI}_{\text{myopic}}(b, c) = \text{VOI}(T(b, c)) = \mathop{\text{E}}_{b' \sim T(b,c)} [R(b', \perp)] .$$

The maximum VOI is the value of full, or "perfect" information (Howard, 1966). In this case, each of the possible future beliefs assigns all probability to a single world state; denote such a belief as $b^*_w = \text{Uniform}(\{w\})$. The distribution over these beliefs, B^*_b is defined $\Pr(B^*_b = b^*_w) = b(w)$. Intuitively, the subjective probability of coming to the belief that

assigns all probability to state w is equal to the subjective probability that w is in fact the true world state. We can then define the value of full information as

$$\mathrm{VOI}_{\mathrm{full}}(b) = \mathrm{VOI}(B_b^*) = \mathop{\mathrm{E}}_{w\sim b}\left[R(b_w^*, \perp)\right] \tag{2.47}$$

2.8.3 Learning to select computations

Given that $\mathrm{VOI}_{\mathrm{myopic}}(b, c)$ and $\mathrm{VOI}_{\mathrm{full}}(b)$ provide lower and upper bounds on the value of information, it follows that the true optimal value of information ($\mathrm{VOI}(B_N)$ in Equation 2.46) is an interpolation between these two values. This suggests the following approximation,

$$Q^{\mathrm{bmps}}(b, c; \mathbf{w}) = w_1 \mathrm{VOI}_{\mathrm{myopic}}(b, c) + w_2 \mathrm{VOI}_{\mathrm{full}}(b) - (\mathrm{cost}(b, c) + w_{\mathrm{cost}}) \tag{2.48}$$

where $w_1 + w_2 = 1$ are the interpolation weights and $w_{\mathrm{cost}} \geq 0$ captures the expected cost of future computations. In specific domains, we may be able to identify additional VOI features that quantify the value of acquiring intermediate amount of information. For example, in Chapter 3 we will use a VOI feature for learning the exact value of a single item in

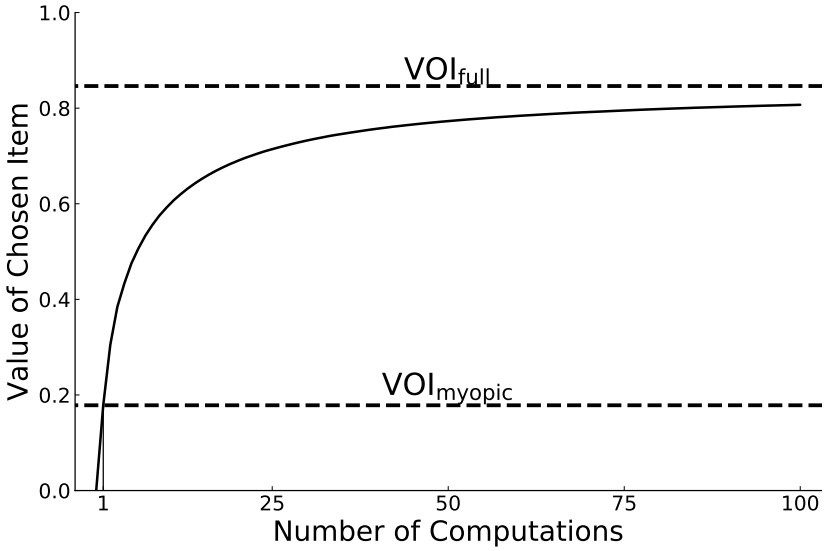

Figure 2.4: Illustration of the value of information features. The solid line shows the average value of the item chosen after different numbers of computations selected by a near-optimal policy for the metalevel MDP defined in Section 3.1.1, assuming no computational costs. The dashed lines show values for two of the VOI features in the initial belief state: $\mathrm{VOI}_{\mathrm{myopic}}$ is the value after one computation and $\mathrm{VOI}_{\mathrm{full}}$ is the asymptotic value after infinite computations.

the choice set. One can add as many such terms as one likes, adding an additional weight for each, and ensuring that they all sum to one. Even with additional features, the approximation makes the very strong assumption that the interpolation weights and expected future cost are constant across belief states, an assumption that will almost certainly not hold. However, it turns out that this rough approximation is sufficient to produce near-optimal metalevel policies in some cases. This policy is defined as

$$\pi^{\text{bmps}}(s; \mathbf{w}) = \text{Uniform}\left(\operatorname*{argmax}_a Q^{\text{bmps}}(s, a; \mathbf{w})\right). \tag{2.49}$$

How should the weights be determined? Treating the approximation literally, the most appropriate strategy would be Q-learning, which tries to find weights that predict that actual returns of the policy. However, because the approximation is so rough, this turns out not to work well in practice. Instead, we can treat the weights as arbitrary parameters of a policy and use policy search to find weights that maximize performance. That is

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}} \mathrm{E}\left[\sum_{t=1}^{N} R(m_t, c_t, w) \;\middle|\; c_t \sim \pi^{\text{bmps}}(m_t; \mathbf{w})\right] \tag{2.50}$$

The key advantage to BMPS over a more generic RL algorithm is that the policy has a very small number of parameters. This allows us to use an efficient search strategy such as Bayesian optimization. Callaway et al. (2018) present an evaluation of BMPS trained with Bayesian optimization on several benchmark tasks, finding that it outperforms the blinkered policy described above as well as a generic DQN. In a cognitive modeling application, consistency may be more important than efficiency; in these cases, we can use exhaustive search algorithms that produce similar results when run multiple times. I describe such a method in Section A.2.

2.9 Summary

In this chapter, I introduced the formal framework of metalevel Markov decision processes. Below I briefly summarize the key points to take away from this section:

- A Markov decision process (MDP) models a sequential decision problem with a set of states the environment can be in, a set of actions the agent can take, a transition function specifying how actions affect the environment state, and a reward function specifying the immediate utility gained by executing each action in each possible state (Section 2.1).

- A metalevel Markov MDP models the problem posed by an agent's internal cognitive environment. Here, states correspond to mental states, actions correspond to cognitive operations, the transition function specifies how operations update mental state, and the reward function specifies both cognitive cost and also the utility of the external action that is taken as a result of cognition. A metalevel MDP also includes a set of world states that determines both the outcome of cognitive operations and also the utility of external actions. (Section 2.2).
- The optimal policy for a metalevel MDP corresponds to an optimal cognitive process, an optimal strategy for selecting which cognitive operation to perform next given the current mental state (Section 2.3).
- We can adopt two views of computation: a mechanical view that treats mental states and actions as exactly analogous to external states and actions, and a Bayesian view that formalize computations as internal experiments that generate information about the world (Sections 2.4 and 2.5). Adopting the latter view allows us to transform a metalevel MDP into an equivalent "marginalized" MDP (Section 2.6).
- Identifying optimal policies for metalevel MDPs can be challenging due to the very large state and action spaces (Section 2.7). Adopting a combination of model-free and model-based techniques can be an effective way to solve this challenging problem (Section 2.8).

In the following three chapters, I show how my colleagues and I have applied this framework to characterize optimal cognitive processes for attention, memory, and planning. As we shall see, this will highlight many of the ways that human cognitive processes are well adapted to both our internal and external environments, revealing new insights about both how our minds work and why they work that way.

Choice of attention—to pay attention to this and ignore that—is to the inner life what choice of action is to the outer. In both cases, a man is responsible for his choice and must accept the consequences, whatever they may be.

W. H. Auden

3

Attention

Fixation patterns in simple choice reflect optimal information sampling[1]

Consider the problem faced by a diner at a buffet table or a shopper at a supermarket shelf. They are presented with a number of options and must evaluate them until they identify the most desirable one. A central question in psychology and neuroscience is to understand the algorithms, or computational processes, behind these canonical simple choices.

Previous work has established two important features of the processes underlying simple value-based choices. First, choices and reaction times are well explained by information sampling models like the diffusion decision model (DDM; Ratcliff & McKoon, 2008; Ratcliff et al., 2016; Milosavljevic et al., 2010) and the leaky competing accumulator model (Usher & McClelland, 2001, 2004). In these models, individuals are initially uncertain about the desirability of each option, but they receive noisy signals about the options' values that they integrate over time to form more accurate estimates. A central insight of these models is that sampling information about unknown subjective values is a central feature of simple choice. Second, visual attention affects the decision-making process. In particular, items that are fixated longer are more likely to be chosen (Shimojo et al., 2003; Armel et al., 2008; Glaholt & Reingold, 2009; Krajbich et al., 2010; Krajbich & Rangel, 2011; Cavanagh et al., 2014; Tavares et al., 2017; Smith & Krajbich, 2019), unless they are aversive, in which case

[1]This chapter is based on the following paper:
Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLOS Computational Biology*, 17(3), e1008863. We thank Ian Krajbich for his help in simulating the aDDM and Bas van Opheusden for suggesting the method for efficiently computing VOI_{full}.

they are chosen *less* frequently (Armel & Rangel, 2008; Armel et al., 2008). These findings have been explained by the Attentional Drift Diffusion Model (aDDM), in which the value samples of the fixated item are over-weighted relative to those of unfixated ones (or equivalently in the binary case, discounting the influence of the unattended item on the drift rate; Krajbich et al., 2010; Krajbich & Rangel, 2011; Smith & Krajbich, 2019; Tavares et al., 2017). See Orquin & Mueller Loose (2013) and Krajbich (2018) for reviews.

These insights raise an important question: What determines what is fixated and when during the decision process? Previous work has focused on two broad classes of theories. One class suggests that decisions and fixations are driven by separate processes, so that fixations affect how information about values is sampled and integrated, but not the other way around. In this view, although fixations can be modulated by features like visual saliency or spatial location, they are assumed to be independent of the state of the decision process. This is the framework behind the aDDM (Krajbich et al., 2010; Krajbich & Rangel, 2011; Tavares et al., 2017) and related models (Gluth et al., 2018; Towal et al., 2013; Thomas et al., 2019).

Another class of theories explores the idea that the decision process affects fixations, especially after some information about the options' values has been accumulated. Examples of this class include the Gaze Cascade Model (Shimojo et al., 2003), an extension of the aDDM in which options with more accumulated evidence in their favor are more likely to be fixated (Gluth et al., 2020), and a Bayesian sampling model in which options with less certain estimates are more likely to be fixated (Song et al., 2019). However, these models have not considered how uncertainty and value might interact, nor have they considered the optimality of the posited fixation process (although see Sepulveda et al., 2020; Moreno-Bote et al., 2020; Ramírez-Ruiz & Moreno-Bote, 2021 for such analyses in simplified settings).

Research on eye movements in the perceptual domain suggests a third possibility: that fixations are deployed to sample information optimally in order to make the best choice. Previous work in vision has shown that fixations are guided to locations that provide useful information for performing a task, and often in ways that are consistent with optimal sampling (Gottlieb & Oudeyer, 2018). For example, in visual search (e.g., finding an 'M' in a field of 'Ns') people fixate on areas most likely to contain the target (Najemnik & Geisler, 2005; Eckstein, 2011); in perceptual discrimination problems, people adapt their relative fixation time to the targets' noise levels (Cassey et al., 2013; Ludwig & Evens, 2017); and in naturalistic task-free viewing, fixations are drawn to areas that have high "Bayesian surprise", i.e., areas where meaningful information is most likely to be found (Itti & Baldi, 2009). The properties of fixations in these types of tasks are captured by optimal sampling models that

maximize expected information gain (Gottlieb et al., 2013; Gottlieb & Oudeyer, 2018). However, these models have not been applied in the context of value-based decision making, and thus the extent to which fixation patterns during simple choices are consistent with optimal information sampling is an open question.

In this chapter, we draw these threads together by defining a model of optimal information sampling in canonical simple choice tasks and investigating the extent to which it accounts for fixation patterns and their relation to choices. In a value-based choice, optimal information sampling requires maximizing the difference between the value of the chosen item and the cost of acquiring the information needed to make the choice. Our model thus falls into a broad class of models that extend classical rational models of economic choice (Savage, 1954; Von Neumann & Morgenstern, 1944) to additionally account for constraints imposed by limited cognitive resources (Lewis et al., 2014; Griffiths et al., 2015; Lieder & Griffiths, 2020; Gershman et al., 2015; Sims, 1998; Caplin & Dean, 2013). However, as is common in this approach, we stop short of specifying a full algorithmic model of simple choice. Instead, we ask to what extent people's fixations are consistent with optimal information sampling, without specifying how the brain actually implements an optimal sampling policy.

Exploring an optimal information sampling model of fixations in simple choice is useful for several reasons. First, since fixations can affect choices, understanding what drives the fixation process can provide critical insight into the sources of mistakes and biases in decision-making. In particular, the extent to which behaviors can be characterized as mistakes depends on the extent to which fixations sample information sub-optimally. Second, simple choice algorithms like the DDM have been shown to implement optimal Bayesian information processing when the decision-maker receives the same amount of information about all options at the same rate (Bogacz et al., 2006; Moreno-Bote, 2010; Drugowitsch et al., 2012; Bitzer et al., 2014; Tajima et al., 2016, 2019; Fudenberg et al., 2018), and this is often viewed as an explanation for why the brain uses these algorithms in the first place. In contrast, the optimal algorithm when the decision-maker must sample information selectively is unknown. Third, given the body of evidence showing that fixations are deployed optimally in perceptual decision making, it is interesting to ask if the same holds for value-based decisions. Given that such problems are characterized by both a different objective function (maximizing a scalar value rather than accuracy) and a different source of information (e.g., sampling from memory Biderman et al., 2020; Bakkour et al., 2019; Wang et al., 2022 rather than from a noisy visual stimulus), it is far from clear that optimal information sampling models will still provide a good account of fixations in this setting.

Building on the previous literature, our model assumes that the decision maker estimates the value of each item in the choice set based on a sequence of noisy samples of the items' true values. We additionally assume that these samples can only be obtained from the attended item, and that it is costly to take samples and to switch fixation locations. This sets up a sequential decision problem: at each moment the decision maker must decide whether to keep sampling, and if so, which item to sample from. Since the model does not have a tractable analytical solution, in order to solve it and take it to the data, we approximate the optimal solution using tools from metareasoning in artificial intelligence (Matheson, 1968; Russell & Wefald, 1991a; Hay et al., 2012; Callaway et al., 2018).

We compare the optimal fixation policy to human fixation patterns in two influential binary and trinary choice datasets (Krajbich et al., 2010; Krajbich & Rangel, 2011). We find that the model captures many previously identified patterns in the fixation data, including the effects of previous fixation time (Song et al., 2019) and item value (Gluth et al., 2018, 2020; Sepulveda et al., 2020). In addition, the model makes several novel predictions about the differences in fixations between binary and trinary choices and about fixation durations, which are consistent with the data. Finally, we identify a critical role of the prior distribution in producing the classic effects of attention on choice (Armel & Rangel, 2008; Armel et al., 2008; Krajbich et al., 2010; Krajbich & Rangel, 2011). Overall, the results show that the fixation process during simple choice is influenced by the value estimates computed during the decision process, in a manner consistent with optimal information sampling.

3.1 Model

We consider simple choice problems in which an agent is presented with a set of items (e.g., snacks) and must choose one. Each item has some value: the utility that the agent would gain by choosing it. However, the agent does not have direct access to these values. Following previous work (Krajbich et al., 2010; Krajbich & Rangel, 2011; Tajima et al., 2016, 2019; Fudenberg et al., 2018), we assume that the agent informs her choice by collecting noisy samples of the items' true values, each providing a small amount of information, but incurring a small cost. The agent integrates the samples into posterior beliefs about each item's value, choosing the item with maximal posterior mean when she terminates the sampling process.

As illustrated in Figure 3.1, we model attention by assuming that the agent can only sample the value of one item at each time point, the item she is fixating on. This sets up a fundamental problem: How should one allocate fixations in order to make good decisions with-

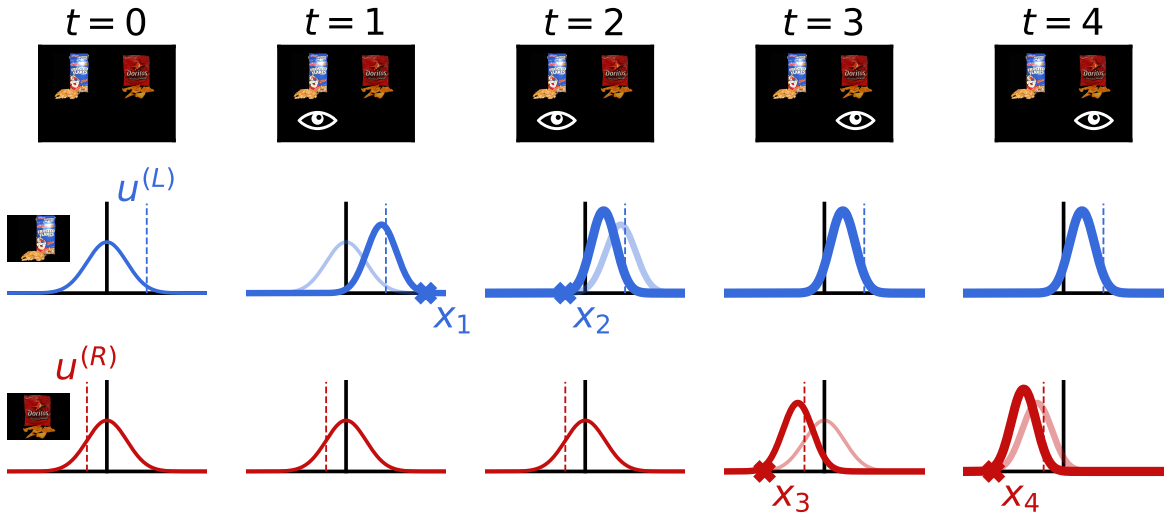

Figure 3.1: Sampling and belief updating in the binary choice task. The top row shows the experimental display, with the fixated item denoted by the eye symbol. The bottom two rows depict the first few steps of the sampling and belief updating process. The decision maker's beliefs about the value of each item are denoted by the Gaussian probability density curves. The true values of each item (dashed lines) are sampled from standard normal distributions; this is captured in the decision maker's initial mental state (first column). Every time step, t, the decision maker fixates one of the items and receives a noisy sample about the true value of that item (x_t marks). She then updates her belief about the value of the fixated item using Bayesian updating (shift from light to dark curve). The beliefs for the unfixated item are not updated. The process repeats each time step until the decision maker terminates sampling, at which point she chooses the item with maximal posterior mean.

out incurring too much cost? Specifically, at each time point, the agent must decide whether to select an option or continue sampling, and in the latter case, she must also decide which item to sample from. Importantly, she cannot simply allocate her attention to the item with the highest true value because she does not know the true values. Rather, she must decide which item to attend to based on her current value estimates and their uncertainty. In the next section, we formalize the model as a metalevel MDP.

3.1.1 Metalevel Markov decision process

To characterize optimal attention allocation, we cast the model as a metalevel MDP in which the mental states correspond to distributions over the value of each item and the computations correspond to fixating on an item and taking a sample of its value. We detail the five components of the metalevel MDP below.

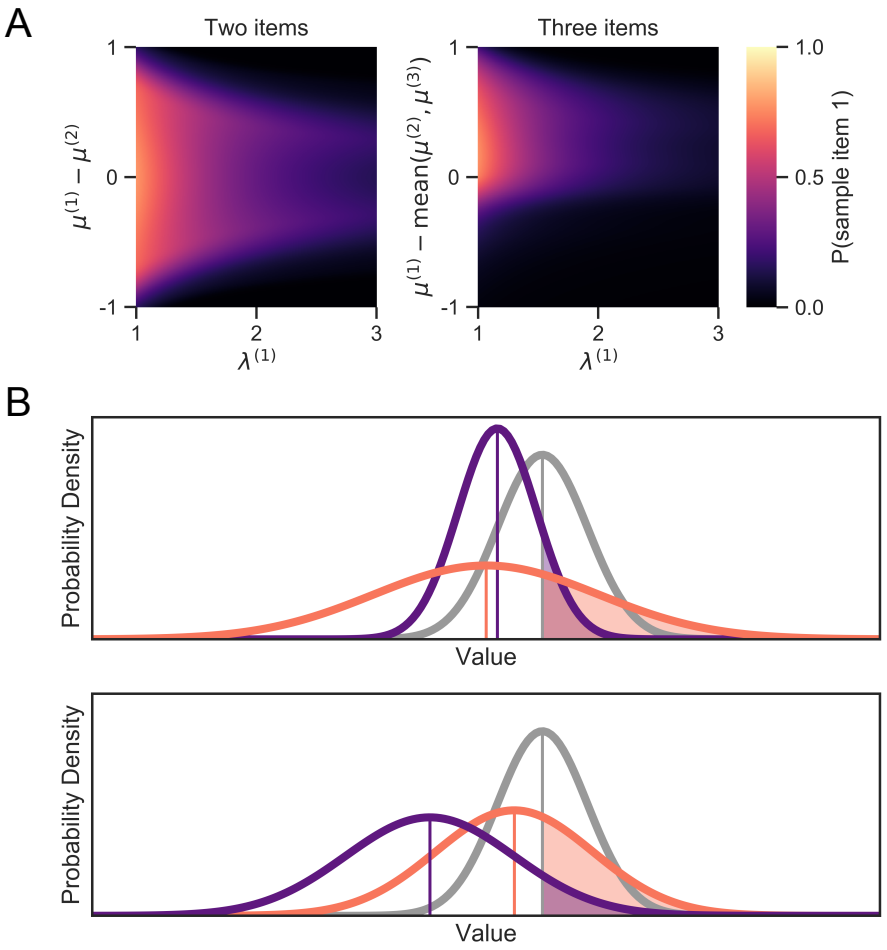

Figure 3.2: Optimal fixation policy. (A) Probability of fixating on item 1 as a function of the precision of its value estimate, $\lambda^{(1)}$, and the mean of its relative value estimate, $\mu^{(1)} - \text{mean}(\mu^{(2)}, \mu^{(3)})$. The heat map denotes the probability of fixating item 1 as opposed to fixating one of the other items or terminating the sampling process. (B) Illustration of the value of sampling. Each panel shows a mental state for trinary choice. The curves depict the posterior distribution over each item's value, and the shaded regions show the probability that the item's true value is higher than the current best value estimate. This probability correlates strongly with the value of sampling the item because sampling is only valuable if it changes the choice (the full value of sampling additionally depends on the size of the potential gain in value, as well as the cost of future samples and the possibility of sampling other items). In each case, it is more valuable to sample the orange item than the purple item because either (top) its value is more uncertain, or (bottom) its value is closer to the leading value.

World states The world state defines the true subjective utility of each item in the choice set. We denote the utility of item i as $u^{(i)}$. The dimensionality of the world state space is equal to the number of items, n.

MENTAL STATES The mental state m corresponds to the agent's belief about the value of each item. Thus, we are fully adopting the Bayesian view of computation (Section 2.4). We assume that the belief distributions are Gaussian, parameterized by mean μ and precision λ (precision is the inverse of variance). The belief at time t about the utility of item i is given by

$$b_t\left(u^{(i)}\right) = \text{Normal}(u^{(i)}; \mu_t^{(i)}, 1/\lambda_t^{(i)}). \tag{3.1}$$

Thus, the mental state specifies a vector of means $\boldsymbol{\mu}_t$ and a vector of precisions $\boldsymbol{\lambda}_t$, with one entry for each item.[2] To model switching costs (see below) we also include the currently fixated item in the mental state, f_t. The mental state at time t is thus defined $m_t = (\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, f_t)$. The dimensionality of the mental state space is $2n + 1$ where n is the number of items.

The initial mental state captures the agent's prior belief about the distribution of values in the environment. We assume that this prior is also Gaussian, with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$ (the role of the prior is discussed in Section 3.1.3). We further assume that no item is fixated at the beginning of the trial. The initial mental state is thus defined $(\boldsymbol{\mu}_0 = \bar{\mu}\mathbf{1}, \boldsymbol{\lambda}_0 = \bar{\sigma}^{-2}\mathbf{1}, f_0 = \oslash)$, where $\oslash$ is a null value.

COMPUTATIONS A computation $c \in \{1, \ldots, n\}$ corresponds to fixating an item, sampling its value, and updating the corresponding estimated value distribution. There are n such computations, one for each item. As in all metalevel MDPs, there is an additional operation $\perp$, which terminates the computation process (here, sampling) and selects an optimal external action given the current mental state (here, choosing the item with maximal posterior mean).

TRANSITION FUNCTION The metalevel transition function specifies the sampling and belief updating process illustrated in Figure 3.1. If the agent selects computation c, she receives a noisy sample of the corresponding item's value,

$$x_t \sim \text{Normal}(u^{(c)}, \sigma_x^2), \tag{3.2}$$

where σ_x^2 is a free parameter specifying the amount of noise in each signal. Note that x_t corresponds to the "outcome" of the computation (o_t in Equation 2.18). The sample is then

[2] Note that it is also possible to represent the mental state as simply the total accumulated evidence and the number of samples taken for each item. We adopt this more mechanical representation in Chapter 4.

integrated into the belief about the item's value by Bayesian inference:

$$\begin{aligned}\lambda_{t+1}^{(c)} &= \lambda_t^{(c)} + \sigma_x^{-2} \\ \mu_{t+1}^{(c)} &= \frac{\sigma_x^{-2} x_t + \lambda_t^{(c)} \mu_t^{(c)}}{\lambda_{t+1}^{(c)}}.\end{aligned} \tag{3.3}$$

Finally, the fixated item is updated ($f_{t+1} = c$), and the estimates for the other items remain unchanged ($\lambda_{t+1}^{(i)} = \lambda_t^{(i)}$ and $\mu_{t+1}^{(i)} = \mu_t^{(i)}$ for $i \neq c$).

The marginal transition function can be specified as the distribution of μ_{t+1}^c given μ_t^c, λ_t^c, and σ_x (all other components of the mental state are updated deterministically). It is given in Equation A.5.

Reward function The reward function describes the cost of sampling and the utility of the chosen item. We assume that there is a fixed cost for each sample taken, as well as an additional switching cost applied when sampling a different item from the last time step. The reward for sampling is thus defined

$$R(m_t, c_t) = -\text{cost}(m_t, c_t) = -(\gamma_{\text{sample}} + 1(c_t \neq f_t)\, \gamma_{\text{switch}}),$$

where γ_{sample} and γ_{switch} are free parameters.[3]

The termination reward is the true value of the chosen item,

$$R(m_t, \perp, w) = u^{(i_t^*)}, \tag{3.4}$$

where the world state w specifies each $u^{(i)}$, and i_t^* is the chosen item (the one with maximal posterior mean):

$$i_t^* = \pi_{\text{act}}(m_t) = \operatorname*{argmax}_i \mu_t^{(i)}. \tag{3.5}$$

The marginal termination reward is simply the maximum posterior mean (see Equation 2.28):

$$R(m_t, \perp) = \mathop{\mathrm{E}}_{w \sim b_t}\left[u^{(i_t^*)}\right] = \max_i \mu_t^{(i)}.$$

[3] We assume that the switch cost is not paid on the first fixation (this is not indicated in the equation for simplicity). However, this assumption has no effect on the optimal policy for reasonable parameter values.

3.1.2 Optimal policy

We assume that the decisions about where to fixate and when to stop sampling are made optimally. That is, we assume that computations are selected by an optimal policy (Equation 2.15), or more precisely, an approximately optimal policy (as described in Section 3.4.1).

What does optimal attention allocation look like? In order to provide an intuitive understanding, we focus on two key properties of mental states: 1) uncertainty about the true values and (2) differences in the value estimates. Figure 3.2A shows the probability of the optimal policy (for a model with parameters fit to human data) sampling an item as a function of these two dimensions (marginalizing over the other dimensions according to their probability of occurring in simulated trials). We see that the optimal policy tends to fixate on items that are uncertain and have estimated values similar to the other items. In the case of trinary—but not binary—choice, we additionally see a stark asymmetry in the effect of relative estimated value. While the policy is likely to sample from an item whose value is substantially higher than the competitors, it is unlikely to sample from an item with value well below. In particular, the policy has a strong preference to sample from the items with best or second-best value estimates.

To see why this is optimal, note that sampling is only valuable insofar as it affects choice, and that the chosen item is the one with maximal estimated value when sampling stops. Thus, the optimal policy generally fixates on the item for which gathering more evidence is most likely to change which item has maximal expected value. There are two ways for this to happen: either the value of the current best item is reduced below the second-best item, or the value of some alternative item is increased above the best item. The former can only happen by sampling the best item, and the latter is most likely to occur by sampling the second-best item because it is closer to the top position than the third-best item (Figure 3.2B bottom). However, if uncertainty is much greater for the third-best item, this can outweigh the larger difference in estimated value (Figure 3.2B top). See Sepulveda et al. (2020) for a more formal justification of value-directed attention in a simplified non-dynamic case.

3.1.3 The prior distribution

Recall that the initial mental state captures the agent's prior belief about the distribution of values in the environment; that is $\mu_0^{(i)} = \bar{\mu}$ and $\lambda_0^{(i)} = \bar{\sigma}^{-2}$. This corresponds to the agent assuming that each item's value is drawn from a prior distribution of true values given by $u^{(i)} \sim \text{Normal}(\bar{\mu}, \bar{\sigma}^2)$. This assumption is plausible if this is the actual distribution of items

that the agent encounters, and she is a Bayesian learner with sufficient experience in the context under study. However, given that these models are typically used to study choices made in the context of an experiment (as we do here), the agent might not have learned the exact prior distribution at work. As a result, we must consider the possibility that she has a *biased prior*.

In order to investigate the role of the prior on the model predictions, we assume that it takes the form of a Gaussian distribution with a mean and standard deviation related to the actual empirical distribution as follows:

$$\begin{aligned} \bar{\mu} &= \alpha \cdot \text{mean}(\text{ratings}) \\ \bar{\sigma} &= \text{std}(\text{ratings}). \end{aligned} \tag{3.6}$$

Here, mean(ratings) denotes the mean value ratings of all items, which provide independent and unbiased measures of the true value of the items (computed across trials in both experiments), and α is a free parameter that specifies the amount of bias in the prior ($\alpha = 0$ corresponds to a strong bias and $\alpha = 1$ corresponds to no bias). As a result, the agent has correct beliefs about the prior variance, but is allowed to have a biased belief about the prior mean. This case could arise, for example, if the average true value of the items used in the experiment differs from the average item that people encounter in their daily lives. This is plausible for the experiments we consider, as items that received negative ratings were excluded from the choice phase (see below).

3.2 Results

We apply the model to two influential simple choice datasets: a binary food choice task (Krajbich et al., 2010) and a trinary food choice task (Krajbich & Rangel, 2011). In each study, participants first provided liking-ratings for 70 snack items on a -10 to 10 scale, which are used as an independent measure of the items' true values. They then made 100 choices among items that they had rated positively, while the location of their fixations was monitored at a rate of 50 Hz. See Appendix A.1 for more details on the experiments.

To compare the model predictions to human fixation behavior, we assume that each sample takes 100ms[4] and that contiguous samples drawn from a single item correspond to a single fixation. We fit the model's five free parameters by maximum likelihood estimation applied to summary statistics for each trial (specifically, we collapse the sequence of fixations

[4]This choice is not important: changing the assumed duration leads to a change in the fitted parameters, but not in the qualitative model predictions.

into proportions of time on each item; see Section 3.4.4 for details). In order to compare the model predictions with the observed patterns out-of-sample, we estimate the parameters using only the even trials, and then simulate the model in odd trials.

Importantly, since the same model can be applied to n-item choices, we fit a common set of parameters jointly to the pooled data in both datasets. Thus, any differences in model predictions between binary and trinary choices are *a priori* predictions resulting from the structure of the model, and not differences in the parameters used to explain the two types of choices.

In order to explore the role of the prior, we also fit versions of the model in which the prior bias term was fixed to $\alpha = 0$ or $\alpha = 1$. The former corresponds to a strongly biased prior and the latter corresponds to a completely unbiased prior.

Because the policy optimization and likelihood estimation methods that we use are stochastic, we display simulations using the 30 top performing parameter configurations to give a sense of the uncertainty in the predictions. All the figures below are based on model fits estimated at the group level on the pooled data. However, for completeness we also fit the model separately for each individual, and report these fits in Appendix A.3. We also describe a validation of our model fitting approach in Appendix A.4.

Basic psychometrics

We begin by looking at basic psychometric patterns. Figure 3.3A compares the choice curves predicted by the model with the actual observed choices, separately for the case of binary and trinary choice. It shows that the model captures well the influence of the items' true values (as measured by liking ratings) on choice.

Figure 3.3B plots the distribution of total fixation times. This measure is similar to reaction time except that it excludes time not spent fixating on one of the items. We use total fixation time instead of reaction time because the model does not account for the initial fixation latency nor the time spent saccading between items (although it does account for the opportunity cost of that time, through the γ_{sample} parameter). As shown in the figure, the model provides a reasonable qualitative account of the distributions, although it underpredicts the mode in the case of two items and the skew in both cases.

Figure 3.3C shows the relationship between total fixation time and trial difficulty, as measured by the relative liking rating of the best item. We find that the model provides a reasonable account of how total fixation time changes with difficulty. This prediction follows from that fact that fewer samples are necessary to detect a large difference than to either detect a small difference or determine that the difference is small enough to be unimportant. How-

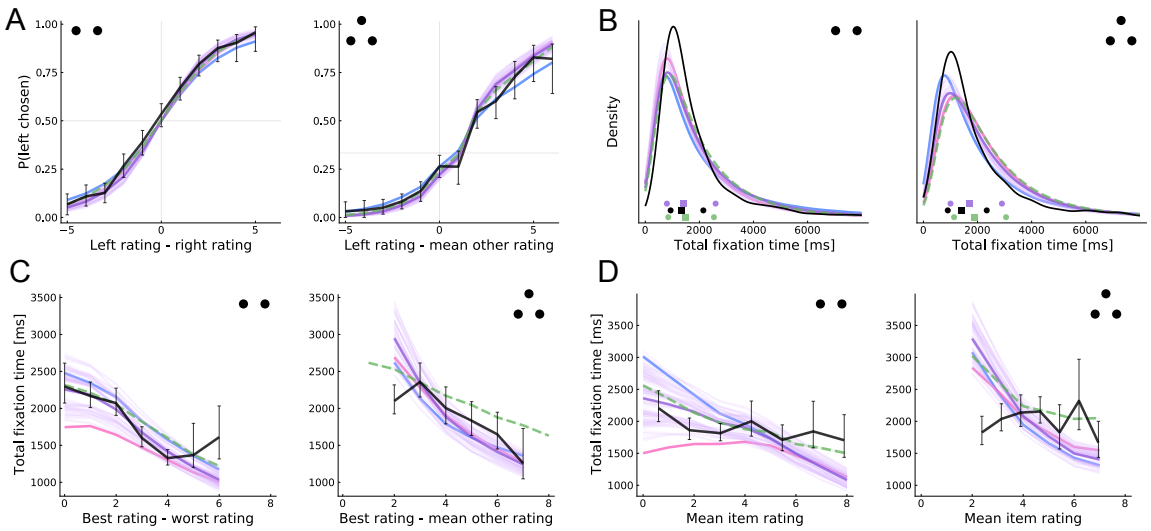

Figure 3.3: Basic psychometrics. Each panel compares human data (black) and model predictions for binary choice (left, two dots) and trinary choice (right, three dots). The main model predictions are shown in purple. The restricted model predictions for the case of a highly biased prior mean ($\alpha = 0$) is shown in blue, and for the case of a highly unbiased prior mean ($\alpha = 1$) are shown in pink. These colors were chosen to illustrate that the main model falls between these two extremes. The aDDM predictions are shown in dashed green. Error bars (human) and shaded regions (model) indicate 95% confidence intervals computed by 10,000 bootstrap samples (the model confidence intervals are often too small to be visible). Note that the method used to compute and estimate the model parameters is noisy. To provide a sense of the effect of this noise on the main model predictions, we depict the predictions of the thirty best-fitting parameter configurations. Each light purple line depicts the predictions for one of those parameters, whereas the darker purple line shows the mean prediction. In order to keep the plot legible, only the mean predictions of the biased priors models are shown). (A) Choice probability as a function of relative rating. (B) Kernel density estimation for the distribution of total fixation time. Quartiles (25%, 50%, and 75% quantiles) for the data, aDDM and main model predictions are shown at the bottom. (C) Total fixation time as a function of the relative rating of the highest rated item. (D) Total fixation time as a function of the mean of all the item ratings (overall value).

ever, the model exhibits considerable variation in the predicted intercept and substantially overpredicts total fixation time in difficult trinary choices.

Finally, Figure 3.3D shows the relationship between total fixation time and the average rating of all the items in the choice set. This "overall value effect" has been emphasized in recent research (Smith & Krajbich, 2019; Krajbich, 2018) because it is consistent with multiplicative attention weighting (as in the aDDM) but not an additive boosting model (e.g., Cavanagh et al., 2014). Bayesian updating results in a form of multiplicative weighting (specifically, a hyperbolic function, c.f. Armel & Rangel, 2008), and thus our model also predicts this pattern. Surprisingly, we do not see strong evidence for the overall value effect

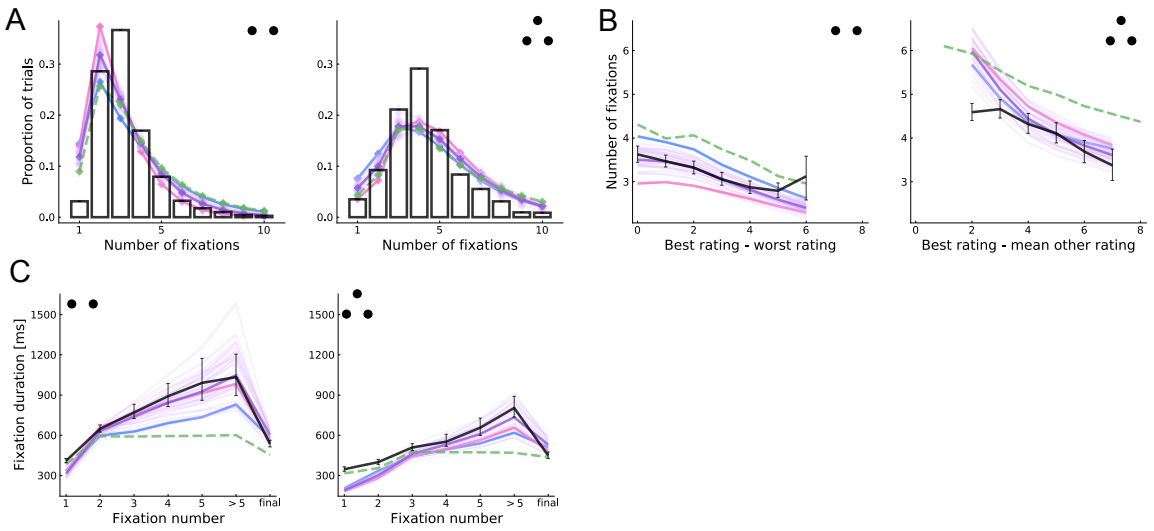

Figure 3.4: Basic fixation patterns. (A) Histogram of number of fixations in a trial. (B) Number of fixations as a function of decision difficulty, as measured by the relative rating of the best item. (C) Duration of fixation by fixation number. Final fixations are excluded from all but the last bin.

in the datasets we consider, but we note that the effect has been found robustly in several other datasets (Smith & Krajbich, 2019; Frömer et al., 2019; Hunt et al., 2012; Polanía et al., 2014; Pirrone et al., 2018). Note that, in the binary case, the predicted overall value effect is symmetric around the prior mean; that is, choices between two very bad items will also be made quickly. Indeed, with an unbiased-prior, the model predicts an inverted-U relationship around the prior mean.

Several additional patterns in Figure 3.3 are worth highlighting. First, all the models make similar and reasonable predictions of the psychometric choice curve and fixation time distributions. Second, the models with some prior bias provide a better account of the fixation time curves in binary choice than the unbiased model, and qualitatively similar predictions to the aDDM. Finally, despite using a common set of parameters, all the models capture well the differences between binary and trinary choice.

Basic fixation properties

We next compare the predicted and observed fixation patterns. An observed "fixation" refers to a contiguous span of time during which a participant looks at the same item. A predicted model fixation refers to a continuous sequence of samples taken from one item.

Figure 3.4A shows the distribution of the number of fixations across trials. The model-predicted distribution is reasonably similar to the observed data. However, in the two-item

case, the model is more likely to make only one fixation, suggesting that people have a tendency to fixate both items at least once that the model does not capture.

Figure 3.4B shows the relationship between the total number of fixations and decision difficulty. We find that the model captures the relationship between difficulty and the number of fixations reasonably well, with the same caveats as for Figure 3.3B.

The original binary and trinary choice papers observed a systematic change in fixation durations over the course of the trial, as shown in Figure 3.4C. Although the model tends to underpredict the duration of the first two fixations in the three-item case, it captures well three key patterns: (a) the final fixation is shorter, (b) later (but non-final) fixations are longer and (c) fixations are substantially longer in the two-item case. The final prediction is especially striking given that the model uses the same set of fitted parameters for both datasets. The model predicts shorter final fixations because they are cut off when a choice is made (Krajbich et al., 2010). The model predicts the other patterns because more evidence is needed to alter beliefs when their precision is already high; this occurs late in the trial, especially in the two-item case where samples are split between fewer items.

Figure 3.4 also shows that the main model provides a more accurate account than the aDDM of how the number of fixations changes with trial difficulty, and of how fixation duration evolves over the course of a trial. One difficulty in making this comparison is that the aDDM assumes that non-final fixation durations are sampled from the observed empirical distribution, conditional on a number of observable variables, and thus the accuracy of its predictions regarding fixation duration and fixation number depends on the details of this sampling. To maximize comparability with the existing literature, here we use the same methods as in the original implementations.

Uncertainty-directed attention

As we have seen, one of the key drivers of fixations in the optimal policy is uncertainty about the items' values. Specifically, because the precision of the posteriors increases linearly with the number of samples, the model predicts that, other things being equal, fixations should go to items that have received less cumulative fixation time. However, the difference in precision must be large enough to justify paying the switching cost. In this section we explore some of the fixation patterns associated with this mechanism.

Figure 3.5A depicts the distribution of relative cumulative fixation time at the beginning of a new fixation, starting with the second fixation. That is, at the onset of each fixation, we ask how much time has already been spent fixating the newly fixated item, compared to the other items. In both cases, the actual and predicted distributions are centered below zero, so

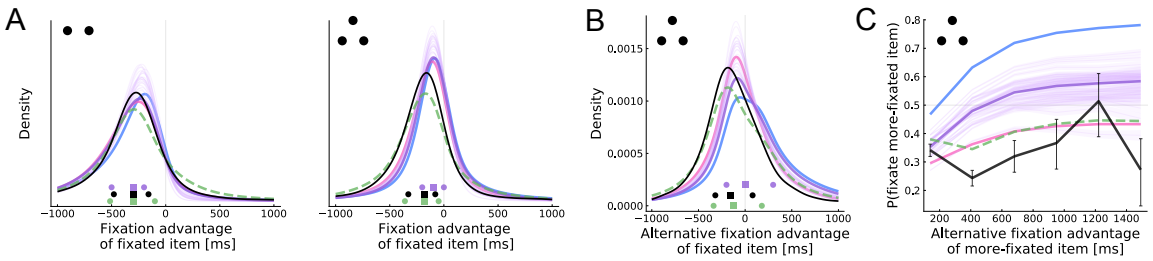

Figure 3.5: Uncertainty-directed attention. (A) Distribution of fixation advantage of the fixated item, computed at the beginning of each new fixation. Fixation advantage is defined as the cumulative fixation time to the item minus the mean cumulative fixation time to the other item(s). First fixations are excluded in this plot. (B) Similar to (A), except that we compare the fixation advantage between the fixated item and the other item that could have been fixated but was not. First and second fixations are excluded in this plot. (C) The probability that the item with greater alternative fixation advantage is fixated, as a function of that advantage.

that items tend to be fixated when they have received less fixation time than the other items. Additionally, the model correctly predicts the lower mode and fatter left tail in the two-item case.

Note, however, that a purely mechanical effect can account for this basic pattern: the item that is currently fixated will on average have received the most fixation time, but it cannot be the target of a new fixation, which drives down the fixation advantage of newly fixated items. For this reason, it is useful to look further at the three-item case, which affords a stronger test of uncertainty-directed attention. In this case, the target of each new fixation (excluding the first) must be one of the two items that are not currently fixated. Thus, comparing the cumulative fixation times for these items avoids the previous confound. Figure 3.5B thus plots the distribution of fixation time for the fixated item minus that of the item which could have been fixated but was not. We see a similar pattern to Figure 3.5A (right) in both the data and model predictions. This suggests that uncertainty is not simply driving the decision to make a saccade, but is also influencing the location of that saccade.

Figure 3.5C explores this further by looking at the location of new fixations in the three-item case, as a function of the difference in cumulative fixation time between the two possible fixation targets. Although the more-previously-fixated item is always less likely to be fixated, the probability of such a fixation actually *increases* as its fixation advantage grows. This counterintuitive model prediction results from the competing effects of value and uncertainty on attention. Since items with high estimated value are fixated more, an item that has been fixated much less than the others is likely to have a lower estimated value, and is therefore less likely to receive more fixations. However, we see that the predicted effect is much

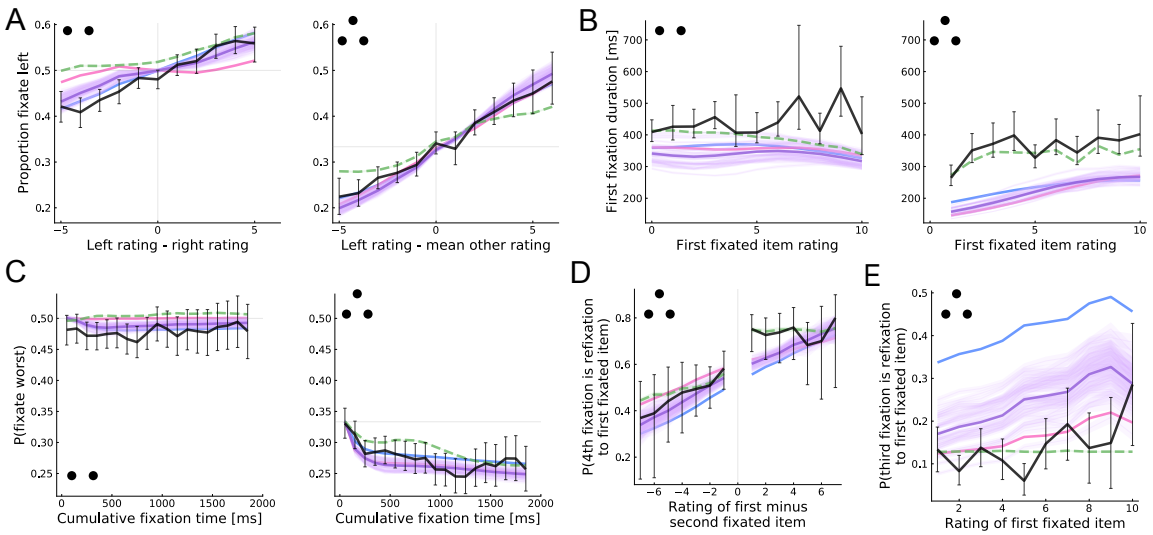

Figure 3.6: Value-directed attention. (A) Proportion of time fixating the left item as a function of its relative rating. (B) First fixation duration as a function of the rating of the first-fixated item. (C) Probability of fixating the lowest rated item as a function of the cumulative fixation time to any of the items. (D) Probability that the fourth fixation is to the first-fixated item as a function of the difference in rating between that item and the second-fixated item. (E) Probability that the third fixation is to the first fixated item as a function of its rating.

stronger than the observed effect, and that the aDDM model provides a better account of this pattern than our main model. However, note that the accuracy of this fit follows from the fact that the aDDM samples fixation locations and durations from the empirical distribution, conditioned on the previous three fixation locations and the item ratings.

Value-directed attention

A second key driver of attention in the optimal policy is estimated value, which directs fixations to the *two* items with the highest posterior means. As illustrated in Figure 3.2A, this implies that fixation locations should be sensitive to relative estimated values in the trinary but not in the binary case.

Although we cannot directly measure the participants' evolving value estimates, we can use the liking ratings as a proxy for them because higher-rated items will tend to result in higher value estimates. Using this idea, Figure 3.6A shows the proportion of fixation time devoted to the left item as a function of its relative rating. Focusing first on the three-item case, both the model and data show a strong tendency to spend more time fixating on higher rated items (which are therefore likely to have higher estimated values). In the two item case, the model simulations show a smaller but also positive effect. This is counterintu-

itive since the model predicts that in the two-item case fixation locations are insensitive to the sign of the relative value estimates (Figure 3.2A). However, the pattern likely arises due to the tendency to fixate last on the chosen item (Figure 3.7A, discussed below).

Figure 3.6B provides an alternative test that avoids confounds associated with the final fixation. It shows the duration of the first fixation, which is rarely final, as a function of the rating of the first fixated item. In the three-item case, both the model and data show longer initial fixations to high-rated items, although the model systematically underpredicts the mean first fixation duration. This prediction follows from the fact that, under the optimal policy, fixations are terminated when the fixated item's estimated value falls out of the top two (below zero for the first fixation); the higher the true value of the item, the less likely this is to happen. In the two-item case, however, the model predicts that first fixation duration should be largely insensitive to estimated value; highly valuable items actually receive slightly *shorter* fixations because these items are more likely to generate extremely positive samples that result in terminating the first fixation and immediately choosing the fixated item. Consistent with this prediction, humans show little evidence for longer first fixations to high-rated items in the binary case.

Previous work has suggested that attention may be directly influenced by the true value of the items (Towal et al., 2013; Anderson, 2016; Gluth et al., 2018). In our model, however, attention is driven only by the internal value estimates generated during the decision making process. To distinguish between these two accounts, we need a way to dissociate estimated value from true value. One way to do this is by looking at the time course of attention. Early in the decision making process, estimated values will be only weakly related to true value. However, with time the value estimates become increasingly accurate and thus more closely correlate with true value. Thus, if the decision maker always attends to the items with high *estimated* value, she should be increasingly likely to attend to items with high *true* value as the trial progresses. Figure 3.6C shows the probability of fixating on the worst item as a function of the cumulative fixation time to any of the items. In both the two- and three-item cases, the probability begins near chance. In the three-item case, however, the probability quickly falls. This is consistent with a model in which attention is driven by estimated value rather than value itself.

The model makes even starker predictions in the three-item case. First, take all trials in which the decision-maker samples from different items during the first three fixations. Consider the choice of where to deploy the fourth fixation. The model predicts that this fixation should be to the first-fixated item if its posterior mean is larger than that of the second-fixated item, and vice versa. As a result, the probability that the fourth fixation is a refixation

to the first-fixated item should increase with the difference in ratings between the first- and second-fixated items. As shown in Figure 3.6D, the observed pattern follows the model prediction.

Finally, the model makes a striking prediction regarding the location of the third fixation in the three-item case. Consider the choice of where to fixate after the first two fixations. The decision maker can choose to fixate on the item that she has not seen yet, or to refixate the first-fixated item. The model predicts a refixation to the first-seen item if both that item and the second-seen item already have high value estimates (leaving the unfixated item with the lowest value estimate). Consistent with this prediction, Figure 3.6E shows that the probability of the third fixation being a refixation to the first-seen item increases with that item's rating. Note that the model with α fixed to zero (corresponding to a strong prior bias), dramatically overpredicts the intercept. This is because this model greatly underestimates the value of the not-yet-fixated item.

Overall, Figure 3.6 shows that our main model provides a better fit to some fixation patterns, whereas the aDDM provides a better fit to others. However, it is important to keep in mind that whereas our model provides predictions for these fixation patterns based on first principles, the predictions of the aDDM for these patterns are largely mechanistic since that model samples fixation locations and durations from the observed empirical distribution. As a result, it is not surprising that Figure 3.6B shows a better match between the aDDM and the data since the "predicted" durations are actually sampled from the observed data conditional on the first item rating.

Choice Biases

Previous work has found a systematic positive correlation between relative fixation time and choice for appetitive (i.e., positively valenced) items (Shimojo et al., 2003; Armel & Rangel, 2008; Armel et al., 2008; Krajbich et al., 2010; Krajbich & Rangel, 2011; Gluth et al., 2020). In particular, models like the aDDM propose that an exogenous or random increase in fixations towards an appetitive item increase the probability that it will be chosen, which leads to attention driven choice biases. Here we investigate whether the optimal model can account for these types of effects.

Importantly, in the type of optimal fixation model proposed here, there are two potential mechanisms through which such correlations can emerge in the optimal model. The first is driven by the prior. If the prior mean is negatively biased, then sampling from an item will on average increase its estimated value. This follows from the fact that sampling will generally move the estimated value towards the item's true value, and a negatively biased

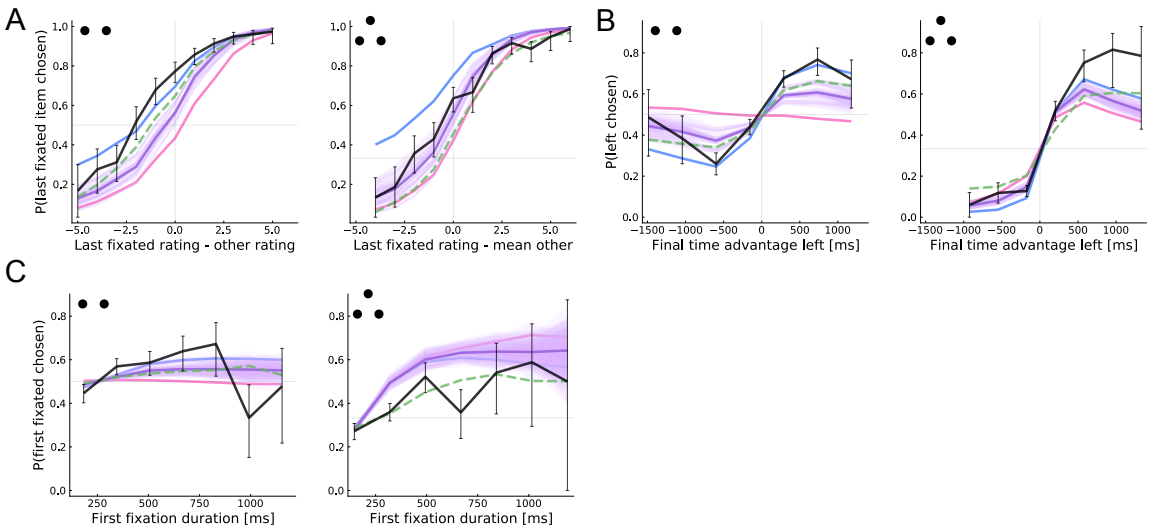

Figure 3.7: Choice biases. (A) Probability that the last fixated item is chosen as a function of its relative rating. (B) Probability that the left item is chosen as a function of its final fixation advantage, given by total fixation time to the left item minus the mean total fixation time to the other item(s). (C) Probability of choosing the first-seen item as a function of the first-fixation duration.

prior implies that the initial value estimate is generally less than the true value. The second mechanism, which is only present in trinary choice, is the result of value-directed attention. Here, the causal direction is flipped, with value estimates driving fixations rather than fixations driving value estimates. In particular, items with higher estimated value are both more likely to be fixated, and more likely to be chosen. Thus, fixations and choice are correlated through a common cause structure. Importantly, the two mechanisms are not mutually exclusive; in fact, our model predicts that both will be in effect for choice between more than two items.

Figure 3.7A shows that there is a sizable choice bias towards the last-seen item in both datasets, as evidenced by the greater-than-chance probability of choosing an item whose value is equal to the mean of the other items. Our model provides a strong quantitative account of the pattern in trinary choice, but substantially underpredicts the effect in binary choice. Interestingly, it predicts a weaker effect than the aDDM in the binary case, but a stronger effect in the trinary case.

To understand this result, it is important to think about the prior beliefs implicit in the aDDM and related models (Krajbich et al., 2010; Krajbich & Rangel, 2011; Gluth et al., 2020). Since these are not Bayesian models, they do not posit an explicit prior that is then modified by evidence. However, the aDDM can be viewed as an approximation to a Bayesian model with a prior centered on zero, as reflected by the initial point of the accumulator

(zero) and the multiplicative discounting (the evidence for the non-attended item is discounted towards zero). The latter roughly corresponds to the Bayesian regularization effect, wherein the posterior mean falls closer to the prior mean when the likelihood is weak (low precision). Given this, our model predicts a weaker effect in the binary case because it has a weaker prior bias ($\alpha = 0.58$) than the one implicit in the aDDM ($\alpha = 0$). Our model predicts a stronger effect in the trinary case due to the value-directed attention mechanism. Critically, although the aDDM accounts for the effect of *true* value on fixations (by sampling from the empirical fixation distribution), only the optimal model accounts for the effects of *estimated* value. Thus, conditioning on true value (as we do in Figure 3.7A) breaks the value-based attention mechanism in the aDDM but not in the optimal model. Finally, note that the optimal model with $\alpha = 0$ provides a good account of the bias in the binary case, but dramatically overpredicts it in the trinary case.

Figure 3.7B shows that the average probability of choosing the left item increases substantially with its overall relative fixation time. As before, in comparison with the aDDM, the optimal model provides better captures the full strength of the bias in the trinary case, but underpredicts the effect in the binary case. The optimal model with α fixed to zero performs best in both cases. Note that the fit of the aDDM is not as close as for similar figures in the original papers because we simulate all models with the observed ratings (rather than all possible combination of item ratings) and we consider a larger range of final time advantage. We replicate the original aDDM figures in Appendix A.5.

Finally, Figure 3.7C shows that the probability of choosing the first fixated item increases with the duration of the first fixation. Importantly, this figure shows that the attention-choice correlation cannot be explained solely by the tendency to choose the last-fixated item. Again, all four models qualitatively capture the effect, with varying degrees of quantitative fit.

3.3 Discussion

We have built a model of optimal information sampling during simple choice in order to investigate the extent to which it can provide a quantitative account of fixation patterns, and their relationship with choices, during binary and trinary decisions. The model is based on previous work showing that simple choices are based on the sequential accumulation of noisy value samples (Ratcliff, 1978; Tajima et al., 2016; Ratcliff & McKoon, 2008; Teodorescu & Usher, 2013; Busemeyer & Townsend, 1993; Holmes et al., 2016) and that the process is modulated by visual attention (Krajbich et al., 2010; Krajbich & Rangel, 2011; Gluth et al.,

2018, 2020; Song et al., 2019; Smith & Krajbich, 2018; Armel et al., 2008). However, instead of proposing a specific algorithmic model of the fixation and choice process, as is common in the literature, our focus has been on characterizing the optimal fixation policy and its implications. We build on previous work on optimal economic decision-making in which samples are acquired for all options at the same rate (Tajima et al., 2016; Fudenberg et al., 2018; Bogacz et al., 2006; Tajima et al., 2019), and extend it to the case of endogenous attention, where the decision maker can control the rate of information acquired about each option. We formalized the selection of fixations as a problem of dynamically allocating a costly cognitive resource in order to gain information about the values of the available options. Leveraging tools from metareasoning in artificial intelligence (Matheson, 1968; Russell & Wefald, 1991a; Hay et al., 2012; Callaway et al., 2018), we approximated the optimal solution to this problem, which takes the form of a policy that selects which item to fixate at each moment and when to terminate the decision-making process.

We found that, despite its simplicity, the optimal model accounts for many key fixation and choice patterns in two influential binary and trinary choice datasets (Krajbich et al., 2010; Krajbich & Rangel, 2011). The model was also able to account for striking differences between the two- and three-item cases using a common set of parameters fitted out of sample. More importantly, the results provide evidence in favor of the hypothesis that the fixation process is influenced by the evolving value estimates, at least to some extent. Consider, for example, the increase in fixation duration over the course of the trial shown in Figure 3.4C, the tendency to equate fixation time across items (Figure 3.5B), and the relationship between the rating of the first fixated item and the probability of re-fixating it (Figures 3.6D and 3.6E). These effects are explained by our model, but are hard to explain with exogenous fixations, or with fixations that are correlated with the true value of the items, but not with the evolving value estimates (e.g., as in Towal et al., 2013; Stojić et al., 2020; Gluth et al., 2018).

Optimal information sampling models may appear inappropriate for value-based decision-making problems, in which perceptual uncertainty about the identity of the different choice items (often highly familiar junk foods) is likely resolved long before a choice is made. Two features of the model ameliorate this concern. First, the samples underlying value-based decisions are not taken from the external display (as in perceptual decisions), but are instead generated internally, perhaps by some combination of mental simulation and memory recall (Biderman et al., 2020; Bakkour et al., 2019; Wang et al., 2022). Second, the model makes the *eye-mind* assumption (Just & Carpenter, 1976; Orquin & Mueller Loose, 2013): what a person is looking at is a good indicator of what they are thinking about. Impor-

tantly, these assumptions implicitly underlie all sequential sampling models of value-based decision-making.

Our model is not the first to propose that the fixation and value-estimation processes might interact reciprocally. However, no previous models fully capture the key characteristics of optimal attention allocation, which appear to be at least approximated in human fixation behavior. For example, the Gaze Cascade Model (Shimojo et al., 2003) proposes that late in a trial subjects lock-in fixations on the favored option until a choice is made, Gluth et al. (2020) propose an aDDM in which the probability of fixating an item is given by a softmax over the estimated values, and Song et al. (2019) propose a Bayesian model of binary choice in which fixations are driven by relative uncertainty. In contrast to these models, the optimal model predicts that fixations are driven by a combination of the estimated uncertainty and relative values throughout the trial, and that attention is devoted specifically to the items with the top two value estimates. Although the data strongly support the first prediction, further data are necessary to distinguish between the top-two rule and the softmax rule of Gluth et al. (2020).

Our results shed further light on the mechanisms underlying the classic attention-choice correlation that has motivated previous models of attention-modulated simple choice. First, our results highlight an important role of prior beliefs in sequential sampling models of simple choice (c.f. Jang et al., 2021). All previous models have assumed a prior mean of zero, either explicitly (Song et al., 2019; Jang et al., 2021) or implicitly (Krajbich et al., 2010; Krajbich & Rangel, 2011; Gluth et al., 2020). Such a prior is negatively biased when all or most items have positive value, as is often the case in experimental settings. This bias is critical in explaining the classic attention-choice correlation effects because it creates a net-positive effect of attention on choice: if one begins with an underestimate, attending to an item will on average increase its estimated value. However, we found that the best characterization of the full behavior was achieved with a moderately biased prior, both in terms of our approximate likelihood and in the full set of behavioral patterns in the plots.

Our results also suggest another (not mutually exclusive) mechanism by which the attention-choice correlation can emerge: value-directed attention. We found that the optimal model with no prior bias ($\alpha = 1$) predicts an attention-choice correlation in the trinary choice case. This is because, controlling for true values, an increase in estimated value (e.g., due to sample noise) makes the model more likely to both fixate and choose an item. This could potentially help to resolve the debate over additive vs. multiplicative effects of attention on choice (Cavanagh et al., 2014; Smith & Krajbich, 2019). While the prior-bias mechanism predicts a multiplicative effect, the value-directed attention mechanism predicts that fixation

time and choice will be directly related (as predicted by the additive model). Although we did not see strong evidence for value-directed attention in the binary dataset, such a bias has been shown in explicit information gathering settings (Hunt et al., 2016) and could be at work in other binary choice settings.

Our work most closely relates to two recent lines of work on optimal information sampling for simple choice. First, Hébert and Woodford (2017; 2019) consider sequential sampling models based on rational inattention. They derive optimal sampling strategies under highly general information-theoretic constraints, and establish several interesting properties of optimal sampling, such as the conditions under which the evidence accumulation will resemble a jump or a diffusion process. In their framework, the decision maker chooses, at each time point, an arbitrary *information structure*, the probability of producing each possible signal under different true states of the world. In contrast, we specify a very small set of information structures, each of which corresponds to sampling a noisy estimate of one item's value (Equation 3.2). This naturally associates each information structure with fixating on one of the items, allowing us to compare model predictions to human fixation patterns. Whether human attention more closely resembles flexible construction of dynamic information structures, or selection from a small set of fixed information structures is an interesting question for future research.

In a second line of work, concurrent to our own, Jang, Sharma, and Drugowitsch (2021) develop a model of optimal information sampling for binary choice with the same Bayesian structure as our model and compare their predictions to human behavior in the same binary choice dataset that we use (Krajbich et al., 2010). There are three important differences between the studies. First, they consider the possibility that samples can also be drawn in parallel for the unattended item, but with higher variance. However, they find that a model in which almost no information is acquired for the unattended item fits the data best, consistent with the assumptions of our model. Second, they use dynamic programming to identify the optimal attention policy almost exactly. This allows them to more accurately characterize truly optimal attention allocation. However, dynamic programming is intractable for more than two items, due to the curse of dimensionality. Thus, they could not consider trinary choice, which is of special interest because only this case makes value-directed attention optimal, and forces the decision-maker to decide which of the unattended items to fixate next, rather than simply when to switch to the other item. Third, they assumed (following previous work) that the prior mean is zero. In contrast, by varying the prior, we show that although a biased prior is needed to account for the attention-choice correlation in binary choice, the data is best explained by a model with only a moderately biased prior

mean, about halfway between zero and the empirical mean.

We can also draw insights from the empirical patterns that the model fails to capture. These mismatches suggest that the model, which was designed to be as simple as possible, is missing critical components that should be explored in future work. For example, the underprediction of fixation durations early in the trial could be addressed by more realistic constraints on the fixation process such as inhibition of return, and the overprediction of the proportion of single-fixation trials in the two-item case could be explained with uncertainty aversion. Although not illustrated here, the model's accuracy could be further improved by including bottom-up influences on fixations (e.g., spatial or saliency biases Towal et al., 2013; Itti & Koch, 2000).

While we have focused on attention in simple choice, other studies have explored the role of attention in more complicated multi-attribute choices (Roe et al., 2001; Noguchi & Stewart, 2018; Russo & Dosher, 1983; Trueblood et al., 2014; Usher & McClelland, 2004; Berkowitsch et al., 2014; Fisher, 2017; Krajbich et al., 2012; Westbrook et al., 2020; Shi et al., 2013; Manohar & Husain, 2013). None of these studies have carried out a full characterization of the optimal sampling process or how it compares to observed fixation patterns, although see Gabaix et al. (2006) and Yang et al. (2015) for some related results. Applying the metalevel MDP framework to this important case is a priority for future work. Finally, in contrast to many sequential sampling models, our model is not intended as a biologically plausible process model of how the brain actually makes decisions. Exploring how the brain might approximate the optimal sampling policy presented here, and also how optimal sampling might change under accumulation mechanisms such as decay and inhibition is another priority for future work.

3.4 Methods

The model was implemented in the Julia programming language (Bezanson et al., 2017). The code can be found at
`https://github.com/fredcallaway/optimal-fixations-simple-choice`.

3.4.1 Approximating the optimal policy

As described in Section 2.3, the solution to a metalevel MDP takes the form of a Markov policy π that stochastically selects which computation to take next given the current mental state. The optimal metalevel policy π^* is the one that maximizes expected total metalevel

reward (Equation 2.1). In our model, we can write this as

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathrm{E}\left[\max_i \mu_N^{(i)} - \sum_t^{N-1} \mathrm{cost}(m_t, c_t) \;\middle|\; c_t \sim \pi(m_t)\right].$$

That is, one wishes to acquire accurate beliefs that support selecting a high-value item, while at the same time minimizing the cost of the samples necessary to attain those beliefs.

How can we identify the optimal policy? For small discrete belief spaces, the optimal metalevel policy can be computed exactly using standard dynamic programming methods such as value iteration or backwards induction. These methods can also be applied to low-dimensional, continuous belief spaces by first discretizing the space on a grid (Tajima et al., 2019), and this approach has recently been used to characterize the optimal fixation policy in binary choice (Jang et al., 2021). Unfortunately, these methods are infeasible in the trinary choice case, since the belief space has six continuous dimensions.

Instead, we approximate the optimal policy using a variant of the BMPS algorithm, described in Section 2.8. We assume that computations are selected according to a softmax policy on the BMPS approximation to the optimal action-value function,

$$\pi(c \mid m) \propto \exp\left\{\beta \cdot Q^{\mathrm{bmps}}(m, c)\right\}, \tag{3.7}$$

where β is a free parameter and

$$Q^{\mathrm{bmps}}(m, c) = w_1 \mathrm{VOI}_{\mathrm{myopic}}(m, c) + w_2 \mathrm{VOI}_{\mathrm{item}}(m, c) + w_3 \mathrm{VOI}_{\mathrm{full}}(m) - (\mathrm{cost}(c) + w_{\mathrm{cost}}). \tag{3.8}$$

In the present model, $\mathrm{VOI}_{\mathrm{myopic}}(m, c)$ denotes the expected improvement in choice utility from drawing one additional sample from item c before making a choice, as opposed to making a choice immediately based on the current belief, m. $\mathrm{VOI}_{\mathrm{item}}(m, c)$ denotes the expected improvement from learning the true value of item c, and then choosing the best item based on that information. Finally, $\mathrm{VOI}_{\mathrm{full}}(m)$ denotes the improvement from learning the true value of every item and then making an optimal choice based on that complete information. We provide full derivations of these features in Appendix A.6.

In the original implementation of BMPS (Callaway et al. (2018)), we use Bayesian optimization to identify the weights within this space that maximize total expected metalevel reward. However, for the present model, we found that often a large area of weight space resulted in extremely similar performance, despite inducing behaviorally distinct policies. Practically, this makes identifying a unique optimal policy challenging, and theoretically

we would not expect all participants to follow a single unique policy when there is a wide plateau of high-performing policies. To address this, we instead identify a set of 80 near-optimal policies and assume that human behavior will conform to the aggregate behavior of this set. To identify this set of near-optimal policies, we apply a method based on Upper Confidence Bound (UCB) bandit algorithms (Auer et al., 2002), described in Appendix A.2.

How good is the approximation method? Callaway et al. (2018) showed that this approach generates near-optimal policies on a related problem, with Bernoulli-distributed samples and no switching costs. Note that in the case of Bernoulli samples, the belief space is discrete and thus the optimal policy can be computed exactly if an upper bound is placed on the number of computations that can be performed before making a decision. Although introducing switching costs makes the metareasoning problem more challenging to solve, in the Bernoulli case we have found that they only induce a modest reduction in the performance of the approximation method relative to the full optimal policy, achieving 92% of optimal reward in the worst case (see Appendix A.7 for details). This suggests that this method is likely to provide a reasonable approximation to the optimal policy in the model with Gaussian samples used here, but a full verification of this fact is beyond the scope of the current study.

3.4.2 Implementation of the prior

In the main text, we specified the prior as a property of the initial mental state. However, for technical reasons (in particular, to reuse the same set of optimized policies for multiple values of α), it is preferable to perform policy optimization and simulation in a standardized space, in which the initial mental state has $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\lambda}_0 = 1$. We then capture the prior over the ratings of items in the experiment by transforming the ratings into this standardized space such that the transformed values are in units defined by the prior. Concretely, given an item rating $r^{(i)}$, we set the true value to

$$u^{(i)} = \frac{r^{(i)} - \bar{\mu}}{\bar{\sigma}}, \tag{3.9}$$

where $\bar{\mu}$ and $\bar{\sigma}$ denote the prior mean and standard deviation. Modulo the resultant change in units (all parameter values are divided by $\bar{\sigma}$), this produces the exact same behavior as the naïve implementation, in which the initial mental state itself varies.

There is one non-trivial consequence of using this approach when jointly fitting multiple datasets: The jointly fit parameters are estimated in the standardized space, rather than the space defined by the raw rating scale. As a result, if we transform the parameters back

into the raw rating space, the parameters will be slightly different for the two datasets (even though they are identical in the transformed space). This was done intentionally because we expect that the parameters will be consistent in the context-independent units (i.e., standard deviations of an internal utility scale). However, this decision turns out to have negligible impact in our case because the empirical rating distributions are very similar. Specifically, the empirical rating distributions are (mean $\pm$ std) 3.492 ± 2.631 for the binary dataset and 4.295 ± 2.524 for the trinary dataset. Due to the difference in standard deviations, all parameters (except α, which is not affected) are $2.631/2.524 = 1.042$ times larger in the raw rating space for the binary dataset compared to the trinary dataset. The difference in empirical means affects $\bar{\mu}$, which is $3.492/4.295 = 0.813$ times as large in the binary compared to trinary dataset. However, given our interpretation of α as a degree of updating towards the empirical mean, this difference is as intended.

3.4.3 Model simulation procedure

Given a metalevel MDP and policy, π, simulating a choice trial amounts to running a single episode of the policy on the metalevel MDP. To run an episode, we first initialize the mental state, $m_0 = (\boldsymbol{\mu}_0 = 0, \boldsymbol{\lambda}_0 = 1, f_0 = \oslash)$. The agent then selects an initial computation $c_0 \sim \pi(m_0)$ and the mental state is updated according to the transition dynamics (Equation 3.3). Note that $\pi(c \mid m_0)$ assigns equal sampling probability to all of the items, since the subject starts with symmetrical beliefs. This process repeats until some time step, N, when the agent selects the termination action, $\perp$. The predicted choice is the item with maximal posterior value, $i_N^* = \text{argmax}_i \mu_N^{(i)}$. In the event of a tie, the choice is sampled uniformly from the set of items with maximal expected value given the current mental state. Because the samples are continuous, this can only happen when multiple items have not been sampled yet (still having the prior mean). In practice, this never happens with well-fitting parameter values because the policy never terminates sampling in this case.

To translate the sequence of computations into a fixation sequence, we assume that each sample takes 100ms and concatenate multiple contiguous samples from the same item into one fixation. The temporal duration of a sample is arbitrary; a lower value would result in finer temporal predictions, but longer runtime when simulating the model. In this way, it is very similar to the *dt* parameter used in simulating diffusion decision models. Importantly the qualitative predictions of the model are insensitive to this parameter because σ_x and γ_{sample} can be adjusted to result in the same amount of information and cost per ms.

We simulate the model for two different purposes: (1) identifying the optimal policy and (2) comparing model predictions to human behavior. In the former case, we randomly

sample the true utilities on each "trial" i.i.d. from Normal$(0, 1)$. This corresponds to the assumption that the fixation policy is optimized for an environment in which the agent's prior is accurate. When simulating a specific trial for comparison to human behavior, the true value of each item is instead determined by the liking ratings for the items presented on that trial, as specified in Equation 3.9.

3.4.4 Model parameter estimation

The model has five free parameters: the standard deviation of the sampling distribution, σ_x, the cost per sample, γ_{sample}, the cost of switching attention, γ_{switch}, the degree of prior updating, α, and the inverse temperature of the Boltzmann policy, β. We estimate a single set of parameters at the group level using approximate maximum likelihood estimation in the combined two- and three-item datasets, using only the even trials.

To briefly summarize the estimation procedure: given a candidate set of parameter values, we construct the corresponding metalevel MDP and identify a set of 80 near-optimal policies for that MDP. We then approximate the likelihood of the human fixation and choice data using simulations from the optimized policies. Finally, we perform this full procedure for 70,000 quasi-randomly sampled parameter configurations and report the top thirty configurations (those with the highest likelihood) to give a rough sense of the uncertainty in the model predictions. A parameter recovery exercise (reported in Appendix A.4) suggests that this method, though approximate, is sufficient to identify the parameters of the model with fairly high accuracy. Below, we explain in detail how we estimate and then maximize the approximate likelihood.

The primary challenge in fitting the model is in estimating the likelihood function. In principle, we could seek to maximize the joint likelihood of the observed fixation sequences and choices. However, like most sequential sampling models, our model does not have an analytic likelihood function. Additionally, the high dimensionality of the fixation data makes standard methods for approximating the likelihood (Turner & Sederberg, 2014; van Opheusden et al., 2020) infeasible. Thus, taking inspiration from Approximate Bayesian Computation methods (Sunnåker et al., 2013; Csilléry et al., 2010), we approximate the likelihood by collapsing the high dimensional fixation data into four summary statistics: the identity of the chosen item, the number of fixations, the total fixation time, and the proportion of fixation time on each item. As described below, we estimate the joint likelihood of these summary statistics as a smoothed histogram of the statistics in simulated trials, and then approximate the likelihood of a trial by the likelihood of its summary statistics. We emphasize, however, that we do not use this approximate likelihood to evaluate the per-

formance of the model. Instead, we intend it to be a maximally principled (and minimally researcher-specified) approach to choosing model parameters, given that computing a true likelihood is computationally infeasible.

Given a set of near-optimal policies, we estimate the likelihood of the summary statistics for each trial using a smoothed histogram of the summary statistics in simulated trials. Critically, this likelihood is conditional on the ratings for the item in that trial. However, it depends only on the (unordered) set of these ratings; thus, we estimate the conditional likelihood once for each such set. Given a set of ratings, we simulate the model 625 times for each of the 80 policies, using the resulting 50,000 simulations to construct a histogram of the trial summary statistics. The continuous statistics (total and proportion fixation times) are binned into quintiles (i.e., five bins containing equal amounts of the data) defined by the distribution in the experimental data. For the fixation proportions, the quintiles are defined on the rating rank of the item rather than the spatial location because we expect the distributions to depend on relative rating in the three-item case. Values outside the experimental range are placed into the corresponding tail bin. Similarly, trials with five or more fixations are all grouped into one bin (including e.g., six and seven fixations) and cases in which the model predicts zero fixations are grouped into the one-fixation bin. This latter case corresponds to choosing an item immediately without ever sampling, and occurs rarely in well-fitting instantiations of the model, but happens frequently when γ_{sample} is set too high. For each simulation, we compute the binned summary statistics, identify the corresponding cell in the histogram, and increase its count by one. Finally, we normalize this histogram, resulting in a likelihood over the summary statistics. To compute the likelihood of a trial, $\mathcal{L}(d \mid \theta)$, we compute the binned summary statistics for the trial and look up the corresponding value in the normalized histogram for that trial's rating set.

To account for trials that are not well explained by our model, we use add-n smoothing, where n was chosen independently for each θ to maximize the likelihood. This is equivalent to assuming a mixture between the empirical distribution and a uniform distribution with mixing weight ε. Thus, the full approximate likelihood is

$$\mathcal{L}(D \mid \theta) = \max_{\varepsilon \in [0,0.5]} \prod_{d \in D} \left(\varepsilon \frac{1}{C} + (1 - \varepsilon)\, \mathcal{L}(d \mid \theta) \right),$$

where $C = n \cdot 5^{n+1}$ is the total number of cells in the histogram. Importantly, this error model is only used to approximate the likelihood; it is not used for generating the model predictions in the figures—indeed, it could not be used in this way because the error model is defined over the summary statistics, and cannot generate full sequences of fixations. Thus,

the ε parameter should be interpreted in roughly the same way as the bandwidth parameter of a kernel density estimate (Turner & Sederberg, 2014), rather than as an additional free parameter of the model.

We then use this approximate likelihood function to identify a maximum likelihood estimate, $\hat{\theta} = \text{argmax}\, \mathcal{L}(D \mid \theta)$. Based on manual inspection, we identified the promising region of parameter space to be $\sigma_x \in (1, 5)$, $\gamma_{\text{sample}} \in (0.001, 0.01)$, $\gamma_{\text{switch}} \in (0.003, 0.03)$, and $\beta \in (100, 500)$. We then ran an additional quasi-random search of 10,000 points within this space using Sobol low-discrepancy sequences (Sobol, 1967). This approach has been shown to be more effective than both grid search and random search, while still allowing for massive parallelization (Bergstra & Bengio, 2012).

Note that the optimal policy does not depend on α because the agent believes her prior to be unbiased (by definition) and makes her fixation decisions accordingly. The alternative, optimizing the policy conditional on α, would imply that the agent is internally inconsistent, accounting for the bias in her fixations but not in the prior itself. Thus, we optimize α separately from the other parameters. Specifically, we consider 10,000 possible instantiations of all the other parameters, find optimal policies once for each instantiation, and evaluate the likelihood for seven values of α; these seven values included the special cases of 0 and 1 as well as five additional randomly-spaced values with a random offset (roughly capturing the low-discrepancy property of the Sobol sequence).

We found that the stochasticity in the policy optimization and likelihood estimation coupled with weak identifiability for some parameters resulted in slightly different results when re-running the full procedure; thus, to give a rough sense of the uncertainty in the estimate, we identify the top thirty parameters, giving us both mean and standard deviation for each parameter and the total likelihood.

The parameter estimates for the main model were (mean $\pm$ std) $\sigma_x = 2.6 \pm 0.216$, $\alpha = 0.581 \pm 0.118$, $\gamma_{\text{switch}} = 0.00995 \pm 0.001$, $\gamma_{\text{sample}} = 0.00373 \pm 0.001$, and $\beta = 364.0 \pm 81.2$. The units of these parameter estimates are standard deviations of value (i.e., $\bar{\sigma}$). For the model with $\alpha = 0$, the fitted parameters were $\sigma_x = 3.16 \pm 0.409$, $\gamma_{\text{switch}} = 0.00875 \pm 0.002$, $\gamma_{\text{sample}} = 0.00319 \pm 0.001$, and $\beta = 326.0 \pm 81.2$. And for the model with $\alpha = 1$, they were $\sigma_x = 2.66 \pm 0.272$, $\gamma_{\text{switch}} = 0.0118 \pm 0.002$, $\gamma_{\text{sample}} = 0.00506 \pm 0.001$, and $\beta = 330.0 \pm 97.9$.

Suppose we try to recall a forgotten name. The state of our consciousness is peculiar. There is a gap therein; but no mere gap. It is a gap that is intensely active. A sort of wraith of the name is in it, beckoning us in a given direction, making us at moments tingle with the sense of our closeness, and then letting us sink back without the longed-for term.

William James

4

Memory

Optimal metalevel control of memory recall[1]

Most of us have experienced moments when we could not recall some piece of information but felt that we knew it (feeling of knowing; Hart, 1965), perhaps even sensing that the answer was imminent and only momentarily blocked (tip-of-tongue; Brown & McNeill, 1966). These processes whereby people can examine and make judgments about the content of memory have been termed "metamemory." Different from memory itself, metamemory refers to the higher order processes that monitor and control basic memory processes (Nelson & Narens, 1990). In this chapter, we aim to characterize the functional role of these processes in supporting rapid memory recall.

Most empirical work in metamemory has focused on how people are able to monitor their memory states (Reder & Ritter, 1992; Miner & Reder, 1994; Eakin, 2005) and on the accuracy of metamemory judgments in predicting future recall (Hart, 1965; Vesonder & Voss, 1985; Dunlosky & Nelson, 1992; Dunlosky & Lipko, 2007). Recently, these phenomena have been understood through computational models of signal detection (Jang et al., 2012) and probability theory (Hu, 2021). Less emphasis, however, has been placed on understanding the function of metamemory judgments (Schwartz & Metcalfe, 2017). In a highly influential paper, Nelson & Narens (1990) proposed that the function of metacognitive systems is to allow effective control of ongoing cognition (Figure 4.1). For example, they outlined a theory in which a dynamically updated feeling of knowing is used to inform the decision of

[1]This chapter is based on a working paper, with authors: Callaway, F., Griffiths, T. L., Norman, K. A., & Zhang Q. This research was supported by a grant from Facebook Reality Labs.

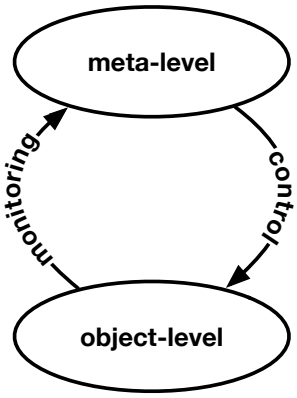

Figure 4.1: Illustration of Nelson and Narens' theoretical framework for metamemory. A "metalevel" process monitors and controls the performance of a basic "object-level" process. Adapted from Nelson & Narens (1990).

when to terminate an unsuccessful recall attempt (Figure 5 in Nelson & Narens, 1990). However, despite this early progress, there is (to our knowledge) still no computational model of how these feeling of knowing estimates might be dynamically generated, nor of how they could be used to control recall efforts. Consequently, despite intuitively suggestive findings such as longer search times for items with high feeling of knowing (Nelson, 1984; Nhouyvanisvong & Reder, 1998; Gruneberg et al., 1977; Lachman et al., 1979), it is unclear to what extent metamemory serves an adaptive function in people.

We believe two challenges have hindered progress on developing computationally explicit theories of metacognitive control of memory recall. First, on the empirical front, commonly-used metamemory paradigms rely on self-report as the primary evidence of people's metamemory. However, the subjective nature of these reports makes it difficult to evaluate the objective utility of metamemory in guiding recall, as we seek to do. Moreover, because the judgments are most often made after retrieval is completed (or abandoned), the causal relationship between metamemory judgments and memory search behavior is unclear (Schwartz, 2001). For example, it is possible that participants report strong feeling of knowing because they spent a long time searching, rather than vice-versa. Indeed, in perceptual decision-making, manipulating response time (while holding accuracy constant) affects confidence judgments (Kiani et al., 2014). On the other hand, rapid feeling-of-knowing judgments made before recall (e.g. Reder, 1987) cannot capture knowledge that only becomes available in the course of recall (Koriat, 1993; Nhouyvanisvong & Reder, 1998). To address this challenge, we developed a metamemory paradigm that allows us to establish a quantitative, objective measurement of memory strength before retrieval. An extension of this paradigm in Experiment 2 additionally allows us to see behavioral signatures of metacognitive control even before retrieval is completed or abandoned, revealing how the dynamic metamemory process unfolds over time. In this way, we can directly test our

model's core predictions about how people will direct their recall efforts depending on the strength of the to-be-recalled memories.

The second challenge is a technical one. In many domains of cognitive science, theoretical progress has been spurred by the development of rational models that optimally solve the problem that the cognitive system is theorized to solve (Savage, 1954; Tenenbaum & Griffiths, 2001; Anderson, 1991; Knill & Richards, 1996; Marr, 1982). Indeed, Anderson & Milson (1989) famously applied this approach to shed light on basic properties of human memory. Metamemory, however, poses an especially thorny type of optimization problem, as it involves a cyclic, "closed-loop" interaction between two cognitive processes (Figure 4.1). It is not obvious how one should quantify the performance of such a system, let alone identify a system that maximizes this performance. To address this challenge, we draw on formal tools developed for metalevel control in artificial intelligence (Russell & Wefald, 1991a; Hay, 2016). These tools have recently been applied to model dynamic metacognitive processes in decision-making contexts, revealing that people's behavior is remarkably consistent with models that optimally trade off utility with cognitive cost (Callaway et al., 2021, 2022b; c.f. Drugowitsch et al., 2012; Tajima et al., 2019; Jang et al., 2021; Chen et al., 2021). By applying these tools to a simple model of memory recall, we can make concrete predictions about the behavior we would expect to see if people can adaptively control their memory processes.

The remainder of this chapter is organized as follows. We begin by reviewing empirical work on metalevel control of memory, focusing on the control of recall. Then, we define an optimal model of metalevel control in memory recall and characterize its predictions. Notably, the model predicts that unsuccessful memory searches will be longer when the target memory is (judged to be) stronger, consistent with the findings of Costermans et al. (1992). Next, we describe a cued-recall experiment that conceptually replicates and extends those findings. We confirm all key qualitative predictions of the model and establish moderate quantitative fit. Our second experiment extends the first by allowing participants to choose between two possible recall targets. This introduces a more complex metalevel control problem of selecting which memory to search for at each moment. Using a keypress-contingent display, we compare the timecourse of attention to each cue with the optimal model's search predictions and again achieve a strong qualitative and moderate quantitative fit. We conclude by discussing implications of the results for metamemory and metacognition research more generally, and identifying interesting directions for further research.

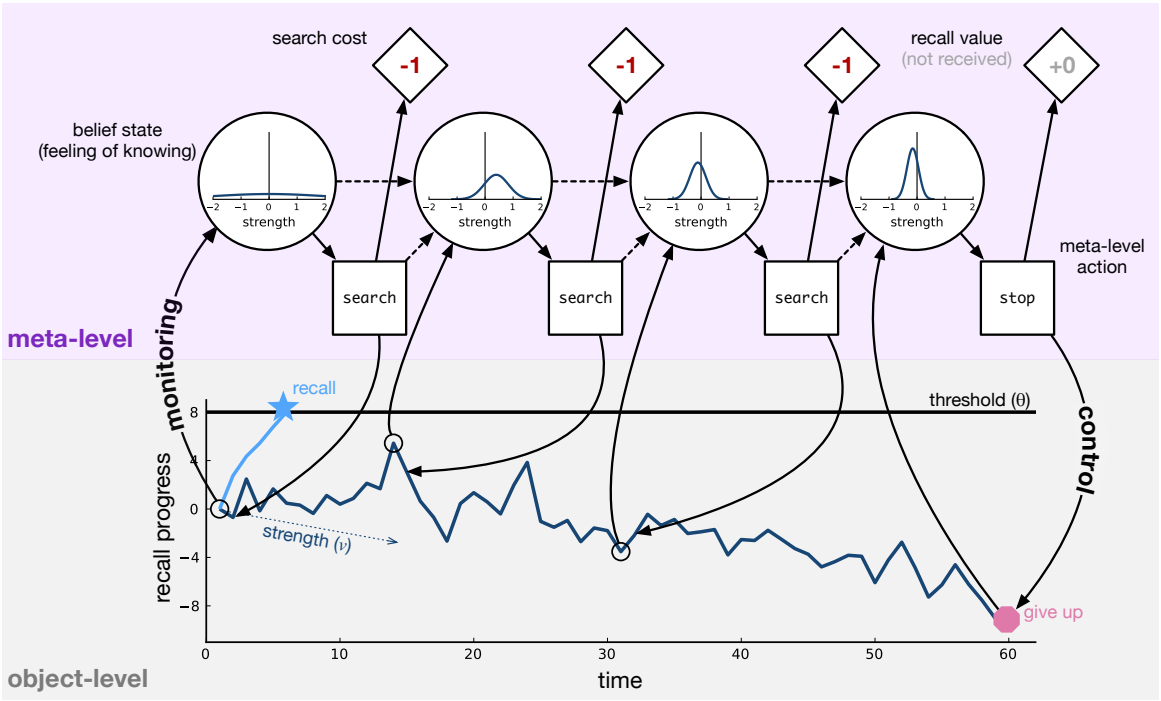

Figure 4.2: A dynamic model of metamemory. Bottom: The object-level recall process is modeled as evidence accumulation. A memory is recalled when a threshold level of evidence is accumulated, with the speed of accumulation (the drift rate) corresponding to the strength of the memory. Top: the metalevel process monitors and controls the object-level recall process. This is modeled as a metalevel MDP where the belief states (circles) correspond to feeling of knowing (estimates of memory strength) and the computations determine whether search continues; the rewards capture the cost of search and the utility of recalling a memory. Solving this MDP yields an optimal policy for determining when to stop searching memory based on partial recall progress.

Empirical evidence for control in memory recall

A number of behavioral studies have already suggested that people are capable of using their ability to monitor their memory in the service of controlling their memory processes. At the acquisition phase, a large body of work has investigated how people preferentially allocate study time depending on how well they have learned different pieces of information (Dunlosky & Hertzog, 1998; Metcalfe, 2009; Gureckis & Markant, 2012). Another substantial literature has addressed how people choose which memories to maintain or forget (Castel, 2007; Williams et al., 2013; Suchow & Griffiths, 2016; Hu et al., 2019). Here, however, we focus specifically on the control of recall.

Most work on metamemory for recall boils down to one essential question: How do people decide whether to (continue to) search for a memory? The initial decision of whether

or not to search at all is often treated as part of a more general *strategy selection* process (Reder, 1988), with memory search being one of multiple possible strategies (along with, e.g., looking up the information in a dictionary). The choice of strategy appears to be driven by an initial feeling-of-knowing (Nhouyvanisvong & Reder, 1998), which is itself driven by surface-level properties of the question, such as familiarity with its terms (Reder & Ritter, 1992). A key finding from this line of work is that people can estimate the probability that they will be able to recall a target faster than they can actually recall it (Reder, 1987). This necessitates some form of metacognitive monitoring, as participants cannot be making judgments of recallability based on the outcome of recall if the former precedes the latter.

Once a memory search has been initiated, how long do people search before giving up? A key finding here is that participants spend longer before giving up on questions for which they report greater feeling of knowing (Nelson, 1984; Nhouyvanisvong & Reder, 1998; Gruneberg et al., 1977; Lachman et al., 1979) or being in a tip-of-the tongue state (Schwartz, 2001). For the latter, participants also show worse performance on a secondary task, indicating more focused processing (Ryan et al., 1982). Feeling of knowing and tip-of-the tongue states are themselves associated with greater subjective familiarity (Reder, 1988), partial recall of the target (Brown & McNeill, 1966; Koriat, 1993; Schacter & Worling, 1985) and the ability to recall given additional information (Gruneberg & Monks, 1974). Together, these results suggest that people are able to accurately identify targets they are likely to recall with further effort, and allocate that effort accordingly.

A related finding, although not one central to control, is that participants give higher confidence judgments when they recall an answer more quickly (Nelson & Narens, 1990). In conjunction with the feeling-of-knowing effects, this produces a striking pattern. Treating both feeling of knowing and confidence as judgments of memory strength, we see opposite relationships between judged strength and response time for successful vs. failed recall. Costermans et al. (1992) demonstrated this pattern in a single study. On each trial, participants were given a general knowledge question. Then, if they provided an answer, they gave a confidence judgment; if they were unable to provide an answer, they instead gave a feeling-of-knowing judgment. Costermans et al. found that, on the recall trials, participants gave higher confidence judgments when they responded more quickly. But on the omission trials, participants gave higher feeling-of-knowing judgments when they responded more *slowly*. In the following section, we will show that both of these findings are consistent with a model in which memory recall follows an evidence accumulation process and search is terminated optimally based on metacognitive monitoring of the rate of progress.

4.1 Model

Following classic theories of metamemory (Nelson & Narens, 1990; Figure 4.1), we specify our model as two interrelated processes operating at different levels. The *object-level* process includes the mechanisms supporting recall itself. Here, we abstract away from the details of memory search, modeling recall instead as a simple evidence accumulation process (Ratcliff & Tuerlinckx, 2002; Sederberg et al., 2008). The *metalevel* process supervises the object-level process; it *monitors* the rate of progress towards recall and *controls* how long the search process is allowed to continue. Here, we assume that the metalevel process is optimal in the sense that it terminates search when the expected costs of search outweighs the expected benefits. The model is illustrated in Figure 4.2; we describe its components below.

Formally, the model has a very similar structure to the model in Chapter 3. Specifically, it uses a similar transition model based on Gaussian evidence accumulation. However, the interpretation is much different. Here, we view evidence accumulation as an abstraction of an underlying memory recall process. Thus, in contrast to Chapter 3, we do not take the Bayesian approach, viewing computation as a process of generating information. Instead, we assume that the computations correspond to searching memory, and the evidence captures how close the memory is to being recalled. This results in a more clear separation between the strategic metalevel component of the cognitive process and the basic object-level component, which describes the cognitive architecture the metalevel component seeks to control.

4.1.1 Object-level process

We model recall as a process of evidence accumulation. Evidence accumulation (or "sequential sampling") models assume that decisions are made by accumulating noisy information over time until a threshold level of evidence is reached. They have been widely applied in the decision-making (Busemeyer & Townsend, 1993; Usher & McClelland, 2001; Ditterich, 2006; Krajbich et al., 2010) and memory (Ratcliff, 1978; Sederberg et al., 2008) literatures, and are successful in accounting for the effects of various experimental manipulations on accuracy and response times during recognition and recall tasks (Ratcliff & Tuerlinckx, 2002; Sederberg et al., 2008; Yonelinas et al., 2010). In our model, the "evidence" captures progress towards recalling a target. Thus, when a threshold level of evidence is reached, the target is recalled (blue star in Figure 4.2).

Concretely, at each time point t, the current recall progress z_t is incremented by a sample

from a Gaussian distribution,

$$z_t = z_{t-1} + x_t \text{ where } x_t \sim \text{Normal}\left(v, \sigma_x^2\right). \tag{4.1}$$

The mean of this distribution, v, controls the rate of accumulation; it is often called the *drift rate* (illustrated as a thin dashed blue arrow in Figure 4.2). In our model, it captures the strength of the memory. The noise σ_x^2 captures the consistency of that progress. The target is recalled when the total progress exceeds a threshold θ.

4.1.2 Metalevel process

The problem of deciding when to cut off an unsuccessful memory search is addressed by the metalevel process. That is, the metalevel process *controls* how long the object-level process is allowed to continue. How should it do so? From a rational perspective, one should keep searching as long as the probability of recall multiplied by the utility of recall is greater than the expected cost of search (Anderson & Milson, 1989). Putting this logic into notation, we can define the optimal metalevel action as

$$c^* = \begin{cases} \textsc{search} & \text{if } p(\text{recall}) \cdot U(\text{recall}) > \text{E}[\text{cost}(\text{search})] \\ \textsc{stop} & \text{otherwise} \end{cases} \tag{4.2}$$

where U stands for utility. The challenge lies in estimating $p(\text{recall})$ and $\text{E}[\text{cost}(\text{search})]$. In our evidence-accumulation model, these values correspond respectively to the probability that the evidence will eventually cross the threshold and the time point at which this occurs.

Intuitively, one could accurately estimate the probability and cost of future recall if one knew the strength of the target memory, v. However, a key assumption of our model—and the metamemory literature more broadly—is that the metalevel process does not have direct access to this information. Instead, we assume that the metalevel process must infer the memory's strength by *monitoring* the object-level process. The existence of such a monitoring process is widely agreed on; however, its precise nature is controversial. In particular, it is unclear to what extent monitoring tracks the underlying memory strength (Hart, 1965), partial recall progress (Koriat, 1993), or superficial cues that happen to be predictive of recall (Reder & Ritter, 1992; Schwartz & Metcalfe, 1992). Resolving this debate is beyond the scope of this chapter. Thus, for simplicity and tractability, we assume that the metalevel process directly observes the state of the object-level process. We emphasize that this is purely a simplifying assumption, and not a claim about how people actually monitor their memory. We

return to this point in the discussion.

Concretely, we assume that the metalevel process observes the current recall progress z_t and the time spent so far t, which provides a complete summary statistic for the entire sequence up to time t. Given this information, the metalevel process then infers a posterior distribution over the strength of the memory,

$$\begin{aligned} p(v \mid z_t, t) &= \text{Normal}\left(v; \mu_t, \sigma_t^2\right) \\ \mu_t = \frac{z_t t \sigma_x^{-2} + \mu_0 \sigma_0^{-2}}{\sigma_t^{-2}} \qquad & \sigma_t^2 = \frac{1}{t\sigma_x^{-2} + \sigma_0^{-2}} \end{aligned} \tag{4.3}$$

where μ_0 and σ_0^2 encode the agent's prior, Normal $\left(\mu_0, \sigma_0^2\right)$. To build intuition, note that with a very weak prior (large σ_0^2), μ_t reduces to z_t/t, the average rate of recall progress. This time-varying belief about the strength of a memory formalizes the concept of *feeling of knowing*.

Given this estimate of memory strength, how should the metalevel process determine whether to continue searching. That is, how should *monitoring* inform *control*? To answer this question, we can pose the model as a metalevel MDP and identify the optimal policy.

4.1.3 Metalevel Markov decision process

To characterize optimal metalevel control of memory recall, we cast the model as a metalevel MDP in which the mental state captures partial recall progress and the computations correspond to searching for a target memory. We detail the five components below.

World states The world state defines the strength of the target memory, that is, the drift rate v. Admittedly, the term "world state" is not entirely appropriate for this variable, as it really describes a property of the agent's internal cognitive architecture. However, it has the formal role of the world state, capturing information that is not known to the agent, but affects transition function.

Mental states The mental state defines the current level of recall progress as well as how long search has continued for; it can thus be represented as a tuple $w_t = (t, z_t)$. The corresponding belief state is given in Equation 4.3. Note that this belief state corresponds exactly to the mental states in Chapter 3. In this case, we must distinguish between the mental state and the belief state because the transition and reward functions depend on the total memory progress z_t rather than the posterior estimate of memory strength.

COMPUTATIONS A computation corresponds to searching for the target memory, allowing the evidence to accumulate for another time step. The termination operation $\perp$ corresponds to terminating the memory search process. This results in an external action that depends on the current recall progress:

$$\pi_{\text{act}}(m_t) = \begin{cases} \text{recall} & \text{if } z_t > \theta \\ \text{give up} & \text{otherwise} \end{cases} \tag{4.4}$$

That is, if the recall progress is above threshold, executing $\perp$ corresponds to recalling the memory (e.g. reporting it). Otherwise, it corresponds to giving up. We assume that $\perp$ is always executed when the recall progress exceeds threshold, as any reasonable policy would do.

TRANSITION FUNCTION The transition function T captures the evidence-accumulation dynamics of the object-level process, as defined in Equation 4.1. The marginal transition function is

$$\begin{aligned} T(m_{t+1}|m_t, c) &= p(z_{t+1} \mid t, z_t) \\ &= \int p(z_{t+1} \mid z_t, v) p(v \mid t, z_t)\, dv \\ &= \int \text{Normal}\left(z_{t+1} - z_t \mid v, \sigma_x^2\right) \text{Normal}\left(v \mid \mu_t, \sigma_t^2\right)\, dv \\ &= \text{Normal}\left(z_{t+1} - z_t \mid \mu_t, \sigma_x^2 + \sigma_t^2\right). \end{aligned} \tag{4.5}$$

The two substitutions in the third line follow from Equations 4.1 and 4.3, respectively. The final line is a standard property of Gaussian distributions (Murphy, 2007).

REWARD FUNCTION The reward function encodes the benefit of recall and the cost of search. We use the same cost function as in Chapter 3: a fixed cost γ_{sample} for each timestep (we add a switch cost later, when we generalize to multiple memories). The termination reward captures the utility of recall, which is only possible when the recall progress exceeds the threshold.

$$R(m_t, \perp) = \begin{cases} U(\text{recall}) & \text{if } z_t \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

4.1.4 Optimal policy

In the metalevel MDP specified above, the policy corresponds to a strategy for deciding when to give up on recalling a memory. To identify the optimal policy, we apply backwards induction (Section 2.7.1). This method computes the optimal value function of the MDP V^*, which specifies the maximal reward that could be gained in expectation starting from any mental state. To build intuition, we can factorize V^* for the current model into two components, capturing the utility of recall and the cost of search,

$$V^*(m_t) = p(\text{recall} \mid m_t)U(\text{recall}) - (\mathrm{E}\left[t_{\max} \mid m_t\right] - t)\gamma_{\text{sample}}, \tag{4.7}$$

where $t_{\max}$ is the time step on which the item is recalled or the search is terminated. The optimal policy is then defined as

$$\pi^*(m_t) = \begin{cases} \textsc{search} & \text{if } \mathrm{E}_{m_{t+1}\sim T(\cdot\mid m_t,\textsc{search})}[V^*(m_{t+1})] > \gamma_{\text{sample}} \\ \bot & \text{otherwise} \end{cases} \tag{4.8}$$

To understand this equation in comparison to Equation 4.2, note that $p(\text{recall})$ and $\text{cost}(\text{search})$ have each been split into two components, capturing immediate versus future outcomes. The immediate recall probability is encoded in the transition function, $T(\cdot \mid m_t, \textsc{search})$; the immediate search cost is encoded in the reward, $-\gamma_{\text{sample}}$. The expected future outcomes are both integrated into $V^*(m_{t+1})$, as shown in Equation 4.7.

4.1.5 Predictions

As illustrated in Figure 4.3, the model makes two key predictions regarding the relationship between memory strength and response time. First, stronger memories should be recalled more quickly because stronger memories accumulate progress faster and hit the threshold sooner. Note that this prediction is a simple consequence of the object-level process and does not depend on metacognition. Second, stronger memories should be abandoned less quickly. In particular, while the metalevel process can quickly identify very weak memories as such (leading it to terminate search), marginal-strength memories produce ambiguous evidence and it takes more time for the metalevel process to determine that the memory is too weak to justify further search.

Figure 4.3 also highlights that the optimal policy can be represented as a time-varying threshold, such that search is terminated if the progress ever falls below the threshold (c.f. Drugowitsch et al., 2012). In the language of Marr (1982), this can be understood as an

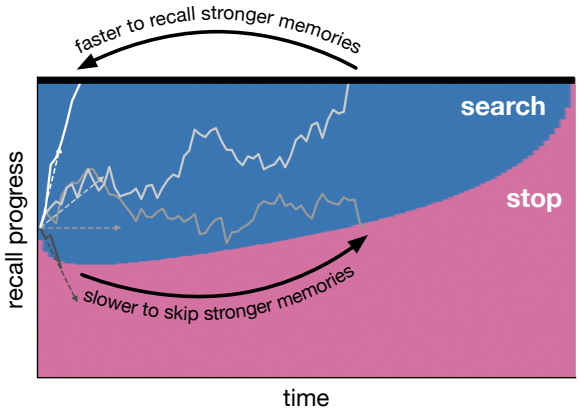

Figure 4.3: Experiment 1 optimal policy and predictions. The optimal policy partitions the state space into two sections, one (blue) in which the policy continues searching, and another (pink) in which it terminates search. The response time for each trial is thus given by the first time point at which the recall progress either exceeds the threshold (recall) or enters the pink region (no recall). In the former case, stronger memories (lighter lines) will result in faster responses because such memories accumulate progress and hit the threshold faster. In the latter case, stronger memories will result in slower responses because such memories can hover in the search region before ultimately hitting the stop region.

algorithmic-level implementation of the computational-level theory outlined above. We return to this point in the General Discussion. Note that the threshold is non-monotonic because a fixed amount of negative progress provides stronger evidence that the memory has low strength if the negative progress was generated more quickly.

4.2 Results

4.2.1 Experiment 1: Optimal stopping

In our first experiment, we sought to replicate and extend the findings of Costermans et al. (1992) in a cued-recall setting. The key finding from the original study was that participants reported higher confidence judgments on trials where they more quickly recalled the answer to a question, but lower feeling-of-knowing judgments when they more quickly reported being unable to recall the answer. Our model can capture both of these effects under the assumption that the metamemory judgments are based on the inferred memory strength at response time (explained below). However, it is also possible that the metamemory judgments reflect a purely post-hoc rationalization of the longer response time, not influencing the decision to stop at all. To avoid this reverse-causality concern, we modified the task such that we could obtain objective measures of memory strength before the critical trials. Specifically, we used a cued-recall paradigm that allowed us to query the same target

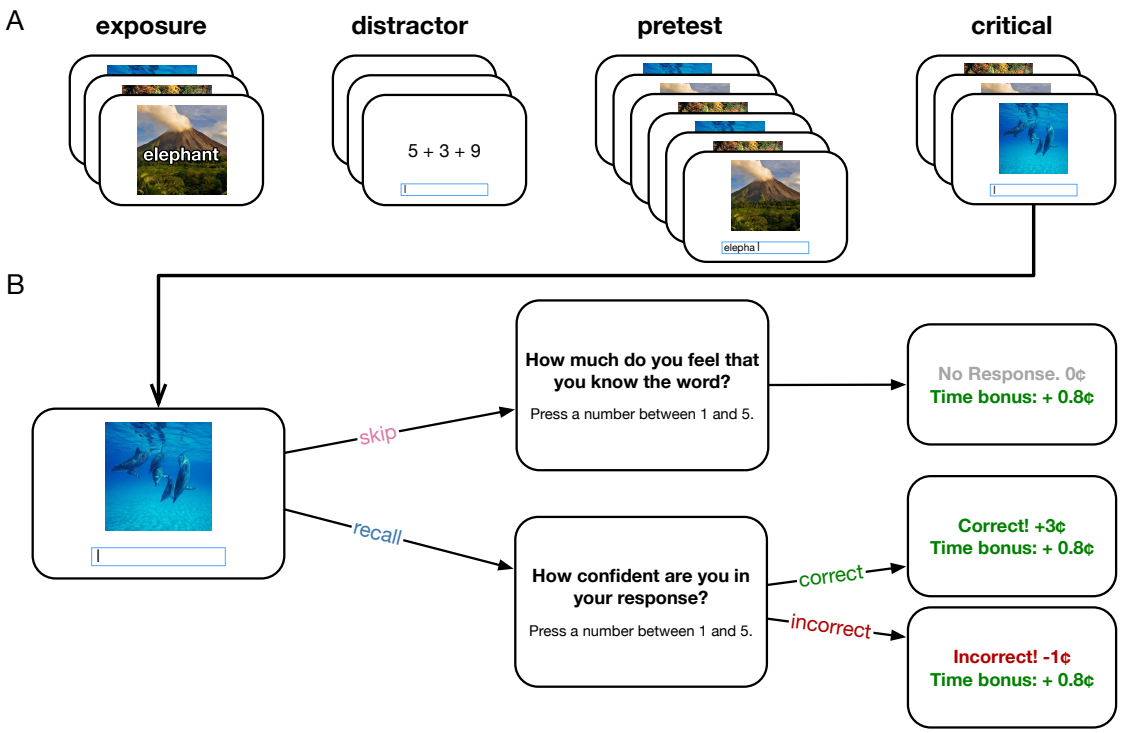

Figure 4.4: Experiment 1 procedure. (A) Participants viewed 40 image-word pairs for two seconds each (exposure). They then completed simple math problems for 60 seconds (distractor). Next, they attempted to recall the word associated with each image, two trials per image (pretest). Finally, they completed one critical trial for each image. (B) Critical trials were similar to pretest trials, except that incorrect responses were penalized. The penalty could be avoided by providing an empty response, "skipping" the trial. After giving a response, participants made a metamemory judgment, confidence if they had entered a word and feeling-of-knowing if they had not. A speed bonus was given regardless of the response.

multiple times. This allowed us to test whether people's stopping decisions depended on their true memory strength, as the optimal policy predicts they should.

The task design is illustrated in Figure 4.4. Participants learned a mapping between images and words in a single round of passive exposure. After a distractor task, they attempted to recall the word for each image in the pretest trials (two trials per pair). These trials provide an objective measure of how well each participant had learned each pair. In the critical trials, we probed participants' metalevel decisions to give up on memory search. To that end, we added a large speed incentive and an error penalty, allowing participants to skip a trial without penalty and still earn the speed bonus. This creates an incentive to quickly identify trials in which the target is unlikely to be correctly recalled. See Section 4.4.1 for details on the procedure.

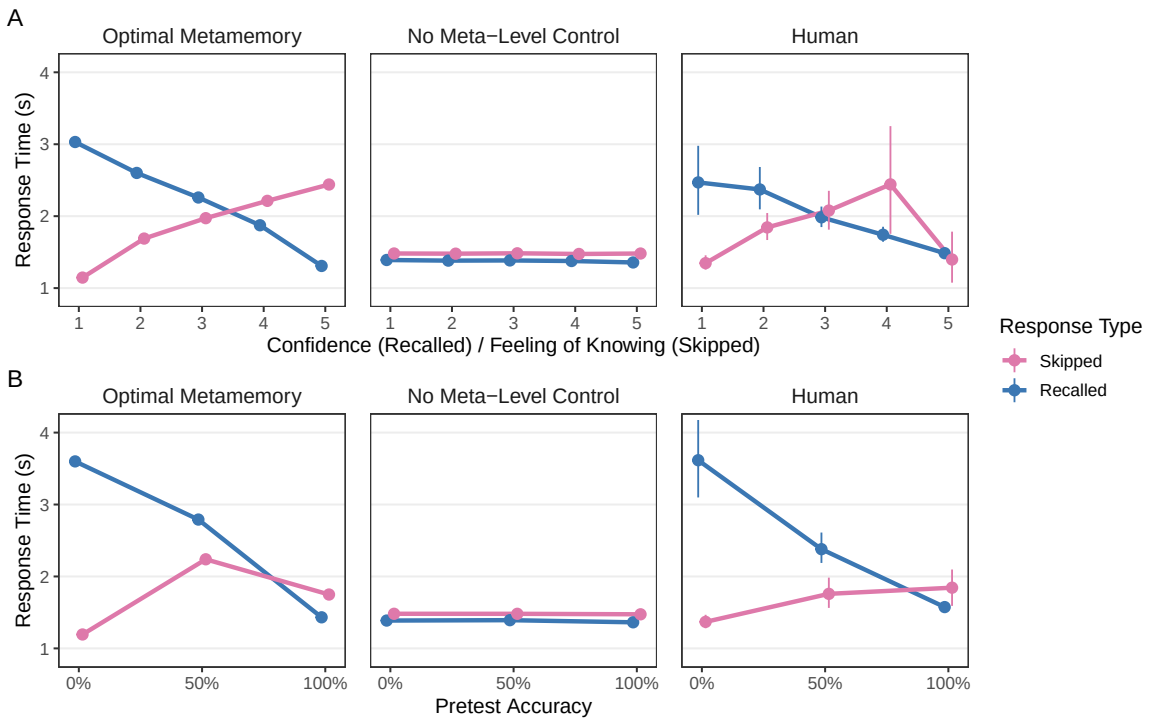

Figure 4.5: Opposing effects of memory strength on time to recall vs. skip a target. (A) Reaction time as a function of metamemory judgment (feeling of knowing for skip trials, confidence for recall trials), separately for trials in which participants correctly recalled the target vs. skipped without responding (errors are excluded). The left panel shows the predictions of the proposed optimal metamemory model, the center panel shows the predictions of a model with the same recall process but no metalevel control (sampling stopping time from the empirical distribution), and the right panel shows human data. The models' metamemory judgments are made based on the inferred memory strength at the end of the trial. (B) The same, but as a function of accuracy rate for the presented cue in the pretest phase. *Note*: Points show means of participant medians and error bars show 95% bootstrapped confidence intervals over participant medians. Each model is treated as one participant (with one million simulated trials). All plotting decisions (including which effects to show, the aggregation method, and axis limits) were pre-registered.

Metamemory judgments and response time

As illustrated in Figure 4.5A (right), we replicated the basic pattern reported by Costermans et al. (1992). Participants were faster to correctly recall targets that they reported greater confidence in ($B = -0.241$, 95% CI [-0.290, -0.193], $t(524.4) = -9.81$, $p < .001$), but slower to skip targets that they reported higher feeling-of-knowing for ($B = 0.385$, 95% CI [0.310, 0.461], $t(189.0) = 10.01$, $p < .001$). The confidence effect indicates that people's metamemory judgments correlate with true memory strength (assuming that stronger memories are recalled more quickly), suggestive of metacognitive monitoring. In contrast,

the feeling-of-knowing effect suggests that participants spent longer trying to recall targets that they thought they were more likely to recall, a form of metacognitive control.

To capture these metacognitive judgments in the model, we assume that the judgment (confidence or feeling of knowing) is made based on the inferred evidence accumulation rate at the end of the trial (see Simulation procedure, above). Unsurprisingly, the optimal model infers a higher accumulation rate when a word is recalled faster, and thus produces higher judgments. More importantly, it gives higher feeling-of-knowing judgments for cues which it took longer to skip. A lesioned model with the same object-level process but no metalevel control (sampling skipping times randomly; described in Section 4.4.1) failed to capture either effect. Interestingly, with some parameter values, the lesioned model can capture either effect in isolation. However, it was not able to predict both effects at once with any parameter configuration (see Appendix B.3).

Pre-test accuracy

While these results are suggestive, the direction of causation is not clear. The metamemory judgment temporally follows the response time; thus, it is entirely possible that participants are actually reporting higher feeling-of-knowing judgments because they spent longer searching. To test whether participants stopping times are truly influenced by an awareness of the memory's strength, we can replace the metamemory judgment with an objective measurement of memory strength, concretely, the proportion of pretest trials in which the target was recalled correctly. As shown in Figure 4.5B, the model predicts a similar pattern: faster recalls and slower skips with increasing pretest accuracy.[2] People were likewise faster to recall targets that they had previously recalled correctly ($B = -0.994$, 95% CI [-1.237, -0.751], $t(237.3) = -8.02, p < .001$). More importantly, they were also slower to skip such targets ($B = 0.405$, 95% CI [0.268, 0.543], $t(226.0) = 5.77, p < .001$). The lesioned model could not produce this effect with any parameter values. These results suggest that participants' decisions to stop searching depended on a metacognitive awareness of how likely they were to recall the target.

[2]The non-monotonic prediction for skip trials is due to a selection effect: the optimal model only skips high-strength memories when it greatly underestimates the strength. This becomes increasingly unlikely as the trial progresses and more evidence is collected. Thus, these "erroneous" skips generally occur quickly (c.f. the "fast errors" phenomenon in decision-making Ratcliff & Rouder, 1998).

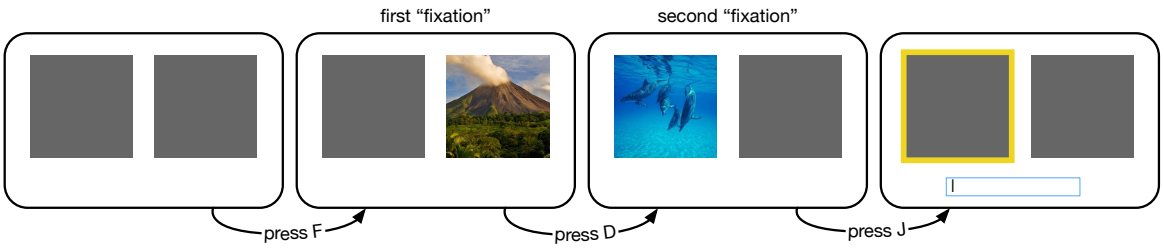

Figure 4.6: Experiment 2 critical trials. Participants were presented with two images on each trial and were instructed to recall the word associated with either of them. Only one cue was visible at a time and participant could flip between them with the D and F keys. At any point they could press J or K to select an image for recall, at which point they had five seconds to enter the associated word. We refer to the periods of time in which one cue was contiguously visible as “fixations.”

Summary

In this experiment, we found that participants were faster to recall targets with higher strength but slower to give up on targets with higher strength. Importantly, this pattern held for both a subjective measure of strength (replicating Costermans et al., 1992) as well as an objective measurement of strength (accuracy on the pretest trials). The latter is critical because it shows that people spent longer searching for memories that they were actually more likely to recall, thus demonstrating the objective utility of metamemory in guiding recall. Furthermore, because pretest accuracy is defined before the critical trials, this measure is not subject to the reverse-causality concern that response times are driving feeling of knowing rather than vice versa (Schwartz, 2001).

The full pattern of results was qualitatively consistent with the optimal model, which terminates search when the expected value of search falls below the expected cost, and reports a Bayesian estimate of strength as feeling of knowing or confidence. Perhaps more importantly, the results could *not* be captured by a model without metalevel control. This provides computational support for the intuition that the correlation between search time and feeling of knowing is a distinctive signature of an adaptive metamemory process.

4.2.2 Experiment 2: Optimal targets selection

In our first experiment, we considered a very simple form of metamemory, the decision of how long to search memory before giving up. Control of memory is not limited, however, to such a simple kind of decision. Instead, successful recall often requires deciding between multiple strategies for finding an answer (Reder, 1988). Going further, Koriat (2000) has characterized recall as a form of problem solving, with a metalevel process “coordinating be-

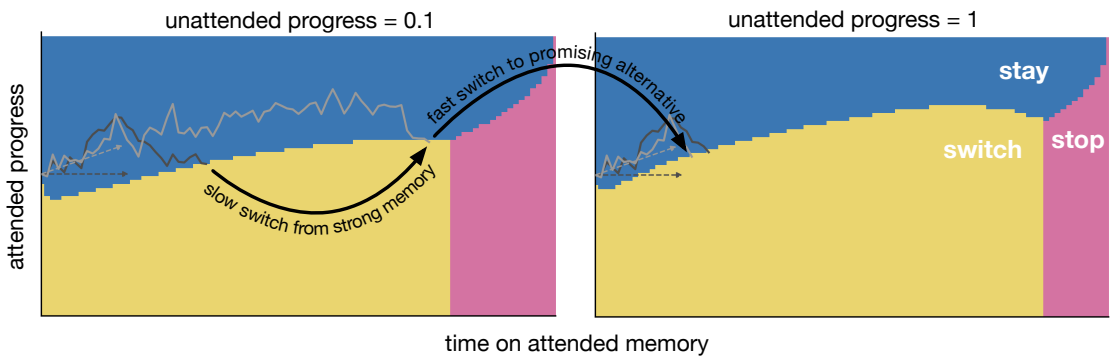

Figure 4.7: Experiment 2 optimal policy and predictions. With two possible memories to recall, the optimal policy partitions the state space into three sections where it is optimal to either: continue searching for the currently attended memory (blue), switch to the other memory (yellow) or give up (pink). The optimal policy depends on the recall progress and time spent on both memories; here we show two slices of the full four-dimensional state space, setting the time spent on the unattended memory to 10 timesteps and its progress to either 0.1 (left) or 1 (right). The gray lines show example progress traces for a weak (dark gray) and moderately strong (light gray) attended memory. The arrows highlight two key features of the optimal policy: it is slower to switch when the currently attended memory is strong (vs. weak), but faster to switch if the unattended memory has already shown promising recall progress (vs. if it has not).

tween different operations directed toward the recovery of the elusive memory" (p. 334). A careful characterization of the operations underlying recall (let alone how they are chosen) is beyond the scope of this chapter. Nevertheless, in our second experiment, we sought to characterize a core aspect of this richer form of metamemory: monitoring multiple target memories and allocating retrieval efforts between them.

On each trial, participants were presented with two cues and could recall the target for either one of them (Figure 4.6). They thus had to make a metacognitive decision about which of the two possible target to search for at each moment. In order to observe how this selection process unfolded over time, we used a keypress-contingent display, such that only one cue was visible at any moment. This provides process-tracing data similar to eye-tracking, but in a format amenable to online presentation.

Predictions

As detailed in Section 4.4.2, we extended the model to the multiple-memory case by creating a separate evidence accumulation process for each target memory. The metalevel process decides which accumulator is allowed to progress at each time point. The optimal policy (illustrated in Figure 4.7) can thus be characterized by when it decides to switch between the two memories (and also when it decides to terminate search, as before).

In general, the optimal policy attends to the memory that it believes can be recalled soonest, as this will incur the least cost. In our experiment, attending to a memory is operationalized by looking at (or "fixating") the associated cue. Thus, the basic prediction is that the cues for stronger memories will receive a greater share of the total fixation time. More specifically, the model will be slower (and less likely) to switch away from a strong memory, but faster to switch when the other memory is strong.

Inspecting model simulations, we also discovered a surprising feature of the optimal policy. Its final fixations are longer than its non-final fixations. This prediction is surprising because we see the opposite pattern in decision-making tasks (e.g., Krajbich et al., 2010, discussed further in the General Discussion). Why does the optimal policy make this prediction? We can understand long final fixations as sort of "rational commitment" behavior, in which the model effectively commits to recalling one memory before it is actually recalled. After committing to a memory, the model continues to attend to it until it is either recalled or its inferred strength drops well below the competitor. The latter occurs only rarely. Thus, commitment tends to happen on final fixations, and final fixations are therefore longer. To see why this is rational, note that constant switching between the two memories is wasteful, as it can take up to twice as long as if one had immediately committed to one memory. On the other hand, immediately committing to the first cue could lead to getting stuck on an out-of-reach memory. Thus, the model only commits to a cue after becoming reasonably confident that the cue is strong.

Attention is drawn to stronger cues

The critical model predictions concern participants' "fixation" behavior in the double-cued-recall trials, that is, the sequence of key presses they made to alternately display the two images. The most basic prediction of the optimal model is that participants should attend more to the cue with stronger memory. Intuitively, the target for the stronger cue can be recalled faster, and so time spent looking at this cue is more productive. Indeed, as illustrated in Figure 4.8A, participants spent substantially more time looking at cues that were stronger than the other available cue ($B = 0.191$, 95% CI [0.182, 0.199], $t(572.0) = 42.40$, $p < .001$). However, this pattern is also shown (to a lesser extent) by a lesioned model that randomly switches between the cues. This is due to two properties of the object-level recall process. First, the last fixation is always on the cue whose target is recalled. Second, stronger cues are more likely to be recalled. Together, this implies that stronger cues are more likely to be fixated last, and thus receive more fixations (and more fixation time) on average.

Inspecting the timecourse of attention across the trial (Figure 4.8B) reveals a more nu-

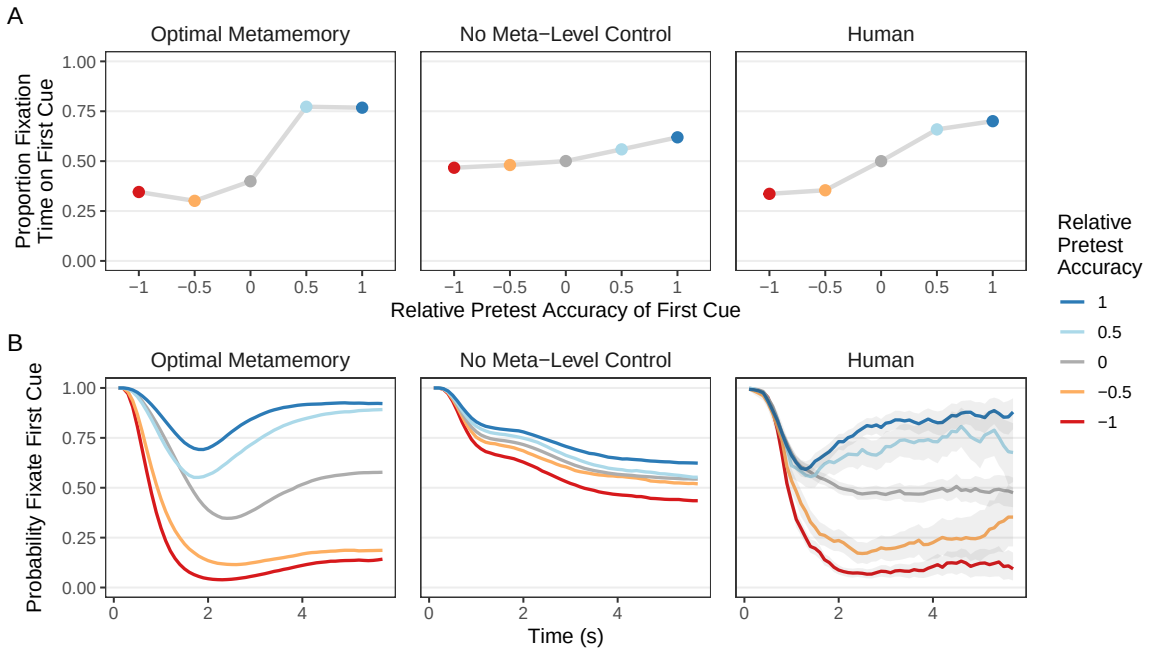

Figure 4.8: Attention is drawn to stronger cues. (A) The proportion of total viewing time allocated to the first-seen cue image as a function of the difference in pretest accuracy of the first- and second-seen cues. Trials for which the second cue was never shown are excluded. Note that the 95% confidence intervals are too small to be distinguishable. (B) The probability that the first-seen cue is currently displayed over the course of the trial, split by relative pretest accuracy.

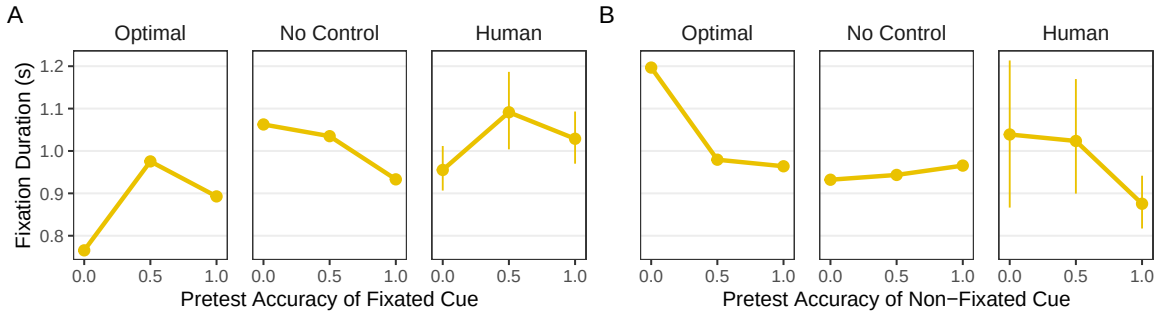

Figure 4.9: Non-final fixation durations. (A) The duration of non-final fixations as a function of the pretest accuracy of the currently fixated cue. (B) The same, but for the pretest accuracy of the cue that is *not* currently fixated (first fixations excluded).

anced picture. People tend to quickly check both cues (as indicated by the the initial dip in probability of fixating the first cue). Then, if the second cue is stronger (red lines), they continue fixating it. But if the first cue is stronger (blue lines), they switch back to it. From about one to three seconds, participants show an increasingly strong tendency to fixate the

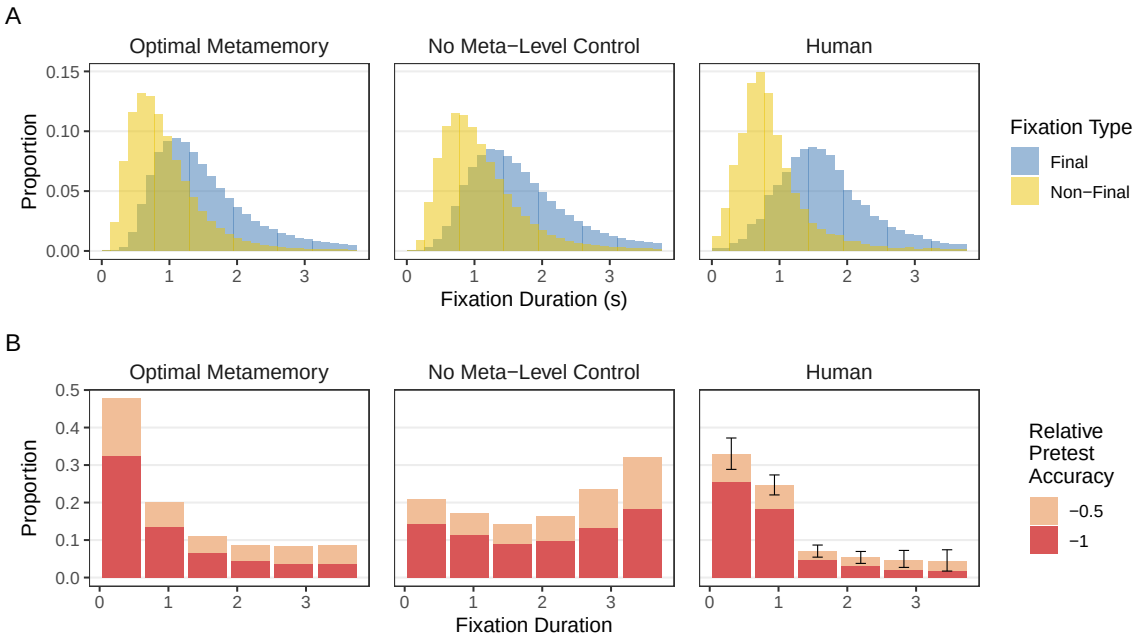

Figure 4.10: Rational commitment to stronger cues. (A) The distribution of final and non-final fixation durations. (B) The proportion of fixations on items with lower pretest accuracy than the alternative, separately for fixations of different durations (first fixations excluded). The error bars indicate 95% confidence intervals computed over the total proportion (including -0.5 and -1) for each participant.

stronger cue, and this tendency remains stable for the remainder of the trial. The optimal model shows a similar pattern, although its tendency to fixate the stronger cue emerges faster (as indicated by the earlier divergence of fixation probability for different relative strengths). In the lesioned model, a slight tendency to fixate the stronger cue emerges in the first second, and remains small for the remainder of the trial. This is due to the last-fixation confound discussed above.

Non-final fixation durations

Although the lesioned model was not able to capture the strength or timecourse of the tendency to fixate stronger cues, the fact that it can produce the effect at all casts some doubt on the interpretation of the pattern in human data. Thus, for our next analysis, we inspected the duration of individual non-final fixations, which are clearly not subject to the last-fixation confound. As illustrated in Figure 4.7, the optimal policy's decision to terminate a fixation by switching to the other cue depends on the recall progress of both targets; it is slower to switch when the currently attended cue is generating rapid progress, but faster

to switch when the unattended cue has already generated substantial progress. The model thus predicts that non-final fixation durations will increase[3] with the pretest accuracy of the fixated cue, but decrease with the pretest accuracy of the non-fixated cue. As illustrated in Figure 4.9, both of these predictions were confirmed: participants' non-final fixations increased with the pretest accuracy of the fixated cue ($B = 0.106$, 95% CI [0.075, 0.137], $t(230.7) = 6.65, p < .001$) and decreased with the pretest accuracy of the non-fixated cue ($B = -0.456$, 95% CI [-0.578, -0.334], $t(59.7) = -7.32, p < .001$; first fixations excluded). The lesioned model predicts no such effect. In fact, it is incapable of predicting either effect under any parameter setting that achieves accuracy levels comparable to our participants (with very low accuracy, it can capture the effect through a selection mechanism; see Appendix B.3).[4]

Rational commitment

For our final analysis, we tested the model's "rational commitment" behavior, in which the model effectively commits to recalling one memory before it is actually recalled. One observable consequence of this is that the model's final fixations are longer than their non-final fixations, as the commitment decision can occur when the memory is still well below threshold. Consistent with this, Figure 4.10A shows that our participants final fixations were indeed longer than their non-final fixations ($B = 0.872$, 95% CI [0.813, 0.930], $t(435.8) = 29.29, p < .001$). However, this figure also shows that the lesioned model can capture this effect as well. It is able to do this through a "random commitment" mechanism: by assuming a high-variance switching-time distribution, it occasionally samples a very long fixation duration, which is likely to end in recall.

Did participants' long final fixations reflect rational commitment or random commitment? The key distinction between these types of commitment is that it is only rational to commit a memory that is at least as strong as the alternative. Therefore, only with a random

[3] To be exact, the model predicts a non-monotonic effect such that fixations are longest for intermediate strength cues. This is due to the same selection effect that produces a non-monotonic prediction in Figure 4.5B. Switching away from a strong cue suggests that the model incorrectly estimated its strength (because progress happened to be slow at first). This becomes increasingly unlikely as the fixation progresses and more evidence is collected. Such "erroneous" switches are thus more likely to occur quickly.

[4] Again, the lesioned model predicts a reverse effect due to selection effects similar to those discussed in Experiment 1 and the previous footnote. For the fixated cue's strength, it is exactly the same logic as in Experiment 1. A strong cue is likely to be recalled quickly; it will only be switched away from when a short switching time is sampled. For the non-fixated cue's strength, the logic is more complex. Because we condition on one of the memories being recalled, the non-fixated cue being weak implies that the fixated cue is strong. Following the same logic as before, this in turn implies that a short switching time was sampled (as otherwise the cue would have been recalled).

commitment strategy will one allocate long fixations to cues that are weaker than the alternative. Figure 4.10B thus shows the frequency with which the weaker memory is fixated, separately for fixations of different length. Importantly, we do not limit this analysis to final fixations as this would exclude the cases where the lesioned model sampled a long fixation duration on a low-strength cue, thus selecting for the cases where the lesioned model was "accidentally" rational. In both the optimal model and the human data, we see that long fixations are unlikely to be directed to weaker memories. In particular, the probability of fixating the weaker memory significantly decreased with fixation duration ($B = -1.482$, 95% CI [-1.651, -1.313], $z = -17.22$, $p < .001$; logistic regression, excluding trials where the cues had equal pretest accuracy). In contrast, under the lesioned model, the weaker cues are actually more likely to receive long fixations (because strong cues are likely to be recalled quickly, cutting off long fixations). The fact that people show a strong tendency to not direct long fixations to weak cues despite this selection effect suggests that their commitment decisions were indeed rational.

Note that we we did not pre-register this final analysis because our initial, less flexible, implementation of the lesioned model that sampled switching times from the empirical distribution could not capture the long final fixation effect (see Appendix B.1, Figure B.5A). However, the effect also holds in a previously collected dataset, which we did not use when developing Figure 4.10B or the accompanying statistical test. This provides a quasi-confirmatory test of the exploratory analysis in the previous paragraph. See Appendix B.2 for details.

Summary

In this experiment, we found evidence of a richer form of metalevel control of memory. Specifically, when presented with two cues associated with different target memories, our participants directed their attention towards the cue associated with the stronger memory. This was reflected in the overall proportion of fixation time, the timecourse of fixations, and the duration of individual non-final fixations. The latter two results are especially important because they are behavioral signatures of metacognitive control during the retrieval process itself (before recall is completed or abandoned). To our knowledge, this is the first empirical demonstration of a dynamic metamemory process unfolding over time.

4.3 Discussion

In this chapter, we presented an optimal model of metalevel control for cued recall. The model consists of a metacognitive process that monitors the progress of a basic recall pro-

cess, and optimally controls how long the process is allowed to continue (either terminating recall or switching to recall of a different memory). In two experiments, we showed that human behavior is qualitatively consistent with the predictions of this model. In Experiment 1, we replicated and extended the findings of Costermans et al. (1992), showing that people were faster to recall targets with higher strength (as indicated by both subjective judgments and objective performance) but slower to give up on targets with higher strength. In Experiment 2, we showed that people also attend to stronger memories when they can choose between two possible targets. Together, our results suggest that people can estimate the strength of a memory they are trying to recall and use this information to adaptively control their retrieval efforts.

4.3.1 An explicit instantiation of a classic theory

Our results contribute to the metamemory literature by providing (to our knowledge) the first computationally explicit instantiation of the classic theory of metamemory proposed by Nelson and Narens (1990) in the context of memory recall. According to this theory, metacognition involves the interaction between a metalevel process and an object-level process, where the metalevel process monitors the state of the object-level process and controls it accordingly. In the context of memory recall, they proposed a verbal model in which feeling of knowing is generated by "an evaluation in terms of whether there has been sufficient progress to continue", with the process terminating in an omission error when this feeling of knowing no longer "exceeds the FOK threshold for claiming to know the answer" (p. 137). They went on to suggest that this model could account for the correlation between feeling of knowing and response time on omission trials.

Here, we have formalized this classic model as a sequential decision problem in which the metalevel process executes a sequence of actions (continuing or terminating memory search) to maximize reward (utility of recall minus cost of search) given the observed state of the object-level process (recall progress). By formalizing this problem as an MDP, we could identify the optimal metacognitive control policy using standard dynamic programming techniques. This allowed us to generate quantitative predictions about the observable behavior we would expect to see if people were indeed using a rational metamemory system to control how long they search for an elusive memory. By confirming these predictions in an experiment, we have contributed quantitative support for this classic theory, which was previously supported only by intuitive qualitative predictions.

Formalizing Nelson and Naren's model as an MDP also allows us to create a conceptual link between metacognition and reinforcement learning (RL; Sutton & Barto, 2018). Specif-

ically, we map the metalevel and object-level processes onto the concepts of agent and environment, respectively. That is, we model metacognition as a metalevel agent interacting with an object-level environment in much the same way as one would model, e.g., a mouse (the agent) searching for food in a maze (the environment). RL has become a major theoretical foundation in the psychology and neuroscience of decision-making (Niv, 2009; Dayan & Daw, 2008; Glimcher, 2011). However, for the most part, it has been used to model the interaction between agents and *external* environments. Applying RL to model the metacognitive interaction between an agent and its *internal* environment (c.f. Simon, 1955) creates the opportunity to transfer much of what we know about how people learn to act effectively in the world to understand how people learn to think effectively in their own minds (Lieder et al., 2018).

4.3.2 Rational analysis for metamemory

Outside of metamemory, the intellectual roots of our model lie in Anderson's (1990) rational analysis, with early work demonstrating that the forgetting behavior commonly observed in lab settings is not a weakness or peculiarity of the memory system, but instead a reflection of rational adaptation to the statistics of the environment (Anderson & Milson, 1989). The key idea underlying both these models and ours is that memory (and cognition more generally) can be treated as an optimization problem. More recently, researchers have emphasized that this optimization problem must also account for the constraints imposed by our limited computational resources (Griffiths et al., 2015; Howes et al., 2009; Lewis et al., 2014; Gershman et al., 2015). This approach, sometimes called resource-rational analysis, has generated insight into a wide variety of cognitive processes (see Lieder & Griffiths, 2020, for a review).

Focusing on memory, a large body of work has shown that many apparent memory biases actually reflect optimal statistical reasoning under the constraint of noise or capacity limitations (Gershman, 2021). For example, in reconstruction tasks, people draw on prior knowledge of a stimulus category to adjust for memory imprecision (Huttenlocher et al., 2000), using more abstract categories for less familiar stimuli (Hemmer & Steyvers, 2009). In working memory tasks, people are sensitive to cost-benefit trade-offs when choosing how many items to encode (Howes et al., 2016), how to allocate encoding resources across items (Yoo et al., 2018), and how much total resource to allocate (van den Berg & Ma, 2018). Researchers have also begun to explore the implications of the constraints that emerge from more detailed models of memory. For example, Zhang et al. (2022) characterize the optimal order in which to recall items from a list, assuming that items are stored and recalled with

the context maintenance and retrieval (CMR) model (Polyn et al., 2009; discussed further below). They find that the optimal policy is to start from the beginning of the list and then sequentially recall forwards, providing a rational account of the often-observed primacy and forward asymmetry effects (Zhang et al., 2022).

Focusing on metamemory in particular, two recent models of judgments of learning are based on signal detection theory (Jang et al., 2012) and Bayesian inference (Hu, 2021), both of which have rational bases in probability theory. However, these models only attempt to explain how metamemory judgments are produced, not how they are used—a critical component in a complete rational analysis. We are aware of four computational models that explicitly address the function of metamemory judgments, and all four take a rational approach. Metcalfe (1993) propose a model in which feeling-of-knowing judgments are used to adaptively control the weight with which new memories are encoded in a holographic memory representation, thus controlling variance in the representation. Bennett et al. (2017) present a model in which feeling of knowing determines which questions participants attempt to answer, assuming that they take questions they have high feeling of knowing for. Hu et al. (2019) takes a more explicitly rational approach, proposing a model of cognitive offloading in which a rational agent decides whether or not to use an external memory aid by comparing the expected increase in recall probability with the reduction in payoff. Most similar to our own work, Suchow & Griffiths (2016) proposed an MDP model of working memory maintenance, in which an agent selects actions that increase the strengths of different memories, as encoded in the state. This model could account for the findings of Williams et al. (2013), in which directing participants to forget certain items reduced recall accuracy for the cued item and increased accuracy for un-cued items.

Our model draws on several specific ideas from the work described above, in addition to the core principle of rationality. Like Anderson & Milson (1989), we frame the decision to terminate search as a cost-benefit comparison (Equations 4.2 and 4.8). Like van den Berg & Ma (2018), we jointly consider the problem of how much resource to allocate (here, the resource being time) as well as how to split that resource between items. Like Hu (2021), we model metamemory judgments as the product of Bayesian inference about the strength of a memory. And like Suchow & Griffiths (2016), we model the resource allocation problem as an MDP. Our work contributes to this literature by synthesizing these previous insights, and showing how they can be applied in a new domain, cued recall.

Importantly, like the other resource-rational models reviewed above, our model aims only to characterize the optimal solution to the problem of cognition under limited resources. We have not attempted to explain how the mind might actually approximate that

solution. We do note, however, that executing the optimal policy for our model does not require one to continually perform Bayesian inference over the memory strength. Indeed, as illustrated in Figure 4.3, the optimal policy in the one-cue case is defined by a single boundary, such that recall is terminated if the evidence ever falls below it. Although we identified the shape of this optimal boundary with computationally intensive, model-based methods, it could be well-approximated by simple model-free learning mechanisms. Understanding how people learn effective metacognitive strategies is a subject of ongoing research (Lieder et al., 2018; Jain et al., 2019; Callaway et al., 2022a; Binz et al., 2022; He & Lieder, 2022).

4.3.3 Imperfect monitoring and cue familiarity

While our results make a strong case for the existence of an adaptive metamemory control system in the human mind, we do not claim to have characterized the precise nature of this system. Indeed, for simplicity and tractability, we made several assumptions that are most likely inaccurate. Perhaps most critically, we assumed that the metalevel process has direct access to the state of the object-level process (the recall progress). This form of monitoring can be seen as a simplified form of Koriat's (1993) accessibility model. Specifically, our simplification does not allow for the possibility that incorrect partial information is recalled, which eliminates any divergence between feeling of knowing and recall progress. While this assumption greatly simplified the implementation of the model and has precedence in previous metamemory models (Suchow & Griffiths, 2016; Hu et al., 2019), perfect monitoring is intuitively implausible and it is inconsistent with the sensitivity of feeling-of-knowing judgments to spurious information generated during recall (Koriat, 1993) and manipulations that do not affect accuracy (discussed below).

Beyond imperfect monitoring of recall progress, the metalevel process could rely on other sources of information about a memory's strength. An influential such theory suggests that feeling-of-knowing judgments are based on *cue familiarity*, that is, the degree to which one feels that they recognize the prompt or question that triggered the memory search (Metcalfe et al., 1993). Much evidence supports this view. For example, participants give higher feeling-of-knowing judgments to arithmetic problems that visually resemble previously seen problems (Reder & Ritter, 1992). Similarly, priming words in a question increases feeling-of-knowing without increasing recall (Reder, 1987, 1988; Schwartz & Metcalfe, 1992). Indeed, this fact, along with the observation that priming the memory target does not increase feeling of knowing, has led some researchers to suggest that feeling of knowing does not depend at all on partial recall (Reder & Ritter, 1992; Schwartz & Metcalfe, 1992; but c.f., Jameson et al., 1990; Narens et al., 1994). On the other hand, the observation that people

in tip-of-tongue states are able to correctly recall some aspects of the memory target, such as the number of syllables (Brown & McNeill, 1966) or semantic associations (Koriat, 1993; Schacter & Worling, 1985) suggests that partial recall may influence metamemory, although perhaps only in later stages of the recall process (c.f., Nhouyvanisvong & Reder, 1998).

It is thus plausible that metamemory draws on both partial recall progress and other factors such as cue familiarity. Indeed, this possibility has been formalized and supported by models in which a metalevel and object-level signal are imperfectly correlated (Fleming & Daw, 2017; Jang et al., 2012; Hu, 2021). Intuitively, the metalevel signal in these models contains some information about recall progress (or "processing experience") and some information about the underlying memory strength from other sources (e.g. cue familiarity), with both components corrupted by independent noise. Our model can be seen as a special case of this class of models, where the amount of information from other sources and the noise are both fixed to zero. Allowing the noise to be non-zero would result in a formalization of Koriat's (1993) accessibility model, in which metamemory is a noisy readout of actual recall progress. Further allowing additional sources of information besides recall would yield a complete model in which control is jointly informed by recall progress and ancillary cues. Unfortunately, both of these generalizations make the metalevel belief state depend on the full trajectory of signals (as opposed to just the sum), making the optimal policy intractable to solve.[5] Nevertheless, exploring the implications of imperfect monitoring and cue familiarity for metalevel control (with suitable approximations to the optimal policy) is a critical direction for future research.

Finally, we emphasize that our we do not take our empirical results to support any specific theory of monitoring. Although we have only implemented a model that monitors partial recall, we believe that similar predictions could be made by a model whose monitoring component is based purely on cue familiarity. For this reason, we have taken care not to claim that our results suggest that people can monitor the progress of memory recall. Our results simply suggest that people are capable of inferring the *strength* of a memory. Rigorously evaluating computational models of exactly how people make that inference will require additional data. Developing a paradigm to produce that data is another critical direction for future research.

[5] To see why this is, note that when the recall progress is not directly observed, the agent must infer a posterior distribution over it. This posterior must account for the continual observation that the boundary has not been crossed, and the likelihood of that event depends on the order of the signals. For example, if a large positive signal is followed by a large negative signal, the fact that the threshold was not crossed after the first signal indicates that the large positive signal was not accompanied by a similarly large change in progress. If the order is reversed, no such inference is drawn, and so the estimated recall progress across the two time steps will be higher.

4.3.4 Beyond cued recall

Although we have focused on cued recall, the decision about when to terminate search is present in almost all naturalistic recall tasks, including, for example, free recall. The basic principle of our model is that this decision should be made by balancing the benefit and probability of recall with the cost of search, a principle we inherit from Anderson & Milson (1989). This is in contrast to existing models of free recall such as the context maintenance and retrieval (CMR) model, which assumes that search is terminated either when a target memory is retrieved or after a fixed amount of time has passed (Polyn et al., 2009; Lohnas et al., 2015). Alternatively, according to the Search of Associative Memory (SAM) model, search is terminated after a fixed number of retrieval attempts that do not result in recall of a new word (Raaijmakers & Shiffrin, 1981). In both models, the decision to terminate recall is based on a general stopping rule independent of the state of the memory system. Integrating a rational metacognitive stopping rule into these models is an interesting direction for future research. Empirically validating such a model will likely require additional data, as participants in free recall experiments are typically given a fixed amount of time to recall as many items as possible; the decision to terminate search cannot be observed in this setting.

Metamemory goes beyond simply deciding when to terminate search. In Experiment 2, we considered the problem of arbitrating between multiple externally provided memory cues. In more naturalistic settings, however, people may need to generate their own cues. Indeed, in some cases, people generate information that is incidental to the information they are searching memory for, for example, recalling where people live in an attempt to recall their names (Williams & Hollan, 1981). These incidental pieces of information can then be used as "stepping stones on the way to the sought-after target" (Koriat, 2000, p. 334). This metaphor highlights the sense in which memory recall is a sequential decision problem. In our framework, each generated cue would correspond to a metalevel action, with some actions serving only to bring one to a mental state where one can generate a more useful cue. This suggests a fascinating direction for further research: How do people know which "stones" are going in the right direction?

4.3.5 Beyond memory

Our results also contribute to the literature on metacognition more broadly. Beyond memory, a wide variety of functional roles for metacognition have been proposed, including the regulation of perception (Deroy et al., 2016), judgment (Polanía et al., 2019; Lebreton et al., 2015), decision-making (Yeung & Summerfield, 2012; De Martino et al., 2013), learning

(Frömer et al., 2021; Nassar et al., 2012), information seeking (Boldt et al., 2019; Desender et al., 2018), and social interaction (Frith, 2012). Some of these roles have been incorporated into computational models that formally describe how confidence might inform decision about, e.g., when to opt out of a difficult trial (Kiani & Shadlen, 2009), when to change one's mind given additional evidence (Folke et al., 2016), how to interpret error feedback (Frömer et al., 2021), or where to fixate in visual search (Stewart et al., 2022). However, these models often assume that the functional role of metacognition is *static*. Although the mechanism underlying confidence judgments may be dynamic (typically being based on evidence accumulation Vickers, 1970; Pleskac & Busemeyer, 2010; Moreno-Bote, 2010), any resulting metacognitive control occurs in a separate stage, typically as a single action, and not feeding back onto the object-level process in a dynamic, interleaved way. In an influential review, Yeung & Summerfield (2012) identified this as a critical gap to be explored in future research.

In the intervening decade, this gap has already begun to be filled. In particular, a substantial body of work has explored the within-trial dynamics of metacognition in controlling a decision-making process. These models track posterior distributions over an evidence accumulation rate and use this information to make optimal decisions about when to stop accumulating evidence (Drugowitsch et al., 2012; Woodford, 2014; Bitzer et al., 2014; Fudenberg et al., 2018; Tajima et al., 2019) or how to allocate attention between multiple options (Chapter 3; Jang et al., 2021). However, to our knowledge, this type of model has been applied exclusively in decision-making contexts. Here, we have shown how a model with very similar structure can be applied to memory recall.

The key structural difference between the model proposed here and these previous models of dynamic metacognition for decision-making is the assumption of an *exogenous* threshold associated with recall. That is, the amount of "evidence" necessary to recall a memory is not under the agent's control. This contrasts with decision-making models, where both thresholds (for choosing each option) are *endogenous*. The agent can choose an option based on very little evidence if they so desire. This simple change has a profound effect on the role of attention in the model. In the decision-making context, the sole purpose of fixating an option is to collect information about its utility. In the memory context, fixating the cue similarly provides information about its strength; but it also contributes to recalling the associated memory.

This additional functional role for attention (stimulating recall as well as estimating strength) has two important consequences for the model's predictions in the two-memory case. First, the model predicts—and our results confirm—that cues for stronger memories receive more

fixation time. In contrast, as we saw in Chapter 3, optimal models of attention in binary choice predict equal attention to high- and low-value items. Second, the model predicts—and again, our results confirm—that final fixations are longer than non-final fixations. In contrast, evidence accumulation models of attention-guided decision-making predict that final fixations will be shorter than non-final fixations, a prediction that is consistently confirmed in data (Krajbich et al., 2010; Krajbich & Rangel, 2011; Tavares et al., 2017). The fact that such a simple structural change can account for these major qualitative differences in the allocation of attention in value-based choice vs. cued recall suggests the promise of our metalevel MDPs as a general framework for modeling metacognition.

4.3.6 Conclusion

In this chapter, we have developed and experimentally validated a model of optimal metalevel control for memory recall. This model can be seen as a union of three influential frameworks in cognitive science: rational analysis (Anderson, 1990), the two-process model of metacognition (Nelson & Narens, 1990), and reinforcement learning (Dayan & Daw, 2008). Concretely, we characterized a rational metamemory system as the optimal policy for a Markov decision process in which a metalevel agent monitors and controls its object-level environment in order to maximize reward. Although here we have focused on metamemory, we are optimistic that this approach could be applied to model metacognition in other domains as well. We believe that dynamic metacognitive processes such as the one studied here are a critical, but understudied feature of human cognition. We thus hope that our work will encourage other researchers to further develop a rich, computational understanding of these important processes.

4.4 Methods

All data and code supporting this chapter can be found at `https://github.com/fredcallaway/memory`.

4.4.1 Experiment 1

All sample sizes, exclusion criteria, statistical analyses, modeling procedures, and plotting decisions were pre-registered (`https://aspredicted.org/wr9ej.pdf`). After pre-registering, we discovered a conceptual error in our specification of a null model. This led us to remove one plot that we discovered the more flexible null model could capture (indicating that the

plot was not actually a good test of rational metamemory). See Appendix B.1 for details, including the removed plot. Additionally, we previously ran a pre-registered version of this experiment with a smaller sample size and a slightly different analysis (which produced a marginally significant result). See Appendix B.2 for details, including full results with the previous dataset.

Participants

We recruited 612 participants through Prolific with the restriction that they reported current U.S. or U.K. residence, had at least a 95% approval rating, and had not participated in any pilot studies. As pre-registered, we excluded 106 (17%) participants who did not provide a response on more than 90% of critical trials. This yielded 506 participants in our final analysis. The target sample size of 500 participants had over 95% power based on a bootstrapping power analysis conducted on pilot data.

Stimuli

Each participant was randomly assigned 40 images and words, which were arbitrarily paired. The images were randomly sampled from 40 common scene categories (one image per category), selected from the Scene UNderstanding (SUN) database (Xiao et al. 2010). We manually removed photos that contained a person. All images were resized and then cropped to 300 by 300 pixels. The words were selected randomly from those used in Madan (2021), which were themselves selected from the University of South Florida free association norms word database (Nelson et al., 2004).

Procedure

The experiment consisted of four phases: exposure, distractor, pretest, and critical. After learning the mapping between images and words through a single round of passive exposure, participants solved simple arithmetic problems to clear working memory. They then completed the pretest and critical trials, both of which involved cued recall. In the pretest trials, participants were given an image and asked to type in the corresponding word; they were incentivized to be both accurate and fast. These trials provide an objective measure of how well each participant had learned each pair. In the critical trials, we increased the speed incentive and added an error penalty. However, we also allowed participants to skip a trial without penalty, still earning the speed bonus. This creates an incentive to quickly identify

trials in which the target is unlikely to be correctly recalled. We provide further details on the procedure for each of these components below.

Exposure On each exposure trial, participants viewed a word superimposed on the center of an image; the word was printed in white font with black outlines such that it would be clearly legible on any image. The pair was shown for two seconds, with a half-second intertrial interval. Each of the 40 pairs was shown once.

Distractor On each distractor trial, a simple arithmetic problem was presented and participants had three seconds to enter the correct answer. Each problem was an addition of three single-digit numbers. After a response (or timeout), feedback was presented for at least one second. If a response was made before the deadline, the feedback phase was extended such that all trials lasted exactly four seconds. Participants were informed that they would earn one cent for each correct answer. Due to a programming error, participants were incorrectly instructed that they would have five seconds to enter a response; however, no participant reported noticing this discrepancy in the debriefing survey.

Pretest Each pretest trial began with a blank screen and the text "press space when ready." When the participant pressed space, an image, text box, and timer appeared. The timer immediately began counting down from 15 seconds. The trial ended when the participant entered a word (by typing it into the text box and pressing enter) or when the timer expired. If the timer expired, "Timeout" appeared in large red letters. No other trial-by-trial feedback was provided. Participants were instructed that they would receive one cent for each correct answer, as well as a small extra bonus for answering quickly and correctly. (The time bonus was a quarter of a cent multiplied by the proportion of time left when a response was given). At the end of each block, participants received a summary of their performance, separately indicating the amount of bonus money they made from correct responses and from response speed. There were two blocks, and each image was shown once in each block, for a total of 80 trials. The first trial in the first block was a practice trial and was excluded from analysis.

Critical trials The critical trials were similar to the pretest trials, but with a different incentive scheme. The bonus for correct responses was increased to three cents, but a one-cent penalty for incorrect responses was introduced. Additionally, participants could skip a trial by pressing enter without typing a word; this did not incur a penalty. Finally, the speed

incentive was raised to a tenth of a cent for each second left on the timer (i.e., up to 1.5 cents per trial). The speed bonus was given on all trials, including skip and error trials. To ensure that participants understood the incentives, they were required to pass a quiz, affirming that there was a penalty for mistakes, no penalty for skipping, and a time bonus regardless of response type. Participants were additionally encouraged to quickly skip trials for which they didn't know the word.

In order to eliminate typing-related variance in response time, we defined RT as the time between stimulus presentation and the first key press initiating the response. If the input box was ever cleared (presumably because the participant changed their mind about which word to enter) response initiation time is defined as the last key press when the text box was empty (i.e., the beginning of typing the final response). For skip trials, (indicated by submitting an empty response) we use the time the final response was made, ignoring any earlier typing (of which there was usually none).

After a response was given, a metamemory judgment was elicited. When participants gave a response, they were asked "how confident are you in your response?" They then pressed a number between 1 and 5 to indicate that they were "not at all sure", "not so sure", "more or less sure", "nearly sure", or "absolutely sure" their response was correct. If they did not give a response (i.e., they skipped the trial), they were asked "how much do you feel that you know the word?", again pressing a number between 1 and 5. The responses were described as: "I am absolutely sure I do not know the word', "I am rather sure I do not know the word', "I have a vague impression I know the word', "I am rather sure I know the word' and "I am absolutely sure I know the word."

Modeling

Computing the optimal policy We compute the optimal policy by backwards induction. See Puterman (2014) for a general description of the method; here, we report the details necessary to apply the method to our model.

Recall that a mental state in the model is a tuple (t, z_t). Because backwards induction can only be applied in finite state spaces, we begin by discretizing the progress dimension into 100 equally sized bins, ranging from $-\theta$ to θ. Note that θ is the maximum possible value z_t can take. The lower bound of $-\theta$ is an arbitrary choice; we found that the optimal policy for well-fitting parameters always terminated well before this value was reached (e.g., Figure 4.3), suggesting that this imposed lower bound did not meaningfully affect the solution.

We first computed the transition function. To account for the discretization, we computed the probability of transitioning from (t, z_t) to $(t+1, z_{t+1})$ as $\Pr(a < z_{t+1} < b \mid t, z_t)$

where a and b are the boundaries of the bin with z_{t+1} in the center. Because $z_{t+1}|t, z_t$ is Gaussian (Equation 4.5), we could compute this quantity with standard statistical library functions (the Normal CDF). For most bins, the boundaries were $z_{t+1} \pm \theta/100$. The top bin was clipped at $b = \theta$ and the bottom bin was unbounded, with $a = -\infty$. This ensures that the transition probability from each state sums to one.

Next, we initialized the value function for all terminal mental states. The value of states with $z_t = \theta$ is $U(\text{recall})$ and the value of states with $t = 150$ (the maximum trial duration) but $z_t < \theta$ is 0. Then, we iterated backward in time, computing the value of all states with $t = 149$ as the maximum of the expected value of each possible computation,

$$V^*(m) = \max_{c \in \{\text{SEARCH}, \perp\}} Q^*(m, c) \tag{4.9}$$

where

$$Q^*(m, \text{SEARCH}) = \sum_{z_{t+1}} p(z_{t+1}|t, z_t) V^*(t+1, z_{t+1}) - \gamma_{\text{sample}} \tag{4.10}$$

and $Q^*(m, \perp) = 0$. The iteration continues with $t = 148$ down to $t = 1$. After computing Q^* for all mental states and computations, the optimal computation in each state can be quickly identified as

$$\pi^*(m) = \operatorname*{argmax}_{c \in \{\text{SEARCH}, \perp\}} Q^*(m, c) \tag{4.11}$$

SIMULATION PROCEDURE In order to compare the behavior of the model with that of our participants, we simulated experimental data. Simulating a trial corresponds to executing one "episode" of the metalevel MDP. That is, we initialized the state at $m_0 = (t = 0, z_t = 0)$ and then repeatedly applied Equation 4.1 to generate a sequence of states.[6] At each time step, we first checked if the recall threshold has been exceeded, i.e. if $z_t > \theta$. If so, the episode ended and the trial was classified as a recall trial. Otherwise, we determined the optimal computation $\pi^*(m_t)$, defined in Equation 4.8. If the optimal computation was $\perp$, then the episode ended and the trial was classified as a skip trial. Otherwise, we repeated the process, unless the maximum time step had been reached, in which case the episode ended as a skip trial.

The simulated response time was determined based on the final value of t. We assumed that response times reflected both time spent actively searching memory as well as "non-decision time" spent on e.g., perceptually encoding the cue and preparing the motor re-

[6] Note that we simulate data conditional on a known strength v rather than with the transition function that marginalizes over v. This allows us to model multiple trials for a single cue/target pair, as described below.

sponse. For search time, we assumed that each time step took a fixed amount of time, a value we arbitrarily set to 100ms (the predictions of the model do not depend critically on this parameter; we chose 100ms to balance prediction fidelity with model runtime). For non-decision time, we assumed that it was drawn separately for each trial from a Gamma distribution, with parameters fit to data as described below. The simulated response time was the sum of the two components. Note that, for computational reasons, we did not factor the non-decision time into the timeout condition (the maximum timestep of 150 is the maximum trial duration of 15s divided by 100ms). This has a negligible effect on model predictions because timeouts were rare (less than 0.01% of trials) with well-fitting parameter values.

The simulated metamemory judgments (confidence and feeling of knowing) were determined based on the posterior mean μ_t at the final time step, that is, when the target was recalled or the policy terminated search. To account for factors contributing to the judgment besides those captured by our model (e.g., individual differences in scale usage) we first corrupted μ_t with Gaussian noise, arbitrarily setting the variance to $\sigma_x/2$. We then binned the continuous measure into five bins, corresponding to the 1-5 response scale. We set the bin boundaries separately for each judgment type in order to match the proportion of each response in the human data.

In order to capture the relationship between performance in the pretest and critical trials, we simulated both phases using the following procedure. For each simulated word/image pair, we sampled its memory strength ν from the prior distribution Normal $\left(\mu_0, \sigma_0^2\right)$. The parameters of the prior are free parameters of the model. Next, we simulated the two pretest trials for that pair by rolling out two episodes of the metalevel MDP. For these trials, we set $U(\text{recall})$ to the experimentally imposed value of one cent. The search cost γ_{sample} is a free parameter. We then simulated the critical trial for the pair, setting $U(\text{recall})$ to the new value of three cents and increasing γ_{sample} by the experimentally imposed value of 0.01 cents per sample (0.1 cents per second and 100ms per sample).

Parameter estimation The model's behavior is governed by six free parameters: the prior mean and standard deviation, μ_0 and σ_0, the progress noise σ_x, the search cost γ_{sample}, and the mean and shape of the non-decision time (NDT) distribution, μ_{NDT} and α_{NDT}.[7] We set these parameters by maximizing the likelihood of the critical trials at the group level. For fitting, we disregarded the metamemory judgment. Each trial (human or simulated)

[7] We use this parameterization rather than the traditional shape-scale parameterization to aid interpretability. The mean is the scale parameter multiplied by the shape parameter.

was thus defined by a pretest accuracy rate (0%, 50%, or 100%), a response type (skip or recall), and a response time (discretized into 100ms bins from 0ms to 15000ms).

Because the optimal policy does not depend on the the NDT parameters, we treated these separately (described below). For the remaining four parameters, we considered 50,000 configurations sampled pseudo-randomly according to the Sobol sequence (Sobol, 1967; Bergstra & Bengio, 2012) within the range ($\mu_0 \in (-0.5, 0.5),\ \sigma_0 \in (0, 1),\ \sigma_x \in (0, 1),\ \gamma_{\text{sample}} \in (0, 0.05)$). For each configuration, we computed the optimal policy by backwards induction, and then simulated 100,000 critical trials. For each simulated dataset, we constructed a $3 \times 2 \times 151$ histogram over possible trials (3 accuracy rates, 2 response types, and 151 response time bins). To apply the NDT model, we convolved this histogram with a Gamma distribution parameterized by μ_{NDT} and α_{NDT}. Finally, to ensure non-zero probability was assigned to all trials, we mixed the model-predicted distribution with a uniform distribution with weight 10^{-6}. The likelihood of each trial in the human dataset is simply the corresponding entry of this histogram. The total log likelihood is the sum of the log-likelihood for each trial.

Note that the NDT-convolution step does not require simulating the model. We thus optimized the NDT parameters for each configuration of the other four parameters using the Nelder Mead algorithm (Nelder & Mead, 1965). This provides the best possible likelihood for each configuration of the four primary parameters. We then selected the top 5000 such configurations, and approximated the likelihood more precisely, using 1,000,000 simulated trials. The best-performing configuration from this smaller set was then identified as the maximum likelihood estimate (MLE). The MLE was ($\mu_0 = -0.002,\ \sigma_0 = 0.186,\ \sigma_x = 0.139,\ \gamma_{\text{sample}} = 0.014,\ \mu_{\text{NDT}} = 714,\ \alpha_{\text{NDT}} = 8.83$) with negative log-likelihood 70904. Note, however, that this exact number is arbitrary because it depends on the coarseness of the response time discretization.

Lesioned model without metalevel control In order to distinguish between model predictions that depend only on the object-level process vs. those that reflect adaptive metalevel control, we implemented a lesioned version of the model that lacks the control component. In this model, the decision to stop searching is no longer made by the optimal policy; instead, it is made randomly. Concretely, when simulating a trial from this model, we begin by sampling the stopping time from a Gamma distribution. We then apply Equation 4.1 until either (1) the threshold is crossed, resulting in a correct trial as in the optimal model, or (2) the pre-determined stopping time is reached, resulting in a skip trial. This model has all the parameters of the main model except γ_{sample}, which only influences the

optimal policy. It has two additional parameters for the stopping time distribution.

The parameters are fit by maximum likelihood estimation using the procedure described above. The MLE was ($\mu_0 = 0.401$, $\sigma_0 = 0.381$, $\sigma_x = 0.073$, $\mu_{\text{stop}} = 157$, $\alpha_{\text{stop}} = 86.77$, $\mu_{\text{NDT}} = 1349$, $\alpha_{\text{NDT}} = 2.59$) with negative-log likelihood 75005. We also considered a version of the model that samples stopping times directly from the empirical distribution. This model fit the data worse (negative log-likelihood of 77541; see Appendix B.1).

Statistical analyses

All reported regressions are linear mixed-effects models with random slopes and intercepts for each participant. We use logistic regression for binary outcome variables and linear regression otherwise. We report non-standardized regression coefficients throughout, with time in units of seconds, accuracy as a proportion (0 to 1), and judgments in the original 1-to-5 scale. Degrees of freedom are estimated using the Satterthwaite method. As pre-registered, we excluded 106 (17%) participants who skipped more than 90% of non-practice critical trials because these participants were likely not engaging seriously with the task. We additionally excluded one trial with a response time under 30ms (as planned but not explicitly pre-registered).

4.4.2 Experiment 2

All sample sizes, exclusion criteria, statistical analyses, modeling procedures, and plotting decisions were pre-registered (`https://aspredicted.org/xq9nx.pdf`). As in Experiment 1, we used a more flexible lesioned model we originally intended. Due to this change, we elected to fit the parameters of the lesioned model to data rather than using parameters from Experiment 1 as pre-registered. Furthermore, we introduced a new plot and accompanying regression to better distinguish between the models regarding the "rational commitment" prediction. See Appendix B.1 for details. As in Experiment 1, we previously ran a pre-registered version of this experiment; in this case, we reran the experiment because we discovered a conceptual flaw in the original analysis plan. See Appendix B.2 for details, including full results with the previous dataset.

Participants

We recruited 685 participants through Prolific with the restriction that they reported current U.S. or U.K. residence, had at least a 95% approval rating, and had not participated in any related studies (including pilots and Experiment 1). As pre-registered, we excluded 178

(26%) participants who failed to correctly recall a target on more than 50% of critical trials. This yielded 507 participants in our final analysis. The target sample size of 500 participants had over 95% power based on a boot-strapping power analysis conducted on pilot data.

Stimuli

The stimuli were identical to those used in Experiment 1.

Procedure

The procedure was identical to Experiment 1 with two exceptions. First, we lengthened the training phase to include two blocks of exposure (with each cue/target pair shown once) and one intervening block of cued recall that was identical to the pretest block except that each pair was shown only once. Second, the critical trials followed an entirely different design described below.

Critical trials The critical trials employed a modified form of cued recall in which two cues were presented on each trial. At the beginning of each trial, two gray occluders were displayed. Participants could temporarily remove the occluders, revealing the cue image underneath, by pressing the J and K keys. However, revealing one image would hide the other one. The 15-second timer appeared and began counting down when the first image was revealed. At any point, participants could press D or F to select one of the two cues for recall. At this point, both images were hidden, a yellow ring was drawn around the occluder for the selected image, and a text box appeared where they could enter the word associated with the selected image. Correct/incorrect feedback was provided after each response. Unlike Experiment 1, there was no penalty for errors (and hence, no skipping mechanism), no additional time incentive (besides the small incentive for fast correct answers that was also present in the pretest trials), and no collection of meta-memory judgments. Note that the lack of a skipping mechanism means that we cannot distinguish between genuine recall errors and the decision to give up on recall and enter a random guess (i.e., performing $\perp$). For this reason, we only analyze trials in which a target was correctly recalled. We decided to omit the skipping mechanism despite this drawback because we were specifically interested in the switching decisions (not the stopping decisions), and we wished to minimize task complexity.

Modeling

To generalize the model to the case with multiple memories that could be recalled, we assume that each memory is associated with its own independent object-level recall process. Furthermore, we assume that progress can be made on only one memory at a time, with the progress of the non-attended memory being held fixed. This is exactly analogous to Chapter 3.

The state is defined $m_t = (t^{\mathrm{L}}, t^{\mathrm{R}}, z_t^{\mathrm{L}}, z_t^{\mathrm{R}}, f)$, with t^{L} and t^{R} denoting the number of timesteps the "left" and "right" memory have each been attended, z_t^{L} and z_t^{R} denoting their respective progress levels, and $f \in \{\mathrm{L}, \mathrm{R}\}$ indicating which memory is currently attended. When the progress for either memory hits the threshold, the corresponding target is recalled.

There are now three computations, one to search for each memory, and $\perp$. We assume that there is some reconfiguration associated with switching. The cost function is thus exactly the same as in Chapter 3:

$$R(m_t, c_t) = -\mathrm{cost}(m_t, c_t) = -(\gamma_{\mathrm{sample}} + 1(c_t \neq f_t)\, \gamma_{\mathrm{switch}}). \tag{4.12}$$

The termination reward is modified to capture the fact that either memory can be recalled

$$R(m_t, \perp) = \begin{cases} U(\mathrm{recall}) & \text{if } \max\{z_t^{\mathrm{L}}, z_t^{\mathrm{R}}\} \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

The transition function simply applies the original transition function to the currently focused cue (which is updated based on the computation). If the left memory is attended, we have

$$T(m_{t+1}|m_t, c) = p(z_{t+1}^{\mathrm{L}} \mid t^{\mathrm{L}}, z_t^{\mathrm{L}}) \tag{4.14}$$

and similarly if the right cue is attended.

Computing the optimal policy We again computed the optimal policy by backwards induction. We applied the same discretization, and computed the transition function in the same way. Note that, because recall progresses for only one memory at a time, it is not necessary to represent the transition function over the full state space.

To compute the value functions, we began by initializing the value of terminal states to $U(\mathrm{recall})$ if either recall progress exceeded the threshold and 0 if the combined time exceeded 150. We then computed the value at previous time steps by iterating backwards. In this case, for each time step, we must consider all combinations of time spent on each item

that sum to the the time step under consideration. Additionally, we must consider three possible computations. Assuming the left cue is currently attended, their values are

$$\begin{aligned} &Q^*(m, \textsc{left}) = \\ &\quad \sum_{z^{\mathrm{L}}_{t+1}} p(z^{\mathrm{L}}_{t+1} | t^{\mathrm{L}}, z^{\mathrm{L}}_t) V^*(t^{\mathrm{L}} + 1, t^{\mathrm{R}}, z^{\mathrm{L}}_{t+1}, z^{\mathrm{R}}_t, L) - \gamma_{\text{sample}} \\ &Q^*(m, \textsc{right}) = \\ &\quad \sum_{z^{\mathrm{R}}_{t+1}} p(z^{\mathrm{R}}_{t+1} | t^{\mathrm{R}}, z^{\mathrm{R}}_t) V^*(t^{\mathrm{L}}, t^{\mathrm{R}} + 1, z^{\mathrm{L}}_t, z^{\mathrm{R}}_{t+1}, R) - \gamma_{\text{sample}} - \gamma_{\text{switch}} \end{aligned} \tag{4.15}$$

with $Q^*(m, \perp) = 0$ as before. Besides these differences, the procedure is identical to the one-memory case.

Parameter estimation Due to the high dimensionality of the data in this experiment (sequences of fixation durations), maximum likelihood estimation is computationally prohibitive. Although approximate fitting schemes are possible, given that we are not interested in the quantitative fit of the model, we instead used this as an opportunity to test the generalization capabilities of the model (c.f. Krajbich & Rangel, 2011). That is, we simply used the best-fitting parameters from Experiment 1. For the switch cost parameter, which was not present in the Experiment 1 model, we arbitrarily set $\gamma_{\text{switch}} = \gamma_{\text{sample}}$, noting that the predictions do not depend greatly on the exact value of this parameter. However, because the model must predict the duration of each fixation, the original non-decision time model is no longer appropriate. Instead, we assumed that non-decision time was added to each fixation independently. We fit the parameters of this model by maximizing the likelihood of all non-final fixation durations, assuming (for tractability) that they were independent and identically distributed. We excluded final fixations from this fitting procedure because they have different distributional properties (Figure 4.10A). The fitted NDT parameters were ($\mu_{\text{NDT}} = 612,\ \alpha_{\text{NDT}} = 2.79$).

Lesioned model without metalevel control The lesioned model is an extension of the lesioned model from Experiment 1, with an additional mechanism to determine fixation durations. As before, the stopping time was sampled from a Gamma distribution at the beginning of each trial. Similarly, at the beginning of each fixation, the switching time was sampled from a second Gamma distribution. If this time was reached before the memory was recalled or the stopping time was reached, then the model switched to attending the

other cue and sampled a new switching time.

To give the lesioned model the best chance of capturing the qualitative effects, we fit all of its parameters to the behavioral data (in contrast to the Optimal model, which uses parameters fit to Experiment 1). Computing an exact likelihood is intractable in this case; thus, we approximated the likelihood by assuming that the duration of each fixation depends only on the pretest accuracy of the fixated and non-fixated cues, and whether or not the fixation is final. Given this assumption, we estimated the likelihood in the same way as for Experiment 1, with the exceptions that the histogram had size $3 \times 3 \times 2 \times 151$ (3 accuracy rates for each cue, final vs. non final, and 151 response time bins) and that the likelihood was computed per-fixation rather than per-trial. Note that we constructed the histogram using only correct simulated trials, as we exclude error trials from the human data. The MLE was ($\mu_0 = 0.083,\ \sigma_0 = 0.103,\ \sigma_x = 0.148,\ \mu_{\text{stop}} = 4527,\ \alpha_{\text{stop}} = 76.16,\ \mu_{\text{switch}} = 4943,\ \alpha_{\text{switch}} = 0.18,\ \mu_{\text{NDT}} = 813,\ \alpha_{\text{NDT}} = 3.56,$).

Half the variations which are calculated in a tournament game turn out to be completely superfluous. Unfortunately, no one knows in advance which half.

Jan Timman

5

Planning

Rational use of cognitive resources in human planning[1]

One of the hallmarks of human intelligence is our ability to act adaptively in novel and complex environments. It is widely agreed that this ability depends critically on our ability to plan, that is, to use a model of the world to simulate, evaluate, and select among different possible courses of action. Research in psychology (Huys et al., 2015, 2012; Van Opheusden et al., 2017; MacGregor et al., 2001; Keramati et al., 2016; Krusche et al., 2018; Snider et al., 2015), economics (Von Neumann & Morgenstern, 1944; Stahl & Wilson, 1994; Camerer & Ho, 2004) and computer science (Newell & Simon, 1956) has formalized planning as search over a "decision tree", where every decision one might have to make is represented as a branching point (see Figure 5.1). In principle, one can identify the best plan by considering every possible decision point. However, traversing the full decision tree is infeasible because the size of the tree grows exponentially with the number of steps that one looks ahead.

The question of how people are able to effectively plan in the face of such formidable computational obstacles is of great interest for both researchers who wish to understand human intelligence and those who wish to recreate it (Griffiths et al., 2019). In fact, one of the earliest attempts to replicate human-like intelligence in a computer, conducted by Newell and Simon, focused on problems that require thinking multiple steps ahead (Newell & Simon, 1956; Newell et al., 1959, 1972). Even at this early stage, it was immediately recog-

[1]This chapter is based on the following paper:
Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022b). Rational use of cognitive resources in human planning. *Nat Hum Behav*, 6(8), 1112–1125.

nized that the success of human planners (and any hope for success of artificial planners) depended critically on the use of heuristics to circumvent the exponential growth of search trees. Recent work on human planning has largely followed a similar vein, proposing and testing different possible heuristics people could be using to reduce the cost of planning. For example, people might limit the depth of their search (MacGregor et al., 2001; Keramati et al., 2016; Krusche et al., 2018; Snider et al., 2015), "prune" away initially unpromising courses of action (Huys et al., 2012, 2015), or avoid planning altogether by relying on habit or "memoization" (Huys et al., 2015; Kool et al., 2017). Each of these models provides insight into how people circumvent the computational intractability of planning.

Despite these successes, the approach of postulating and testing specific heuristics faces several challenges. First, it is limited by the creativity of the researchers who must generate hypotheses about different possible heuristics people could be using. Second, it does not provide a straightforward way to predict which heuristics will be employed in new situations, or how each individual heuristic should be parameterized (e.g., How deep will someone plan in this environment? How large of a punishment will lead to a branch being pruned?) Finally, although these models are intuitively motivated as making planning more efficient, they do not provide a formal answer to the question of why people use these heuristics (Norris & Cutler, 2021).

These challenges—hypothesis generation, generalizable prediction, and functional explanation—are not unique to planning; indeed, they arise in nearly all areas of cognition. In many domains, progress in addressing these challenges has been made by analyzing optimal solutions to the problem a cognitive system is meant to solve (Marr, 1982; Anderson, 1990). This approach has generated insight into a wide range of problems, including decision-making (Savage, 1954), generalization (Tenenbaum & Griffiths, 2001), categorization (Anderson, 1991; Ashby & Alfonso-Reese, 1995), perception (Knill & Richards, 1996), and information-seeking (Oaksford & Chater, 1994; Gureckis & Markant, 2012). More recently, the notion of optimality has been extended to account not only for the demands imposed by the external environment but also the demands imposed by our own cognitive limitations (Howes et al., 2009; Lewis et al., 2014; Gershman et al., 2015; Griffiths et al., 2015; Lieder & Griffiths, 2020). This approach dates back to Simon (Simon, 1955) and has been especially useful in the domain of decision-making, where it has been used to explain both how long people deliberate (Bogacz et al., 2006; Drugowitsch et al., 2012; Tajima et al., 2016, 2019; Fudenberg et al., 2018) and also what people think about (Callaway et al., 2021; Jang et al., 2021) while making "simple" (i.e., non-sequential) choices. However, to the best of our knowledge, there has been no such analysis in the domain of planning, despite the especially critical role that

computational limitations play in this case (but c.f. Sezener et al., 2019; Mattar & Daw, 2018 for closely related efforts, which we discuss further below).

In this work, we propose an optimal model of planning under computational constraints. Drawing on the field of rational metareasoning in artificial intelligence (Matheson, 1968; Horvitz, 1987; Russell & Wefald, 1991a), we formalize planning as a sequential decision problem in which an agent executes a sequence of cognitive operations to construct a decision tree. Formalizing planning in this way allows us to identify the optimal planning strategy for a given environment as the one that maximizes the expected utility of executing the resulting plan minus the cost of each cognitive operation used to make that plan. This also provides a flexible framework for specifying heuristic planning strategies in a highly precise and composable way. Every model we consider specifies an explicit distribution over the sequence of planning operations that will be executed in any given environment.

To rigorously test the fine-grained predictions of the optimal and heuristic models, we develop a novel process-tracing paradigm that externalizes the cognitive operations underlying planning as mouse clicks, extending the widely used Mouselab paradigm (Payne, 1976) to sequential decision-making problems. In a series of four experiments, we find that our participants use planning strategies that are largely consistent with optimal planning strategies, using previously proposed heuristics when they are adaptive, but adjusting their strategies when the structure of the environment changes. However, we also find systematic deviations from optimal planning, in particular a bias towards considering states in the order in which they would be traversed (forward search). Based on these results, we conclude that human planners use highly adaptive planning strategies, but that these strategies are also shaped by additional constraints that may reflect the specific cognitive mechanisms underlying human planning.

5.1 Model

Following previous work (Huys et al., 2012, 2015; Van Opheusden et al., 2017; Sezener et al., 2019), we model planning as search over a decision tree. That is, we assume that the agent represents possible courses of actions as a tree-structured directed graph, in which nodes correspond to hypothetical future states and edges correspond to actions that bring the agent from one state to another (Figure 5.1B). By constructing such a tree and passing information about future rewards back to the root node (representing the current state), the agent can determine a sequence of actions that maximizes total reward. However, constructing the entire tree is prohibitive in large problems. How should a resource-constrained

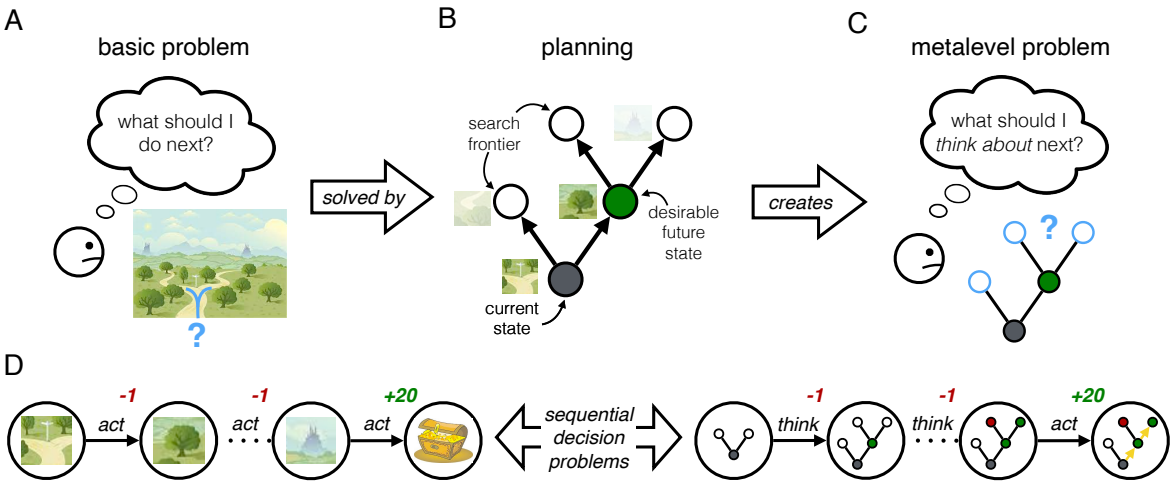

Figure 5.1: Formalizing planning under computational constraints. (A) The basic problem facing an intelligent agent is to take actions that maximize long-run reward. If the agent can predict the consequences of their actions, they can solve this problem by planning. (B) In one version of planning, the agent constructs a decision tree where nodes (circles) represent possible future states and edges (arrows) represent possible actions the agent could take. The agent constructs the tree by iteratively considering possible future states, estimating the reward to be gained there, and expanding the search frontier to include states that could be visited next. Eventually, this procedure will reveal the sequence of actions that maximize reward. But for an agent with limited cognitive resources, exploring the entire tree is usually infeasible. This creates the metalevel problem: (C) Which states should the agent consider—or ignore—in order to achieve the best tradeoff between the costs and benefits of planning? (D) The key observation underlying our model is that the basic problem and the metalevel problem are both sequential decision problems. That is, they require the agent to make a sequence of choices, in which the outcome of each choice depends on which choices were made previously.

agent plan in such a setting?

One intuitive way to conceptualize the problem of resource-constrained planning is in terms of a cost-benefit tradeoff (Daw et al., 2005; Keramati et al., 2011; Shenhav et al., 2013; Kool et al., 2017; Kool & Botvinick, 2018) in which an agent must find an optimal balance between the mental effort or time spent planning and the quality of the resulting decision. This type of model predicts, for example, that people will reduce the depth of planning under time pressure (Keramati et al., 2016). However, this one-dimensional simplification cannot capture the full range of different planning strategies people might employ. In particular, a planning strategy specifies not only the amount but also the direction of planning, that is, which courses of action are explored deeply and which are hardly considered at all (Sezener et al., 2019). To further complicate matters, it is not sufficient (or perhaps even possible) to determine in advance the amount and direction of planning. An adaptive planning strategy will dynamically adjust both based on the partial results of previous planning; for

example, one can only prune away a branch of a decision tree after discovering a large punishment early on that branch (Huys et al., 2012).

To summarize, the problem of planning involves balancing between costs and rewards attained at different time points, by determining in which direction to plan (or to stop planning) based on the outcome of previous planning. That is, in addition to being a method for solving sequential decision problems, planning is itself a sequential decision problem (Figure 5.1). This is exactly the insight captured by the metalevel MDP framework. Below, we define a metalevel MDP model of decision-tree search.

5.1.1 Metalevel Markov decision process

To characterize optimal resource-constrained planning, we cast decision-tree search as a metalevel MDP in which the mental states correspond to partially constructed decision trees and the computations correspond to expanding a node in the tree. We detail the five components of the metalevel MDP below.

A note on terminology So far, we have used the term "state" to refer to states in the *object-level* MDP, that is, the one that defines the planning problem itself. For example, in Figure 5.1, the object-level states are physical locations. Critically, the object-level states are different from the world state (defined below). When it is clear from context, we will continue to use "state" to refer to the object-level states. We will use the term "object-level reward" to refer to the rewards associated with the object-level MDP.

World states The world state defines the object-level reward that the agent would receive in each object-level state. Following Chapter 3, we denote the reward associated with object-level state i as $u^{(i)}$.[2] We assume that the object-level environment has n states; the world state is thus represented as a vector of length n.

Mental state The mental state m corresponds to a partially constructed decision tree. We make the simplifying assumptions that the external environment is itself tree-structured and known to the agent.[3] Thus, the largest possible decision tree has the same graphical

[2] This notion of world state may be counter-intuitive at first, as it is really defining the object-level reward function. Note, however, that this is the natural generalization of the world state in Chapter 3 (which specified the utility of each item in the choice set) to a sequential object-level problem.

[3] This is really quite a simplifying assumption, but also a very challenging one to not make. Hay (2016, Section 5.2) presents a sophisticated recursive method for representing search trees of arbitrary size. Such a method could be used to create cognitive models that do not make this assumption.

structure as the environment itself. The mental state can thus be represented as a vector of length n where each position corresponds to a node in the decision tree. The values $m^{(i)}$ specify either the reward that can be attained at the object-level state i, or a special value $\oslash$, which indicates that the corresponding node has not been expanded yet. In the initial mental state, only the root node (the initial object-level state) has been expanded, always having value 0; all other nodes have value $\oslash$.

COMPUTATIONS A computation c corresponds to expanding a node of the decision tree. This operation determines the cost or reward for visiting an object-level state and integrates that value into the total value of the path leading to that state. There is thus one computation $c^{(i)}$ to expand each node i. In standard decision-tree search, one can only expand nodes that are connected to nodes one has already expanded, the *search frontier*. That is, one can only consider actions in states that are already explicitly represented in the tree. Formally, we define

$$\text{frontier}(m) = \{c^{(i)} \mid m^{(i)} = \oslash \wedge m^{(\text{parent}(i))} \neq \oslash\} \tag{5.1}$$

and limit the set of allowable computations in mental state m to frontier(m). Here, parent(i) is the parent node of the expanded node, corresponding to the state from which the newly considered state can be reached. Note that, for ease of notation, we have defined frontier(m) as the set of allowable computations rather than the nodes themselves.

TRANSITION FUNCTION The metalevel transition function specifies the effect of node expansion on the decision tree. If $c^{(i)}$ is executed in m_t, the resulting mental state is identical to m_t except that the entry for node i is set to the true object-level at node i:

$$m_{t+1}^{(i)} = u^{(i)}. \tag{5.2}$$

The marginal transition function is identical, except that $u^{(i)}$ is sampled from a node-specific distribution $U^{(i)}$ capturing the agent's prior knowledge about the distribution of rewards in the environment. This distribution is a key aspect of the environment that we will manipulate in Experiments 2 and 3.

REWARD FUNCTION The metalevel reward function captures both the cost of node expansion and the quality of the plan that is ultimately executed. We assume that node expansion has a fixed cost, $R(m, c) = \gamma$. To capture plan quality, the reward for the termination operation is the value of the external rewards one will attain while executing the chosen plan

(a sequence of object-level states). We assume that the plan is selected optimally given the current decision tree. For this model, we directly specify the marginal termination reward, which is the maximum expected value of any complete plan[4] (i.e., one ending in a terminal state) given the current decision tree:

$$R(m, \perp) = \max_{p \in \mathcal{P}} V(m, p), \tag{5.3}$$

where $\mathcal{P}$ is the set of possible complete plans (all possible trajectories from the initial state to a terminal state) and V specifies the expected value of a plan:

$$V(m, p) = \sum_{i \in p} \begin{cases} \mathrm{E}[U^{(i)}] & \text{if } m^{(i)} = \oslash \\ m^{(i)} & \text{otherwise.} \end{cases} \tag{5.4}$$

5.1.2 Optimal and heuristic policies

We have now specified all four components of a metalevel MDP for decision-tree planning. However, there are countless possible planning algorithms consistent with this general class. To create a complete model, we must specify one additional component: the strategy one uses to select which nodes to expand, and when to stop expanding nodes. Formally, this corresponds to a policy for the metalevel MDP, a distribution over computations in each possible mental state.

One policy of particular interest is the optimal policy, that is, the one that maximizes the expected total metalevel reward. On a given trial, the total metalevel reward is the external reward attained by executing the chosen plan minus the cost of the node expansions used to construct the plan. The optimal policy thus balances the costs and benefits of search, expanding the nodes that are most likely to improve one's ultimate decision, and only doing so when the expected improvement in decision quality outweighs the cost of expansion. Importantly, the optimal balance depends on the cost of node expansion; the optimal model's behavior is thus governed by one key free parameter (not including parameters of the noise/lapse model used to fit human data; see Section 5.4.5).

[4] Note that a complete plan generally cannot be computed given a partial decision tree. Our assumption of maximizing over the expected value of complete plans corresponds to the assumption that, when the agent reaches a frontier node, they fall back on a default policy that is sensitive only to the expected reward at different states. In Experiments 1-3, this expected reward at all states is zero, and so it can be ignored. In Experiment 4, the expected reward is strictly negative but constant across states, so it corresponds to following the shortest path to the terminal state. This assumption is consistent with the broader assumption that the agent has direct knowledge of the transition structure and reward distributions. Addressing the more realistic case in which these are not known is an important direction for future research.

Early work in rational metareasoning proposed that optimal metalevel policies can be approximated by a myopic one-step lookahead (Russell & Wefald, 1991a; Section 2.7.2). The myopic policy chooses the planning operation that would be most helpful if the agent had to select a plan immediately afterward. Like the optimal model, this model has one key free parameter, the cost of node expansion.

We additionally consider "heuristic" policies based on three classical planning algorithms (Russell & Norvig, 2002). Breadth-first search first considers all immediate successors of the current state, then the successors of those states, and so on. That is, it prioritizes nodes that are close to the initial state. In contrast, depth-first search constructs a full plan to a terminal state before considering any alternative; it prioritizes nodes that are far from the current state. Finally, best-first search prioritizes nodes on promising paths, that is, nodes that lie on the frontier of plans with high expected value.

These classical algorithms specify the order in which nodes are expanded, but are agnostic about how people might decide when to stop planning. Previous research has proposed a number of heuristics people might use to reduce the amount of planning they must do to reach a decision. We consider four such heuristics. The "satisficing" heuristic terminates planning as soon as it finds a path whose expected value exceeds some predefined threshold (Simon, 1955). The "best vs. next" heuristic terminates planning when one path's expected value is sufficiently greater than any other path's (Solway & Botvinick, 2015). As discussed below, these two terms respectively correspond to absolute and relative stopping rules in evidence accumulation models. The pruning heuristic stops considering paths once their value falls below a predefined threshold (Huys et al., 2012). The "depth limit" heuristic only considers states that can be reached in some predefined number of steps (MacGregor et al., 2001; Keramati et al., 2016; Krusche et al., 2018; Snider et al., 2015). For brevity, we will refer to these heuristic mechanisms for limiting the amount of planning as simply "heuristic mechanisms". We assume that people could use any combination of these four mechanisms, resulting in $3 \times 2^4 = 48$ heuristic planning models (three search orders and sixteen combinations of heuristic mechanisms for each). The heuristic models have between 3 and 9 parameters depending on which mechanisms are included (see Section 5.4.5).

5.2 Results

All the models we consider make precise predictions about the exact sequence of node expansion operations a person will execute while planning. The ideal way to test these predictions would be to compare them directly to the node expansion operations performed

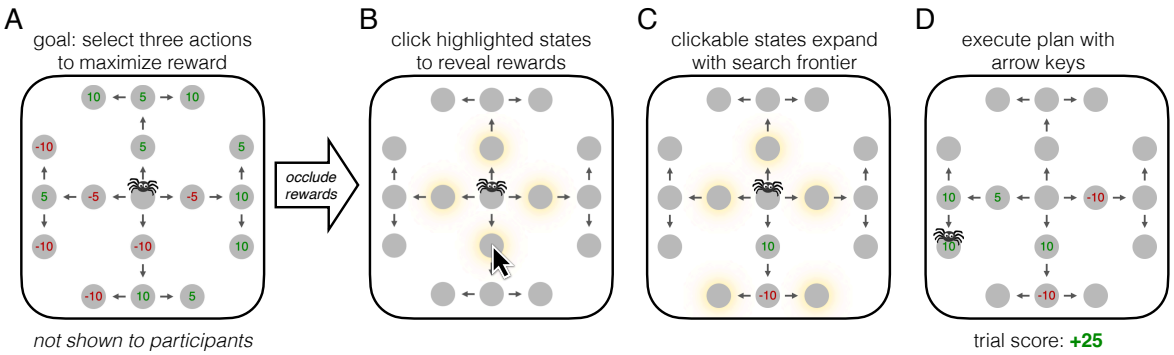

Figure 5.2: Experimental task. (A) Participants are presented with a sequential decision problem displayed as a graph. Gray circles indicate states, arrows indicate actions, and green and red numbers indicate rewards and punishments. (B) Rewards are initially occluded, but can be revealed by clicking on the corresponding state. Only highlighted states can be clicked. (C) The clickable states expand with the search frontier, which includes all states adjacent to either the initial state or an already-clicked state. (D) At any point, participants can execute a plan by pressing a sequence of three arrow keys.

by people. Unfortunately, this is impossible because those operations are internal and unobservable. Early work on human planning addressed this challenge using "think aloud" protocols in which participants narrate their planning process (De Groot, 1965; Newell et al., 1972; Chase & Simon, 1973). However, verbal reports are only indirectly related to the cognitive operations involved in planning and do not lend themselves well to precise quantitative modeling. More recently, researchers have tried to infer people's planning algorithms based only on their external actions (Huys et al., 2012, 2015; Daw et al., 2005; Solway & Botvinick, 2015; Snider et al., 2015; Van Opheusden et al., 2017). However, the precise nature of a person's planning algorithm is generally only weakly constrained by their actions alone, because there are usually many sequences of planning operations that are consistent with each possible choice.

How can we collect fine-grained and precise data on human planning processes? A similar problem faced researchers studying how people make non-sequential decisions. To address this challenge, Payne and colleagues developed the Mouselab paradigm (Payne, 1976; Payne et al., 1988), which traces participants' decision-making processes by requiring them to click to reveal decision-relevant information. In the original paradigm, participants clicked on cells in a table to reveal the payoffs associated with different outcomes of risky gambles. Here, we apply the same idea to multi-step decision problems, with participants clicking to reveal rewards at hypothetical future states.

5.2.1 Mouselab-MDP

In order to directly compare our model predictions with human behavior, we developed a task, "Mouselab-MDP", which makes human planning observable. The task is illustrated in Figure 5.2. On each trial, participants are presented with a route-planning problem, displayed as a graph. Each vertex in the graph (gray circles) corresponds to a future state the participant could visit, and harbors a reward or punishment (-10, -5, $+5$, or $+10$ with equal probability). The edges in the graph correspond to actions the participant can take to travel between states. The goal is to select a sequence of three actions that maximize the total reward. The potential gains and losses are initially occluded, but the participant can reveal them by clicking on the corresponding state, with the constraint that they can only click on states adjacent to the initial state or a previously revealed state. This constraint ensures that participants follow a forward-planning strategy, as has often been assumed in the literature (Huys et al., 2015, 2012; Van Opheusden et al., 2017; MacGregor et al., 2001; Keramati et al., 2016; Krusche et al., 2018; Snider et al., 2015); we remove the constraint in Experiment 3. Each click was followed by a three-second delay.

Importantly, the task involves two types of sequential decision problems, both of which can be modeled as MDPs. The problem of moving the spider in the web is modeled as an MDP with 17 states (gray circles), four actions (key presses), and four possible rewards (-10, -5, $+5$, and $+10$). In contrast, the problem of selecting which potential rewards to consider when planning a route is modeled as a *metalevel* MDP, with over four billion possible states (patterns of revealed rewards), 16 actions (one for revealing each reward), and fourteen possible rewards (one implicit cost for the delay and thirteen possible path values, i.e., -30 to 30 in steps of 5).

Like its predecessor, Mouselab-MDP externalizes the core representations and operations underlying a cognitive process. In particular, our paradigm externalizes the decision tree as the graphical display, the node expansion operation as clicking, and the cognitive cost of that operation as the delay. While it is possible that externalizing a cognitive process in this way might alter the strategy people adopt, the extensive use of the original Mouselab paradigm (Payne et al., 1988; Ford et al., 1989; Payne et al., 1993; Gabaix et al., 2006; Schulte-Mecklenbeck et al., 2011) and the early advances made possible by a less structured form of process tracing (De Groot, 1965; Newell et al., 1972; Chase & Simon, 1973) provide support for using this approach. We return to this point in the Discussion.

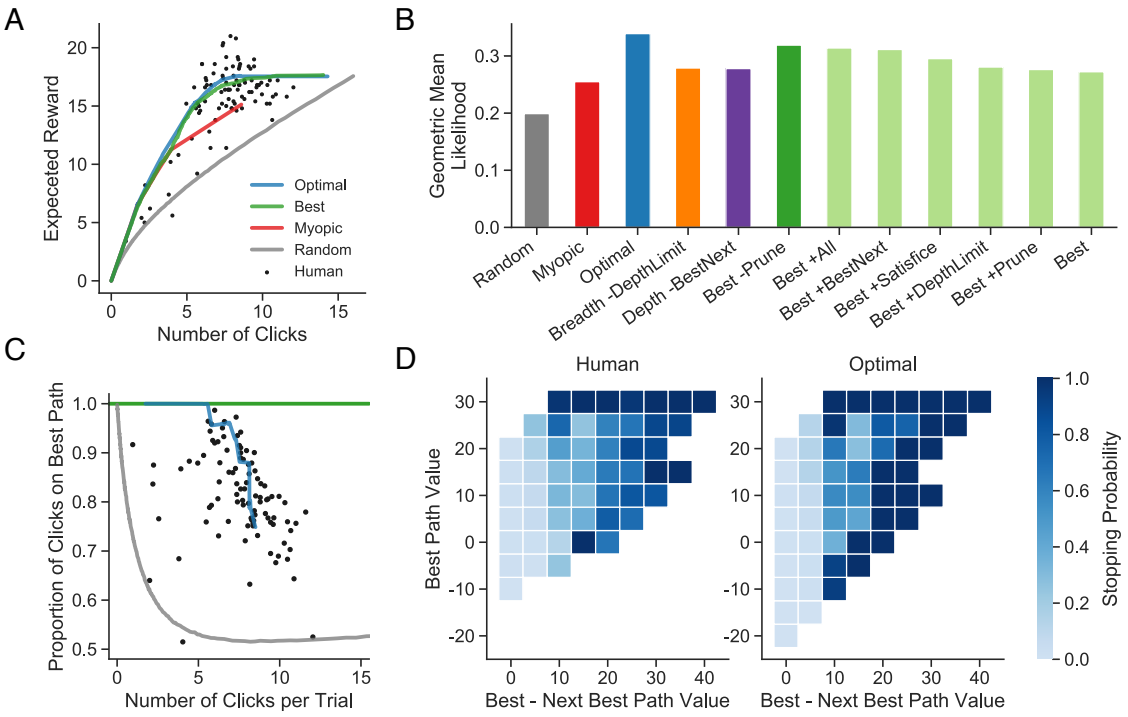

Figure 5.3: Experiment 1 results. (A) Pareto curves. Each point shows the average reward attained and number of clicks made by a participant (black dots) or model (colored lines). Note that with a small number of trials, it is possible to exceed the expected performance of the optimal model by getting lucky. (B) Model comparison. Bars show geometric mean likelihood (the total log-likelihood divided by the number of observations and then exponentiated) estimated on out-of-sample data. For the heuristic models, we indicate which heuristic components are present: +All indicates that all mechanisms are included, -Prune indicates that all mechanisms except for pruning are included. The best-fitting versions each heuristic model are shown in dark bars. Alternative best-first search models are shown in light green. Note that any visually detectable difference corresponds to a large difference in likelihood. (C) Selection rule. Proportion of clicks following a best-first strategy as a function of the average number of clicks per trial for each participant. Colors match panels A and B. Model predictions are made without fitted noise parameters. (D) Stopping rule. Probability that planning is terminated as a function of the value of the best path found yet and the difference in values of the best and next best paths. The right panel shows simulations from the noise-free optimal model. Cases in which all nodes have been clicked and termination is required are excluded.

5.2.2 Experiment 1: Comparing human and optimal planning algorithms

In our first experiment, we sought to test the extent to which human planning is consistent with an optimal planning strategy in a relatively unstructured environment, illustrated in Figure 5.2A.

Overall performance

To evaluate participants' performance, we must consider both the scores they achieved as well as the amount of planning effort (i.e., clicking) that they expended. Figure 5.3A thus shows the average reward and number of clicks each participant made per trial. The blue line shows the Pareto front, the maximum average reward attainable for a given average number of clicks. On average, participants earned 0.92 fewer points than they could have with the same number of clicks. They earned 4.94 more than clicking randomly (95% CI [4.43, 5.44]; Wilcoxon test: $z = 8.40, p < .001$). Confidence intervals are boot-strapped over participants and p-values are two-tailed (see Section 5.4.9).

Selection rule: cost-dependent best-first

We first considered the order in which the model expands nodes. Inspecting simulations of the optimal planning strategy across a range of costs (0.05 to 3.75, the maximum cost for which any planning occurs), we found that the optimal model expands a node on a path that has maximal expected value between 74.6% and 100% of the time, compared to 51.7% in the random clicking model. That is, optimal planning in this environment resembles best-first search. Consistent with this prediction, participants expanded a path with maximal expected value on average 81.5% of the time (95% CI [79.6, 83.3]; Wilcoxon test vs. chance: $z = 8.46, p < .001$).

However, the degree to which optimal planning conforms to best-first search depends on the cost parameter, with a closer match for higher costs. Intuitively, this is because the optimal planning policy expands nodes that are likely to lead to a quick decision. When the cost is high, a plan can be chosen when it is only moderately better than its competitors; the path that currently has maximal value is the most likely candidate. When the cost is low, however, a plan must be exceptionally good to justify stopping early; a path with moderately high value is actually less likely to provide such an outcome, compared to a completely unexplored path. As a result, the optimal model predicts that the degree to which people follow best-first search will decrease with the average number of clicks they make (the most direct behavioral correlate of the cost parameter). Figure 5.3C confirms this prediction (Spearman's $\rho = -0.481$, 95% CI $[-0.66, -0.28]$, $p < .001$). The correlation also arises in the random model because all paths are "best" on the first click. However, controlling for the best-first rate of the random model, we still find a significant correlation ($\rho = -0.347$, 95% CI $[-0.56, -0.12]$, $p < .001$).

Stopping rule: both absolute and relative

By inspecting simulations of the optimal model with a range of costs matching that inferred from human participants, we found that the model was more likely to stop planning when it had found a path with high expected value, consistent with satisficing. However, its stopping decisions were more strongly influenced by the difference between the value of the best path and the next best path. That is, the optimal stopping rule depends primarily on the best path's relative value, but also on its absolute value.

As illustrated in Figure 5.3D, our participants' decisions to terminate planning were also sensitive to both the absolute and relative value of the best path. A mixed-effects logistic regression with random intercepts and slopes for each participant revealed significant effects of both terms (best path value: $\beta = 0.82$, 95% CI [0.69, 0.94], $z = 12.89$, $p < .001$; best vs. next: $\beta = 1.68$, 95% CI [1.52, 1.84], $z = 20.70$, $p < .001$). However, compared to the coefficients for the optimal model ($\beta = 0.99$, 95% CI [0.84, 1.15] and $\beta = 4.64$, 95% CI [4.02, 5.26]), people appear to be under-sensitive to relative value (note that the confidence intervals for the optimal model are not negligible due to the mixed-effects structure; predictors are standardized by their mean and SD in the human data).

These results are broadly consistent with evidence accumulation models of non-sequential decisions, where relative stopping rules (specifically best vs. next) generally perform better, both in terms of fitting data (Ratcliff & Smith, 2004; Teodorescu & Usher, 2013) and maximizing accuracy (McMillen & Holmes, 2006; Bogacz et al., 2006). However, although both the model's and our participants' stopping decisions were primarily driven by relative value, absolute value also played a role. This raises the intriguing possibility that people could be using a hybrid stopping rule in simple value-based choices as well.

Model comparison

Having characterized the qualitative matches and mismatches between participant and optimal behavior in the task, we next sought to quantify the ability of the optimal and heuristic models to predict human behavior quantitatively. We fit our models to participants at the individual level and obtained out-of-sample predictions using five-fold cross-validation. We used the total log-likelihood (LL) across all five folds as a measure of model performance. Note that this metric accounts for the flexibility of the different models without relying on parameter counting (as do AIC and BIC), which can be a poor measure of flexibility (Piantadosi, 2018). Differences in this cross-validated log-likelihood (Δ_{LL}) can be interpreted similarly to differences in AIC: $\Delta_{\text{LL}} = 1$ is roughly equivalent to $\Delta_{\text{AIC}} = 2$.

Figure 5.3B shows the predictive accuracy achieved by each of the models. The optimal model clearly outperforms the random, myopic, breadth-first, and depth-first models (all $\Delta_{LL} > 3981$). In terms of total likelihood, it also outperformed best-first search (all $\Delta_{LL} > 1250$), although 41 participants were best fit by the one of the best-first models vs. 45 by the optimal model (9 by some other model). Importantly, given that the best-first model achieved a near-optimal reward-effort trade-off (Figure 5.3A), a substantial majority of participants were best fit by an optimal or near-optimal model.

5.2.3 Experiment 2: Adapting to the environment

In Experiment 1, we found that participants seemed to use a best-first search strategy that was well-suited to the task environment. However, this does not mean that people always plan in this way. On the contrary, a key prediction of the optimal model is that people adapt their strategy to the structure of the environment. We tested this prediction in Experiment 2.

To investigate the effect of environment structure on human planning strategies, we constructed three new experimental environments (see Figure 5.4A). The environments have the same transition structure (four independent paths with five steps each) but different reward distributions. In the "constant variance" environment all states had the same reward distribution, as in Experiment 1. In the other two environments, most states had low variance; extreme rewards could only be found in one state on each path. In the "decreasing variance" environment extreme rewards were possible only in the first state on each path. In the "increasing variance" environment extreme rewards were possible only in the last state.

We designed these environments to produce clear qualitative differences in the predictions of the optimal model. Specifically, in each environment, the optimal planning strategy resembles a different classical planning algorithm: breadth-first for decreasing variance, best-first for constant variance, and depth-first for increasing variance. As illustrated in Figure 5.4B, each algorithm is approximately optimal in its respective environment, but suboptimal in the other two.

If people indeed adapt their planning strategy to the environment, we should find that, out of these three classical search models, the model that achieves the best reward-effort trade-off should also predict human behavior best. Figure 5.4C confirms this prediction (all $\Delta_{LL} > 446$). For the classical search models, we used the combination of heuristic mechanisms that achieved the best likelihood across all conditions; however, we excluded depth limits from this analysis because they allow the best-first and depth-first models to mimic breadth-first search. With the unrestricted set of heuristic models, the optimal model best

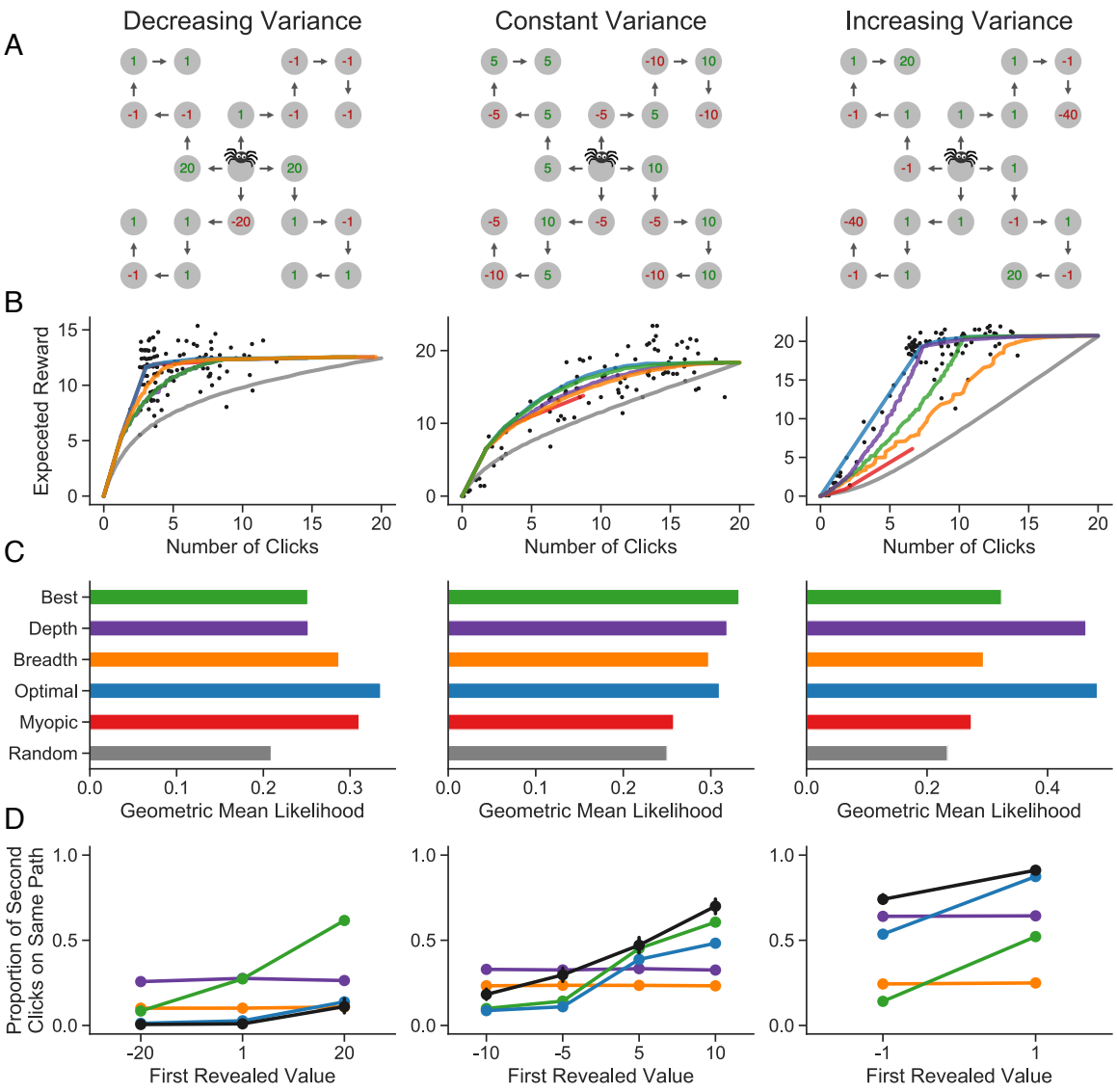

Figure 5.4: Experiment 2 results. Each column shows one experimental condition. (A) Example trials. Large values are found at the beginning of each path (decreasing variance), at any location (constant variance), or at the end of each path (increasing variance). (B) Pareto curves. Each point shows the average reward attained and number of clicks made by a participant (black dots) or model (colors match panel C). (C) Model comparison. Best, Depth, and Breadth refer to the versions of the model that performed best in Experiment 1 (excluding depth limits to prevent other Best and Depth from mimicking Breadth). Of these classical algorithms, the one that achieves the best reward-click trade off (shown in panel B) also best predicts human behavior. (D) Behavioral indicator of planning strategy. Each panel shows the probability of making a second click on the same path as the first, depending on the value revealed by that first click. In each condition, one heuristic model captures the behavioral pattern, but only the optimal model captures the behavior in all conditions. Human data is in black and model colors match panel C. All heuristic mechanisms are excluded (see Figure C.1 for the full models). For human data, points show means and error bars show bootstrapped 95% confidence intervals, both computed across participants.

predicts human behavior in the increasing ($\Delta_{LL} = 606$) and decreasing ($\Delta_{LL} = 1276$) conditions; the best-first model with best vs. next fits best in the constant condition (compared to optimal: $\Delta_{LL} = 2150$).

Figure 5.4D demonstrates the shift in planning strategy with a simple behavioral measure. Considering only trials on which at least two clicks were made, we can ask how often people use their second click to continue down the path that they began with their first, depending on the value revealed by that first click. An overall tendency to continue down the same path is consistent with a depth-first strategy, the reverse tendency is consistent with a breadth-first strategy, and high sensitivity to the revealed value is consistent with a best-first strategy; we illustrate this by plotting the predictions of the basic search models without any heuristic mechanisms. Participants in each condition show the same pattern as the adaptive search order.

5.2.4 Experiment 3: Backwards planning

In the previous experiments, we constrained participants' planning strategies to variations of decision tree search by only allowing them to click on states adjacent to the initial state or a previously-clicked state. However, people may sometimes use planning strategies that are not constrained in this way. For example, they may plan backward from a goal as in means-ends analysis (Newell et al., 1972) or they may even consider states in arbitrary order (Sutton, 1990). Experiment 3 thus investigated a broader class of possible planning algorithms by lifting the forward-planning constraint, allowing participants to click any state at any point.

As in Experiment 2, we used environments with decreasing, constant, and increasing variance. For this experiment, we employed the transition structure from Experiment 1 and decreased or increased the reward variance exponentially with depth. The constant variance condition used the same reward distribution as Experiment 1. See Figure 5.5A for examples.

The key prediction of the optimal model is that participants will adopt a backward-planning strategy in the increasing variance condition, considering terminal states first and then working towards the initial state. Consistent with this prediction, participants in this condition were most likely to click a terminal state first (Figure 5.5D, right).

However, we also see a systematic deviation from the optimal model predictions. In the constant variance case (Figure 5.5D, center), the model is completely neutral between depth-one and depth-two states because they provide equivalent information about the optimal path. In contrast, participants showed a strong tendency to click a depth-one state first. More generally, participants in the constant-variance condition showed a consistent bias

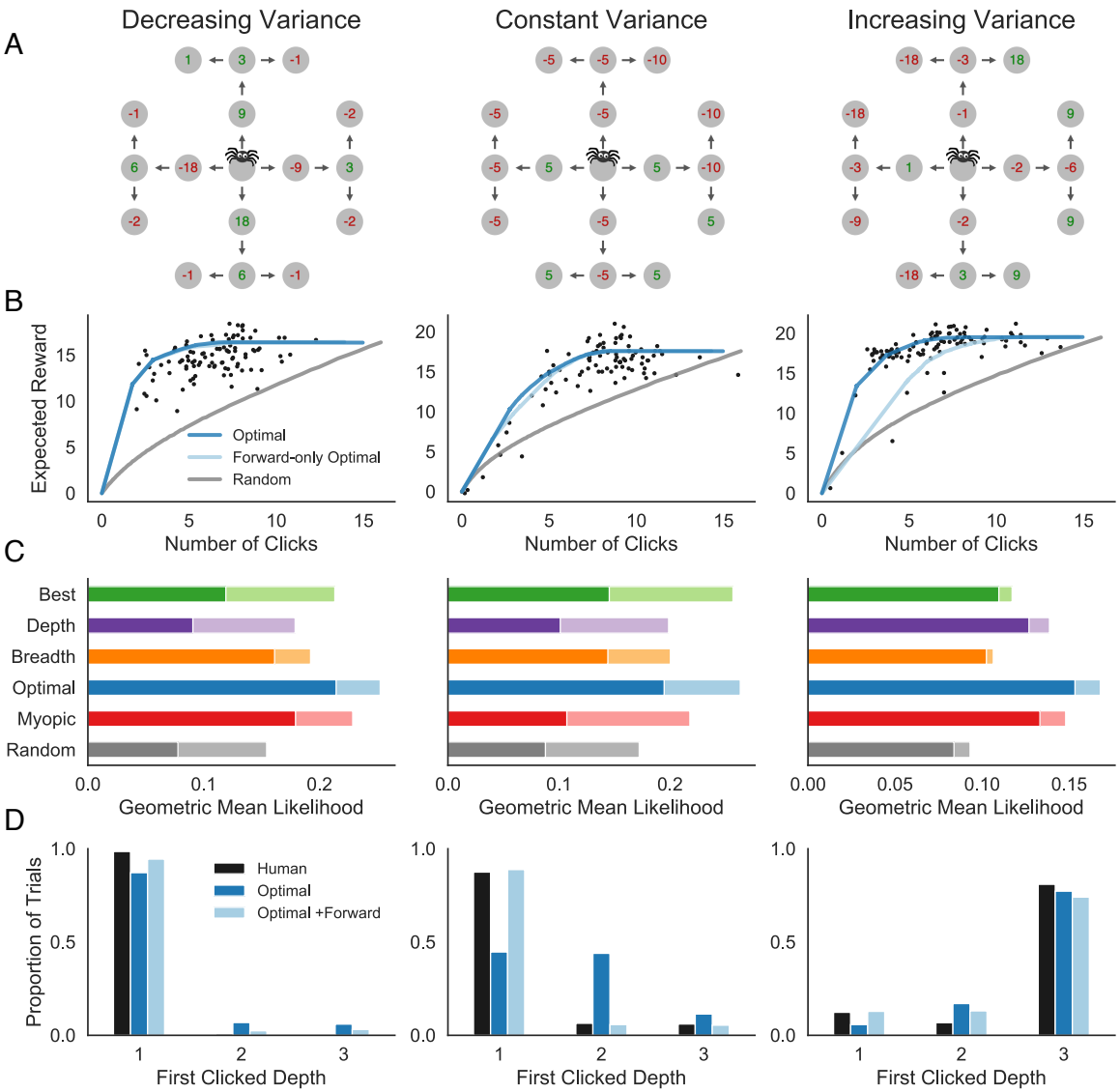

Figure 5.5: Experiment 3 results. (A) Example trials. Each condition is characterized by a different location-dependent reward distribution with standard deviation linearly increasing, decreasing, or remaining constant with depth. (B) Pareto curves. The light blue line shows the optimal model restricted to plan forwards. (C) Model comparison. Light bars show the performance of the corresponding model with a fitted degree of forward-search bias (including the no-bias model and forward-only model as special cases). (D) Behavioral indicator of forward and backward planning. Each panel shows a histogram of the depth of the first clicked state, in the data and in simulations from the optimal model with and without a forward-search bias. Although participants use forward-search by default (center), they switch to backward search when the environment encourages this strategy (right).

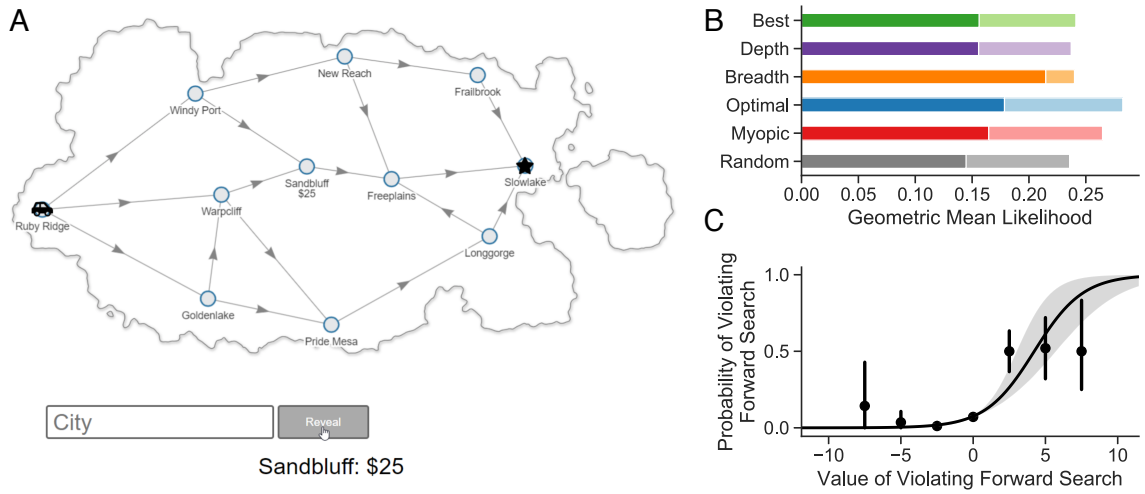

Figure 5.6: Experiment 4 results. (A) Task: participants acted as a travel agent, attempting to find a low-cost route from a start city to a goal city. They could reveal the price of passing through each city using a textual search interface. (B) Model comparison. Light bars show models augmented with a forward-search bias. (C) The probability of a participant inspecting a city without a revealed parent (i.e., violating forward search) as a function of the value of doing so. This value is defined as the maximal Q value for expanding a node not on the frontier minus the maximal Q value for expanding a node on the frontier. The line shows a logistic regression fit and points show binned means. The shaded regions and error bars show 95% confidence intervals.

for forward search, clicking a state whose parent had already been revealed 92.4% of the time compared to 75.5% in the noise-free optimal model simulations (95% CI [86.2, 94.4], Wilcoxon test vs. optimal $z = 5.32, p < .001$). Importantly, however, such a bias was not maladaptive as indicated by the strong performance of a strictly-forward planning strategy (Figure 5.5B center).

Figure 5.5B shows that augmenting the models with a forward-search bias improves predictive accuracy considerably. Whether or not we incorporate the bias, the optimal model predicted human behavior best in every condition (with bias: all $\Delta_{\text{LL}} > 509$). Note that it is not clear how to extend pruning and depth limits when non-adjacent nodes on a single path can be expanded; thus, we do not include these mechanisms for this analysis.

5.2.5 Experiment 4: Planning a road trip

In Experiment 4, we tested the ability of the optimal model to generalize to a new task environment. In this new task, illustrated in Figure 5.6A, participants acted as travel agents, planning a route from an initial city to a goal city and minimizing the price of hotels that must be visited along the way. Participants were informed that hotels could cost $25, $35,

$50, or $100 (with equal probability), but to see the actual price of the hotel in a city they had to type its name into a search box.

Although the task has the same formal structure as that used in Experiment 3 (allowing us to use the same models), there are three important dimensions on which the new task differs from the previous ones. First, rather than allowing participants to plan an arbitrary path, they were required to reach a specific destination; second, the transition structures were not limited to trees—that is, there could be multiple ways to reach a given state; third, the distribution of possible costs did not have a mean of zero, making it necessary to account for expected future cost when estimating the value of an incomplete plan. This task thus provides a non-trivial test of the model's ability to generalize.

As illustrated in Figure 5.6B, the optimal model most accurately predicted human behavior when the bias for forward search was taken into account ($\Delta_{\text{LL}} = 295$). Interestingly, the forward search bias is so important for capturing behavior that when we remove it, the breadth-first model (which follows forward search by default) performs best.

However, the tendency towards forward search was not without exception. Participants violated forward search by looking up a city without a revealed parent 7.2% of the time. Figure 5.6C shows that these exceptions were not random: participants were more likely to violate forward search when doing so was more valuable (logistic regression with random slopes and intercepts for each participant: $\beta = 2.48$, 95% CI [1.59, 3.37], $z = 5.48$, $p < .001$).

5.3 Discussion

In this paper, we proposed a rational model of resource-constrained planning and compared the predictions of the model to human behavior in a new process-tracing paradigm. Our results suggest that human planning strategies are highly adaptive in ways that previous models cannot capture. In Experiment 1, we found that the optimal planning strategy in a generic environment resembled best-first search with a relative stopping rule. Participant behavior was also consistent with such a strategy. However, the optimal planning strategy depends on the structure of the environment. Thus, in Experiments 2 and 3, we constructed six environments in which the optimal strategy resembled different classical search algorithms (best-first, breadth-first, depth-first, and backward search). In each case, participant behavior matched the environment-appropriate algorithm, as the optimal model predicted.

The idea that people use heuristics that are jointly adapted to environmental structure and computational limitations is not new. First popularized by Herbert Simon (Simon,

1955), it has more recently been championed in ecological rationality, which generally takes the approach of identifying computationally frugal heuristics that make accurate choices in certain environments (Gigerenzer, 2008; Gigerenzer & Gaissmaier, 2011; Todd & Gigerenzer, 2003; Gigerenzer & Goldstein, 1996). However, while ecological rationality explicitly rejects the notion of optimality (Gigerenzer & Todd, 1999), our approach embraces it, identifying heuristics that maximize an objective function that includes both external utility and internal cognitive cost. Supporting our approach, we found that the optimal model explained human planning behavior better than flexible combinations of previously proposed planning heuristics in seven out of the eight environments we considered (see Table C.1).

Why did the optimal model generally explain human behavior better than the heuristic models? One possibility is that the optimal model has a more sophisticated stopping rule, informed by the full distribution of possible rewards, not just the expected values of different paths. Indeed, augmenting the heuristic models with distributional variants of the best vs. next and satisficing rules improved fit substantially (see Appendix C.3). However, the optimal model still achieved a better fit in all but two cases (constant variance in Experiments 2 and 3).

The increasing variance environments in Experiments 2 and 3 provide an especially interesting test of the model. In these environments, distal rewards are more extreme than proximal ones, and so the optimal model considers these states as soon as possible. In contrast, a classic finding is that people tend to neglect long-term consequences (O'Donoghue & Rabin, 1999), suggesting that people might fail to consider those distal states in their planning. We found that people's clicking was consistent with the optimal model. In Experiment 2, they ignored small short-term losses to more quickly find large long-term rewards (Figure 5.4D), and when we lifted the forward-planning constraint in Experiment 3, people considered the final states first (Figure 5.5D). A potential reason why people were more far-sighted in our experiments than they are in some real-world situations is that our experiment allowed them to learn about the structure of the decision environment and adapt their decision strategy to it through intensive practice with immediate, reliable performance feedback that is often unavailable in the real world (Kahneman & Klein, 2009). Consistent with this, people did show a strong bias to consider proximal rewards first when the environment did not strongly incentivize a different strategy (Figure 5.5D, center).

The ways in which our participants deviated from the optimal model are equally—if not more—informative than the ways in which they were consistent (Norris & Cutler, 2021). Using the approach of resource-rational analysis (Griffiths et al., 2015; Lieder & Griffiths, 2020), we can use the observed discrepancies to generate hypotheses about additional con-

straints (internal or external) that shape human planning strategies. That is, people's cognitive resources might be more limited than the model assumes and they may be adapted to an environment that differs from our artificial experimental task in important ways.

We found the most striking deviation from the optimal model's predictions in Experiments 3 and 4, where we observed a strong bias for forward search when it was not adaptive (nor clearly maladaptive, see Figure 5.5B). This suggests that people's default representation of plans is temporally ordered, and that representing or computing information which does not fit this temporal structure is cognitively costly. There are two reasons such a representation might be preferred. First, in many (but not all) natural environments, the set of states one could feasibly reach is not clear in advance; one can only discover such states by forward search. In these cases, the standard assumption that people can only search in the forward direction (Keramati et al., 2016; Van Opheusden et al., 2017; Huys et al., 2012; Snider et al., 2015) may be appropriate. Second, in many domains, people likely have generative models of the world (Battaglia et al., 2013; Jara-Ettinger et al., 2016); given such models, one can directly simulate the consequences of an action, but one must infer what action could have led to a given consequence. In these cases, forward search will be less costly than backward search, but still possible; this is consistent with Figure 5.6C.

One important limitation of our work is that externalizing planning, as our task does, may alter the internal process that we wish to measure (Lohse & Johnson, 1996). Nevertheless, there are at least five reasons to believe that the present results already reveal something important about human planning. First, the paradigm is a direct extension of the Mouselab paradigm, which has been widely used in the multi-attribute and risky-choice literature (Payne et al., 1988; Ford et al., 1989; Payne et al., 1993; Gabaix et al., 2006; Schulte-Mecklenbeck et al., 2011). Second, our Experiment 1 results replicate previous findings that suggest that participants use a best-first strategy (Van Opheusden et al., 2017)—or, similarly, avoid nodes following large losses (Huys et al., 2012)—in the absence of environmental structure that a different algorithm could exploit. Third, we found that people show a bias for forward search even when the task does not require or even encourage it; this suggests that participants are carrying over a strategy that they have developed for naturalistic planning (where such a bias is likely adaptive, as discussed above). Fourth, recent work has noted a parallel between planning and information-seeking (our task could be characterized as the latter), suggesting that similar neural mechanisms may underlie both behaviors (Hunt et al., 2021). Finally, measuring how people plan in the absence of working memory constraints provides a useful comparison point for future work investigating how these constraints shape human planning strategies.

Comparing human and optimal planning in a more naturalistic paradigm is thus a critical step in future research. One promising approach is to use reaction time in a secondary task as a signal of previous planning (e.g., choosing between a subset of actions Ongchoco et al., 2019, replanning after a random teleportation Ho et al., 2020, or determining whether a specific state falls on the optimal path Solway et al., 2014). Another approach would be to use eye- or mouse-tracking with a display that only reveals the reward at future states, but not the transition function. However, deploying these paradigms would also require augmenting the model to account for constraints on working memory and imperfect knowledge of the transition function—important but challenging directions for future work.

A second limitation of our work is that we only consider deterministic environments. This assumption greatly simplifies the task of identifying optimal strategies; in particular, it ensures that it is optimal to do all planning before taking any actions, allowing us to avoid the complexities associated with interleaving planning and action. Although we enforced this plan-then-act structure in our main experiments, a follow-up experiment (reported in Appendix C.2) found that participants rarely violate this ordering when allowed to do so (3.9% of trials). However, in stochastic environments, planning far ahead may be wasteful because an unexpected transition can render much of that planning irrelevant. In such cases, it may be optimal to take an action and see its result before planning further ahead (discussed further in Section 6.2.3). Investigating how people adapt their planning strategies in unpredictable environments is thus an important direction for future work.

A third limitation is that we only consider problems with small, unstructured state spaces. This contrasts with early work exploring human planning in massive state spaces with rich internal structure, such as propositional logic (Newell et al., 1972). Although this limitation applies equally to most recent empirical studies of human planning, future work should explore the strategies people use to plan efficiently in more complex environments.

Taken together, these three limitations put important limits on the conclusions we can draw from our results. Although we have shown that human planning can be quite close to optimal in simple environments without working memory constraints, it remains unclear whether people will be able to plan as effectively in more complex domains when working memory is limited. Nevertheless, our results do suggest that models of efficient use of limited cognitive resources may be a good starting place when developing theories of planning in these more naturalistic conditions.

A final limitation of our work is that we do not provide a process-level theory for how people are able to approximate optimal planning. One plausible hypothesis is that people use a myopic approximation, considering the immediate value of expanding a node while

disregarding the potential for future node expansions. Indeed, such an approximation has been employed in two recently proposed models of human planning (Mattar & Daw, 2018; Sezener et al., 2019). However, we found that this model generally performed poorly, both in terms of reward (Figures 5.3A and 5.4B) and predicting human behavior. Another hypothesis is that people learn effective planning strategies through experience (Lieder & Griffiths, 2017; Krueger et al., 2017). However, the mechanisms that allow this learning to proceed so rapidly given the large state spaces of metalevel MDPs are still not well understood.

Over the past decades, the assumption that humans are well-adapted to their environment (Marr, 1982; Anderson, 1990) has facilitated rapid progress in many psychological domains (Oaksford & Chater, 1994; Gureckis & Markant, 2012; Tenenbaum & Griffiths, 2001; Anderson, 1991; Ashby & Alfonso-Reese, 1995; Savage, 1954). However, the constraints imposed by the external environment are insufficient to explain many key features of human cognition (Rahnev & Denison, 2018; Lieder & Griffiths, 2020). By additionally considering the constraints imposed by our limited cognitive resources—that is, our internal environments—we can apply the tools of rational modeling to a much broader set of cognitive phenomena (Griffiths et al., 2015; Lieder & Griffiths, 2020). In this work, we have presented the beginning of such an analysis for planning. We anticipate that more precise characterizations of the cognitive constraints that shape planning will yield a correspondingly deeper understanding of this remarkable human ability.

5.4 Methods

All experiments can be viewed exactly as they were given to participants and in abbreviated form at https://webofcash.netlify.app. All experiments were approved by the institutional review board of Princeton University, and all participants gave informed consent. Each participant could only participate in one experiment (including pilots). For all experiments, we aimed to collect 100 participants per condition. We did not conduct a formal power analysis because all our hypothesis tests were highly significant in pilot samples. All reported statistics, model comparisons, and figures were pre-registered.[5] We describe deviations from the pre-registered analysis plan in Appendix C.1.

All data and code supporting this chapter can be found at `https://github.com/fredcallaway/`

[5]Experiment 1: `https://aspredicted.org/jd8rs.pdf`
Experiment 2: `https://aspredicted.org/w4kt2.pdf`
Experiment 3: `https://aspredicted.org/2cr5k.pdf`
Experiment 4: `https://aspredicted.org/wq87z.pdf`

rational-resources-planning.

5.4.1 Experiment 1

We recruited 104 participants from Prolific who reported fluency in English, resided in the United States, and had a 95% approval rating (this number excludes participants who accepted the study but did not move past the second instruction page). We excluded 6 participants because they failed a quiz following the instructions and 3 participants who did not complete the experiment for some other reason, leaving 95 participants in the analysis. Participants who completed the experiment or failed the quiz received $1.50 for participation. Those who completed the experiment additionally earned a performance-dependent bonus of (mean $\pm$ sd) $2.43 $\pm$ $0.42 for 22.5 $\pm$ 6.6 minutes of work.

Main Task In the main task of Experiment 1 (see Figure 5.2), participants navigated a cartoon spider through a directed graph in which each vertex (the gray circles) harbored a gain or loss, with the goal of maximizing the total payoff accrued along the selected route. All rewards were independently drawn from a discrete uniform distribution over the values $\{-10, -5, +5, +10\}$. At the beginning of each trial, all rewards were occluded; however, participants could click on nodes adjacent to the starting location or to an already-revealed node to reveal the value. After each click, there was a three-second delay during which no additional clicks could be made. To visually convey these constraints, nodes were highlighted whenever they could be clicked. At any point, participants could stop clicking and move the spider from the starting node using the arrow keys. After each arrow key press, the spider moved to an adjacent node, the value of that node was revealed (if not already revealed), and its value was added to a total shown in the top right. Clicking was disabled after the first move, and the trial ended when the participant reached a terminal node (i.e., one with no outgoing edges).

Procedure The experiment began with an instruction phase in which participants completed increasingly complex versions of the task. First, they were told the basic goal of selecting paths to maximize the amount of "money" acquired, and completed three trials with the rewards fully revealed. Second, they were told the reward distribution and shown ten examples where they did not make choices. Third, they completed one trial with the rewards occluded (i.e., guessing randomly). Fourth, they were told that they could click nodes to reveal the values, and completed three trials in which they had to make at least five clicks.

Finally, they were told the conversion between in-game currency and their bonus (1 US cent for every 2 points) and completed three practice trials of the full task.

After completing the instructions, participants took a multiple-choice quiz that asked about the reward distribution, the rules for inspecting nodes, and the points-to-bonus conversion. Participants who failed the quiz were shown a review screen with all the necessary information and were given another chance to complete the quiz. If they failed the quiz three times, they were dismissed. Otherwise, they progressed to the main phase of the experiment where they completed 25 trials of the main task. They were given an initial endowment of 100 points to minimize the chance that they would ever have a negative score.

5.4.2 Experiment 2

All aspects of the design were identical to Experiment 1 except where noted otherwise. We recruited 313 participants. We excluded 4 who failed the quiz and 11 who did not complete the experiment, leaving 298 participants in the analysis. Participants received \$1.50 plus a bonus of \$2.18 $\pm$ \$0.74 for 23.6 $\pm$ 10.4 minutes of work.

Main Task The main task of Experiment 2 had the same basic structure as that in Experiment 1, but with a different graph and reward structure (see Figure 5.4A). The graph had a single choice point at the first move (four options) followed by four forced moves. The reward distributions depended on a between-participant condition. In the constant variance condition, it was the same as in Experiment 1. In the other two conditions, most nodes were -1 or $+1$ with equal probability, but four nodes had an extreme distribution. For increasing variance, the terminal nodes (farthest from the initial location) had values of $+20$ with $2/3$ probability and -40 with $1/3$ probability. For decreasing variance, the nodes closest to the initial location had value $+1$ with $3/5$ probability, and either $+20$ or -20 with roughly $1/5$ each, slightly skewed towards -20 to make the expected reward 0 (.185 and .215). These distributions were selected to make the optimal planning strategy closely resemble depth-first and breadth-first search in the increasing and decreasing variance conditions respectively.

Procedure The procedure was identical to Experiment 1 except that we replaced the bonus question with a question asking on which nodes the maximal reward could be found.

5.4.3 Experiment 3

All aspects of the design are identical to Experiment 2 except where noted otherwise. We recruited 319 participants. We excluded 11 who failed the quiz and 17 who did not complete the experiment, leaving 291 participants in the analysis. Participants received $\$1.50$ plus a bonus of $\$2.49 \pm \$0.43$ for 21.3 ± 7.5 minutes of work.

Main Task The task had the same basic structure and graph as Experiment 1. The key difference from previous experiments is that we lifted the restriction that only nodes adjacent to the initial state or already-revealed nodes could be revealed. That is, participants could reveal any unrevealed node at any point. The graph was the same as in Experiment 1. The reward structure varied by condition. In the constant variance condition, it was identical to Experiment 1. In the increasing variance condition, the reward distribution for depth 1 nodes was uniform over the values $\{-2, -1, +1 + 2\}$. The possible values at later depths were scaled by 3^d; that is, the range and standard deviation increased by a factor of 3 from each depth to the next, up to $\{-18, -9, +9 + 18\}$ at the depth 3 leaf nodes. In the decreasing variance condition, the situation was exactly reversed: depth 1 nodes could take values in $\{-18, -9, +9 + 18\}$, and the values decreased by a factor of 3 with each step down to $\{-2, -1, +1 + 2\}$ at the leaf nodes.

Procedure The procedure was identical to Experiment 2.

5.4.4 Experiment 4

We recruited 137 participants from Prolific who reported fluency in English, resided in the United States, and had a 95% approval rating. We excluded 7 who failed the quiz and 37 who did not complete the experiment, leaving 93 in the analysis. Due to a technical error, instruction progress was not recorded, hence the larger number of incompletes. Participants received $\$1.75$ plus a bonus of $\$0.99 \pm \$0.13$ for 18.2 ± 7.8 minutes of work.

Main Task Participants assumed the role of a travel agent planning a road trip. On each trial, the participant saw a map of an island with eleven cities represented as circles and roads represented as arrows. Participants were instructed that the client wants to travel from a given starting location to a goal location. Each "day" they can move along any single arrow between two cities and each "night" the client has to stay in a hotel at a price that varies between cities. Participants were informed that hotels could cost $\$25$, $\$35$, $\$50$, or

$100, and that all values were equally likely. To reveal the price of the hotel in a given city, participants had to type its name into a text box. They could uncover any number of prices, in any order, and they could submit their recommended route at any moment. At this point the total cost was computed; this value was subtracted from a budget of $300 and the participant's bonus for the trial was 1 cent for each $10 remaining.

Procedure The experiment began with an instruction phase in which the task was explained through verbal instructions and images. Participants were required to complete a quiz (in no more than three attempts) before continuing. Each participant then performed 8 trials, the first of which was a practice trial that did not count towards their bonus payment.

5.4.5 Model specifications

Each of our candidate models corresponds to a parameterized family of metalevel policies. For all models, the policy is specified as a four-step generative process. First, if the frontier is empty (i.e., all nodes have been clicked or pruned), the model executes the termination operation, $\perp$. Second, if the frontier includes at least one node, then a random legal computation is executed with some probability, ε. Otherwise (step 3), the model executes $\perp$ with probability $p^M_{\text{stop}}(m)$; the form of this function depends on the model, M. Finally (step 4), if the model did not act randomly or terminate, then it selects a node to expand, each node having probability $p^M_{\text{select}}(m, c)$ The models are thus defined by stochastic stopping and selection rules.

The heuristic models (best-first, depth-first, and breadth-first) all use a common stopping rule that incorporates both the relative and absolute value of the best path identified so far. The stopping probability is a logistic function of a weighted linear combination of these terms, that is,

$$p^H_{\text{stop}}(m) = \frac{1}{1 + \exp\left\{-f_{\text{stop}}(m)\right\}}, \tag{5.5}$$

where

$$f_{\text{stop}}(m) = \beta_{\text{satisfice}} \cdot V_{\text{best}} + \beta_{\text{bestnext}} \cdot (V_{\text{best}} - V_{\text{next}}) + \theta_{\text{stop}}. \tag{5.6}$$

V_{best} and V_{next} are the expected values of the best and second-best paths given the current mental state, θ_{stop} sets the midpoint of the logistic function, and the β's control the contribution of each term to its slope. This implementation allows the model to both flexibly interpolate between a relative and an absolute stopping rule and also to vary the precision in the application of the rule. For example, a classic "hard" satisficing rule can be created by

setting $\beta_{\text{satisfice}}$ to a very large number, β_{bestnext} to zero, and θ_{stop} to $-\theta \cdot \beta_{\text{satisfice}}$ where θ is the aspiration level. This results in

$$p_{\text{stop}}^{\text{SATISFICE}}(m) = \frac{1}{1 + \exp\left\{-\beta_{\text{satisfice}}(V_{\text{best}} - \theta)\right\}}, \tag{5.7}$$

that is, a logistic function of the expected value of the best path with slope $\beta_{\text{satisfice}}$ and intercept θ.

We defined the selection rule for each heuristic model so that its policy approximates the corresponding classical search algorithm. To do this, we defined

$$p_{\text{select}}^{H}(m, c) = \frac{\mathbf{1}(c \in \text{frontier}(m)) \cdot \exp\left\{\beta_{\text{select}} \cdot f_{\text{select}}^{\text{ALG}}(m, c)\right\}}{\sum_{c' \in \text{frontier}(m)} \exp\left\{\beta_{\text{select}} \cdot f_{\text{select}}^{\text{ALG}}(m, c')\right\}}, \tag{5.8}$$

where $f_{\text{select}}^{\text{ALG}}(m, c)$ denotes a node-scoring function for each algorithm; specifically,

$$\begin{aligned} f_{\text{select}}^{\text{BEST}}(m, c^{(i)}) &= V(m, i) = \max_{p \in \{\mathcal{P} | i \in p\}} V(m, p) \\ f_{\text{select}}^{\text{DEPTH}}(m, c^{(i)}) &= \text{depth}(i) \\ f_{\text{select}}^{\text{BREADTH}}(m, c^{(i)}) &= -\text{depth}(i). \end{aligned} \tag{5.9}$$

We chose these node scoring functions to ensure that in the limit $\beta_{\text{select}} \to \infty$, the model's selection rule is deterministic and exactly matches the corresponding algorithm. Pure best-first search always expands a node with maximal expected value, pure depth-first search always expands the deepest node in the tree, and pure breadth-first search always expands every node at each depth before expanding any at the next depth. However, to account for variability in human selection decisions, we allow for $\beta_{\text{select}} \in [0, \infty)$.

The random model takes the same form as the heuristic models, with $f_{\text{select}}^{\text{RAND}}(m, c) = 0$ and $f_{\text{stop}}(m) = \theta_{\text{stop}}$. This is equivalent to a fixed stopping probability and random selection. In the random model the probability of choosing computations at random is set to zero ($\varepsilon = 0$) because this step is redundant.

For the optimal model, we define both the stopping and the selection rules using the optimal state-action value function, Q_γ, of the metalevel MDP with computational cost γ. We

computed the Q function using dynamic programming. The stopping rule is

$$p^O_{\text{stop}}(m) = \frac{\exp\left\{\beta_{\text{stop}} \cdot Q_\gamma(m, \bot)\right\}}{\sum_{a' \in \text{frontier}(m) \cup \{\bot\}} \exp\left\{\beta_{\text{stop}} \cdot Q_\gamma(m, c')\right\}} \tag{5.10}$$

and the selection rule is

$$p^O_{\text{select}}(m, c) = \frac{\exp\left\{\beta_{\text{select}} \cdot Q_\gamma(m, c)\right\}}{\sum_{c' \in \text{frontier}(m)} \exp\left\{\beta_{\text{select}} \cdot Q_\gamma(m, c')\right\}}. \tag{5.11}$$

Note that if $\beta_{\text{select}} = \beta_{\text{stop}}$, this corresponds to a single softmax over the full action space. However, we use separate inverse temperature parameters for stopping and selection to match the flexibility of the error model used by the optimal model to that of the heuristic models.

The myopic model has the same form, but the Q_γ function is replaced by a myopic one-step approximation (Russell & Wefald, 1991a), which we denote Q^{myopic}_γ. For the termination operation, this approximation is exact because the trial ends after this action is executed and thus $Q_\gamma(m, \bot) = Q^{\text{myopic}}_\gamma(m, \bot) = r(m, \bot)$. For expansion, the myopic model approximates the Q value as the expected value of stopping at the next time step (after expanding a node) minus the expansion cost, that is

$$Q^{\text{myopic}}_\gamma(m, c) = \mathop{\mathrm{E}}_{s' \sim T(\cdot|m,c)}[r(s', \bot)] - \gamma. \tag{5.12}$$

5.4.6 Pruning and depth limits

To model pruning (Huys et al., 2012, 2015) and depth limits (Keramati et al., 2016; Snider et al., 2015), we assume that each time a participant expands a node, she may choose to eliminate the corresponding branch from further consideration. Because both mechanisms ultimately involve removing a branch of the decision tree, we refer to them as value-based and depth-based pruning, respectively. If a path is pruned, all unexpanded nodes on that path are removed from the frontier, preventing the model from selecting these nodes. Note that pruning also acts as a secondary stopping rule because all models stop whenever the frontier is empty.

We assume that the value-based and depth-based pruning mechanisms operate independently. For each one, the probability of pruning a just-expanded node is defined as a logistic

function of the expected value or tree depth of the node. Value-based pruning is defined

$$p_{\text{prune}}^{\text{VALUE}}(m,i) = \frac{1}{1+\exp\left\{-\beta_{\text{prune}}^{\text{VALUE}} \cdot (\theta_{\text{prune}}^{\text{VALUE}} - V(m,i))\right\}} \tag{5.13}$$

where $V(m,i)$ is the value of the best path that includes node i, defined in Equation 5.9. Thus, a path is increasingly likely to be pruned the further below $\theta_{\text{prune}}^{\text{VALUE}}$ its expected value is. Depth-based pruning is defined

$$p_{\text{prune}}^{\text{DEPTH}}(m,i) = \frac{1}{1+\exp\left\{-\beta_{\text{prune}}^{\text{DEPTH}} \cdot (\text{depth}(m,i) - \theta_{\text{prune}}^{\text{DEPTH}})\right\}}. \tag{5.14}$$

Thus, a path is increasingly likely to be pruned the further past the depth limit it is. Finally, the complete heuristic model contains both forms of pruning operating independently, resulting in

$$p_{\text{prune}}(m,i) = 1 - \left(1 - p_{\text{prune}}^{\text{VALUE}}(m,i)\right) \cdot \left(1 - p_{\text{prune}}^{\text{DEPTH}}(m,i)\right). \tag{5.15}$$

Unfortunately, implementing this model exactly requires creating (and marginalizing over) a new latent state variable that specifies which nodes have been pruned. To avoid the formidable computational challenges associated with fitting such a model, we follow Huys et al. (Huys et al., 2012, 2015) and use a mean-field approximation. Specifically, we assume that the stochastic decision of whether to prune each branch is resampled at every time step based on its current expected value, treating the set of pruned nodes at each time step as independent. When computing the stopping and selection probabilities (Equations 5.5 and 5.8), we marginalize over all possible frontiers that could result from different combinations of pruning decisions, weighing each by its probability according to Equation 5.15.

5.4.7 Backward planning and forward-search bias

In Experiments 3 and 4, we modified the metalevel MDP to allow planning algorithms that do not correspond to traditional decision-tree search. The formalism described above is maintained with one exception: $\text{frontier}(m)$ in Equations 5.8, 5.10, and 5.11 is replaced with $\text{unexpanded}(m) = \{c^{(i)} \mid m^{(i)} = \oslash\}$. Although the metalevel state and action spaces are formally the same, we now interpret a metalevel state as a partially computed value function and a metalevel action as computing the reward at a future world state and also integrating this information into the value of its ancestor states (we assume an acyclic transition function).

However, because we found that participants still showed a strong tendency for forward search, we augmented the selection rule of all models with a forward-search bias, $\beta_{\text{forward}} \cdot \mathbf{1}(c \in \text{frontier}(m))$. For the heuristic models, this term was added to f_{select}. For the optimal and myopic models, it was added inside the exponentiation in the numerator and denominator of Equation 5.11.

5.4.8 Model fitting and evaluation

We fit all models by maximum likelihood estimation at the individual level, cross-validated across trials. We used five folds in all experiments except Experiment 4, where we used seven folds because there were only seven trials (excluding the practice trial). For each participant, model, and fold, we optimized the model's free parameters by minimizing the negative log-likelihood on the training set, using the L-BFGS algorithm with 100 random starting points sampled from a plausible range. The lapse rate ε was constrained to be no less than .01 to prevent extremely low test likelihoods (a simple form of regularization). For the optimal model, we optimized the cost parameter on a grid (0 to 4 in steps of .05) because dynamic programming is not easily differentiated. We then computed the log-likelihood of each computational action in the test set (node expansions and terminations). The total log-likelihood of the data under each model is the sum of the log-likelihoods in each test set.

5.4.9 Statistical analyses

Analyses on human data were performed on all test trials for all participants who passed the exclusion criterion. For comparison to the optimal model, we conducted analyses on a simulated dataset using costs fit to participant data, but removing decision noise (setting $\varepsilon = 0$, $\beta_{\text{stop}} = 10^5$, and $\beta_{\text{select}} = 10^5$).

Regression analyses were performed using the "lme4" R package with default settings (Bates et al., 2015). We included random intercepts as well as random slopes for each fixed effect. Confidence intervals were produced using the default Wald method. Note that, to allow for direct comparison of the model and participant coefficients, we also use mixed-effects regression for the model; in this case, we used the participant that the model's cost parameter was fit to as the group identifier.

All other analyses were performed over participant means. Thus, we report mean proportions rather than total proportions. Confidence intervals were produced by bootstrapping over participants. Wilcoxon and Spearman tests were performed using the "scipy" Python

package with default settings.

6

Conclusion

We have become acquainted with a general framework for modeling cognition as a sequential decision problem: metalevel Markov decision processes. We saw how the framework can be applied to derive rational mechanistic models in three different domains: attention, memory, and planning. And in each case, we found that human behavior showed substantial qualitative alignment with the optimal metalevel policy. Taken together, the results suggest that human cognitive processes are well-adapted to the internal environments in which they operate. More importantly, by formally characterizing the problems posed by those mental environments, and their optimal solutions, we have developed a richer understanding of human cognition in each of these domains.

The breadth of domains covered in Chapters 3-5 is substantial, but the models share many core features. Although Chapters 3 and 4 considered very different tasks (attention in choice and memory recall), they relied on essentially the same state space and transition structure, based on Gaussian evidence accumulation. This parallels the breadth of domains in which evidence accumulation models have been applied. Our model of planning (Chapter 5) employed a very different type of state space (decision trees), but it shares with Chapter 3 the idea that decision-making can be understood as gathering information about rewards (c.f. Tajima et al., 2016; Sezener et al., 2019). All three chapters rely on probabilistic models that specify how the effects of computations relate to an unknown world state. Although these similarities may undercut the claimed generality of the approach, they can

also be viewed as a strength. By specifying models using a common framework, it is easier to transfer concepts and computational tools between domains.

6.1 Kindred efforts

The idea of modeling cognitive processes as optimal solutions to sequential decision problems is not unique to this dissertation. There are at least three major strands of research in this area. I briefly review these strands below, noting similarities and differences to our approach.

6.1.1 Optimal evidence accumulation

As mentioned in the introduction, there is a long history of modeling optimal speed-accuracy tradeoffs using evidence accumulation models. Early work (e.g., Bogacz et al., 2006; Vul et al., 2009) focused on binary choices with accuracy-based rewards (e.g., +1 for correct responses) and known evidence coherence (i.e., the strength of evidence supporting the correct response is always the same). In these cases, the optimal stopping rule is given *sequential probability ratio test* (SPRT): continue collecting evidence until a threshold level of evidence is reached either for or against the hypothesis. However, when any of these assumptions are violated, the SPRT (and by extension) is no longer optimal.

By explicitly modeling evidence accumulation as a sequential decision problem, Drugowitsch et al. (2012) were able to derive the optimal stopping rule when evidence coherence varies from trial to trial.[1] This model adopts the same basic evidence accumulation and belief updating dynamics as we used in Chapter 3 and 4, with a single accumulator tracking relative evidence for one choice vs. the other. They showed that the optimal policy corresponds to a threshold that changes over time (rapidly expanding and then slowly collapsing). Building on this work, Tajima et al. (2016) characterized the optimal decision thresholds for value-based choices, where both importance and difficulty depend on the difference in value between the choice options (c.f. Fudenberg et al., 2018). Going further, Tajima et al. (2019) characterizes the optimal solution for three-alternative choices, where the mental states and thresholds reside in an three-dimensional space (including time).

Despite this richness, all the models mentioned above assume a very simple cognitive architecture, in which there are only two possible cognitive operations: gather more evidence,

[1] This problem is more complex because one cannot simply sum up the log likelihoods for each piece of evidence (which could support option A or B). One must maintain a full belief state over the drift rate, which specifies both the correct answer and the coherence (higher absolute values correspond to higher coherence).

or stop. Building on the Tajima et al. model, Jang et al. (2021) derived the optimal policy for the case when the agent can only sample from one item at a time (more precisely, when the value of the attended item produces less noisy samples). Without switching costs, Fudenberg et al. (2018) showed that it is optimal to perfectly balance attention to each item (alternating at each time step), but in the presence of switching costs, the optimal policy takes on a more interesting structure. As we showed in Chapter 3, the optimal policy takes on even richer structure when there are more than two alternatives, preferentially attending to items with high estimated value. However, the high dimensionality of the state space makes it impossible to exactly identify the optimal policy for trinary choice using backwards induction (the method employed in the Drugowitsch, Tajima, and Jang papers; described in Section 2.7.1). We were only able to (approximately) identify the optimal policy in this case using the BMPS algorithm (Section 2.8), which we developed specifically to solve metalevel MDPs. This highlights the value of using a formalism that is tailored to the specific type of sequential decision problems posed by cognition.

Another line of work in economics has also sought to characterize optimal evidence accumulation in cases where the agent can control the nature of evidence sampled (Woodford, 2014; Hébert & Woodford, 2017). The key distinguishing feature of these models is the assumption of a very flexible cognitive architecture in which the agent can gather an arbitrary form evidence at each time step (formally a conditional distribution of a signal given the true world state). Paradoxically, allowing for such flexibility actually makes it easier[2] to identify the optimal policy, as it can be derived analytically. However, this flexibility may limit the ability of these models to account for human cognitive processes, which are likely adapted to more restrictive architectures.

6.1.2 POMDP models of visual search

Another area where optimal sequential models are often found is in visual search. In these tasks, participants are asked to find a target object hidden among distractors in an image. Early work in this area suggested that people fixate on areas where the target is most likely to be (Najemnik & Geisler, 2005). This strategy can be understood as a myopic policy (in precisely the sense of Section 2.7.2) for a metalevel MDP, as it maximizes the chance of identifying the target on the very next time step. However, as we have seen, myopic policies are often suboptimal. Recognizing this, Butko & Movellan (2008) reformulated the "ideal observer" model of Najemnik and Geisler as an "information-gathering POMDP", a special case of POMDPs in which the state does not change and actions serve only to generate in-

[2] Assuming you have the math training of an Econ PhD.

formation. This is exactly the same restriction made by metalevel MDPs (see Section 2.5.4). However, rather than using a reward function capturing the cost of fixations and the reward for finding the target (as we would do in the metalevel MDP approach), they assumed a reward function that directly rewards reduction of entropy in the belief state. Nevertheless, they found that the learned policy found the target faster than the ideal observer model.

Taking an approach more similar to ours, Acharya et al. (2017) present a POMDP model in which the reward exactly captures the incentives of the task and the opportunity cost of time. Indeed, this model can be viewed as a metalevel MDP in which fixations correspond to computations (as in Chapter 3). Applying the model to the distractor-ratio task, they found that the optimal fixation policy better captured qualitative patterns in human fixation data compared to the ideal observer model. In parallel, Hoppe & Rothkopf (2019) conducted an experiment specifically designed to distinguish myopic and planned eye movements (this involved using oddly shaped search regions and allowing participants to make exactly one or two fixations). They model the task as an MDP and find that the optimal policy predicts people's average fixation locations almost shockingly well, predicting exactly when and how people deviate from the myopic ideal observer model.

6.1.3 POMDP models of decision-making

Most similar to our work, researchers have recently begun to use POMDPs to model decision-making processes. Building on the models of visual search described above, Chen et al. (2017) present a model of how people direct their eye fixations when making a decision based on information presented in a table. They find that both people and the optimal policy use decision strategies that collect intermediate amounts of information compared to classical strategies (such as take-the-best Gigerenzer & Goldstein, 1996). This parallels our own predictions and findings in a metalevel MDP model of multi-attribute choice (Gul et al., 2018).

In all the POMDP models mentioned so far, the "computations" correspond to gathering information from an external environment. To some extent, this can also be said of the models in this dissertation (although, in Chapters 3 and 4, there is good reason to believe that people's fixations are more indicative of internal processing than true information gathering). In very exciting new work, Chen et al. (2021) present a POMDP model of risky choice in which the actions correspond to truly internal operations such as making ordinal comparisons of payoff probabilities and computing a noisy estimate of expected value. They find that, when the expected value computation is very noisy, the optimal policy relies more on the more robust comparison operations. This improves performance, but yields

systematic choice biases (decoy effects). Although all of the work reviewed so far takes a fully Bayesian view of computation (making no distinction between mental states and belief states), Oulasvirta et al. (2022) propose a general POMDP framework in which an agent interacts with an "internal environment" which is defined by "mental states", suggesting that the framework could be applied more generally (although I am not aware of any concrete examples of this more general case).

In light of all the work reviewed in this section, we might ask again: why bother with metalevel MDPs? The answer, I think, comes down to what level of generality we would like to have in a framework for developing optimal sequential models of cognition. Models of evidence accumulation are typically posed in very specific terms, directly specifying the dynamic programming problem for a given decision-making task. On the other end, the POMDP models are posed in very general terms, as just one instance of the problem of interacting with a dynamic and partially observable environment. Metalevel MDPs taken an intermediate approach. They are general enough to model many different types of cognitive processes within one framework, but they are specific enough to *only* model cognitive processes (or more generally, computational processes). In particular, metalevel MDPs formally distinguish between internal and external states and actions. Although this can be limiting (as discussed below), it has important conceptual and technical benefits. Emphasizing the latter, the POMDP models of decision-making discussed above use artificially simplified state spaces, for example, treating continuous features as binary (Chen et al., 2017), or assuming a small number of operations per episode (Chen et al., 2021)). This is likely because the general-purpose reinforcement learning approaches they employ would not have performed well in larger state spaces. Using metalevel MDPs, we were able to identify optimal policies in richer cognitive architectures than would be feasible when treating the problem as a generic POMDP.

6.2 Points of weakness and avenues for growth

There are several dimensions on which the framework could be extended in future work.

6.2.1 Learning the metalevel homunculus

One of the most significant challenges for the metalevel MDP framework is the problem of infinite regress. As mentioned in the Introduction (Section 1.4.1), the framework assumes that people are *metalevel rational*, meaning that they choose computational actions to optimally balance a cost-benefit tradeoff. However, it does not explain how those choices are

themselves made. In this way, the framework assumes a "metalevel homunculus": an unbounded, perfectly rational agent that always knows just what thought to think next (c.f. Hazy et al., 2006; Botvinick & Cohen, 2014). Explaining (or perhaps explaining away) the homunculus is critical for these models to provide a complete explanation of how people effectively allocate computational resources.

Perhaps the most plausible theory of the metalevel homunculus is that it is learned. Indeed, learning has been a critical aspect of many models of metalevel control in psychology. The earliest examples are found in cognitive architectures like ACT-R and SOAR, which have simple mechanisms for learning when to apply different production rules (Laird et al., 1986). However, like early models of strategy selection (Shrager & Siegler, 1998), this type of learning was primarily associative. The idea that a metalevel controller explicitly learns to maximize reward was proposed in later strategy selection models (Erev & Barron, 2005; Rieskamp & Otto, 2006; Lieder & Griffiths, 2017). Reinforcement learning (RL) has also been proposed as the mechanism by which people learn which goals to pursue (Cushman & Morris, 2015), and when to retrieve and store information in working memory (O'Reilly & Frank, 2006; Todd et al., 2008) and episodic memory (Lu et al., 2022).

Metalevel MDPs provide a natural framework in which to cast theories of metacognitive RL. However, explicitly formalizing the metalevel control problem in this way also reveals a major challenge: learning good policies in metalevel MDPs is "exceptionally difficult" (Hay, 2016). For example, my colleagues and I have shown that a widely used deep reinforcement learning method (a DQN; Mnih et al., 2015) is unable to learn a good policy in a simplified form of the attention metalevel MDP from Chapter 3 (Callaway et al., 2018). In a more complex problem (the game of Hex), Hay (2016) showed that a learned metalevel policy can outperform a standard Monte Carlo tree search algorithm (Kocsis & Szepesvári, 2006), but only when the number of iterations is very limited (20 or less). Why is it so challenging to learn policies for metalevel MDPs? There are at least three reasons: metalevel MDPs are characterized by (1) large state spaces, (2) stochastic dynamics, and (3) sparse rewards. Reward sparsity is perhaps the most challenging, as it induces an extreme temporal credit assignment problem. In each episode, the agent takes many computations, but receives only a single external reward. It is thus not clear which computations should receive "credit" for large rewards.

One way to make learning easier in the presence of sparse rewards is to adjust the reward function by adding *shaping rewards* (or "pseudo rewards") that provide more immediate feedback about the value of each (mental) action that is executed (Ng et al., 1999). For example, shaping rewards derived from the optimal metalevel value functions can accelerate

human learning in the Mouselab MDP task used in Chapter 5 (Callaway et al., 2022a). Although people would not have access to such perfect shaping rewards in the real world, they may have access to simpler but still useful surrogates.[3] That is, people may experience a reward when they have a "good thought" even if it does not immediately lead to an external action (Gopnik, 1998). Indeed, people appear to place value on external information that cannot be acted on (Eliaz & Schotter, 2007; Gottlieb & Oudeyer, 2018), and researchers have begun to formalize the specific qualities of information that people value (Markant & Gureckis, 2014; Markant et al., 2016). This suggests a fascinating research question: what factors elicit the subjective experience of having a good thought?

6.2.2 Partially observable minds

A key structural assumption in the framework is the distinction between the mental state and the world state. The defining features of the mental state are that it can be affected by computation, and that it is directly accessible to the agent. However, these two features do not need to coincide. It is possible that we do not have complete access to aspects of our mental state that we can nevertheless control to some extent. This possibility was suggested by Suchow & Griffiths (2016), who proposed a POMDP model of working memory maintenance. In this model, the agent selects mental actions to increase the activation level of a selected memory, but the current activation levels can only be imperfectly measured.

To allow for partial observability in a metalevel MDP, we can simply assume that the agent does not have direct access to the mental state, but instead has access to an incomplete observation of that state. An interesting question arises as to how these observations should be used. From a POMDP perspective, the observations should be integrated over time into a metalevel belief state (a distribution over mental states). However, this complicates the metalevel controller considerably, creating technical challenges in finding optimal solutions, and exacerbating the homunculus problem described above. An alternative approach, employed by Suchow & Griffiths (2016), is to assume that the metalevel decisions are made based only on the current observation. This yields a simpler and perhaps more psychologically plausible model. However, because observations are not Markovian, one can no longer use standard dynamic programming techniques to find the optimal policy. This further motivates the development of model-free strategies for learning metalevel policies.

[3] Hay (2016) propose one intuitively appealing shaping reward based on the difference in the expected value of taking an external action in the mental states before and after executing a computation. However, the results were, in his terms, "not yet as good as we'd like".

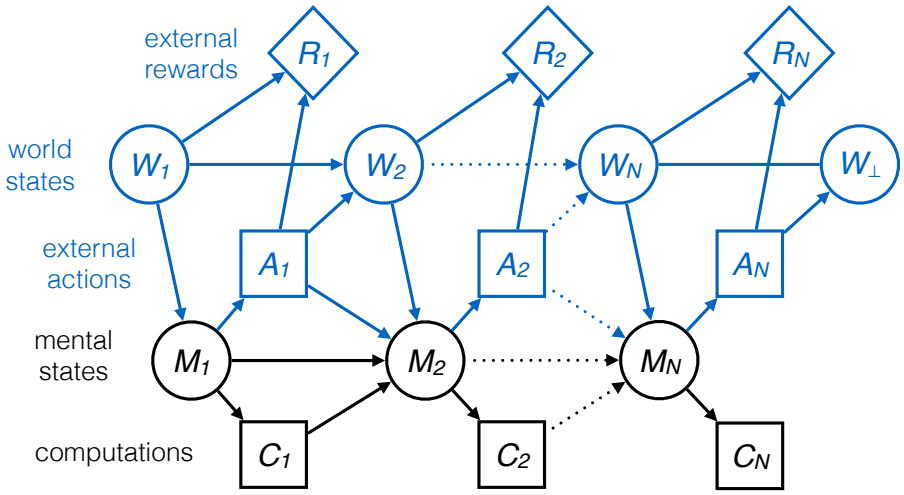

Figure 6.1: Interleaved metalevel MDP. To formalize the problem of interleaved computation and action, we can augment the metalevel MDP framework to include sequences of world states, actions, and external rewards. Note that there is no cognitive cost in this formalism, as any opportunity costs should be captured in the external reward function. The elements that capture the external environment are indicated in blue.

6.2.3 Interleaved computation and action

A second key structural assumption in the framework is that each metalevel episode occurs within a single external timestep. That is, the world state does not change, and only one external action is executed (at the end of the episode). This has two implications. First, it assumes that the agent can think as long as they want without having to worry about the world state changing. Second, it assumes that information considered while choosing one action cannot influence the selection of future actions. Although these assumptions hold approximately in many experimental and naturalistic tasks, they will not hold in fast-paced dynamic problems (e.g. jay walking) or in sequential problems with stochastic dynamics (e.g. chess).

To model interleaved computation and action, we essentially replace the single world state and action with a full external MDP that progresses in lockstep with the metalevel MDP (c.f. joint-state MDPs; Russell & Wefald, 1991b; Parr & Russell, 1998; Hay, 2016). Figure 6.1 illustrates one way this could work. At each timestep the agent selects both a computation and an action (the action may be to simply sit still while thinking). The next world state and reward depend on the current state and action, as in a standard MDP. As in standard metalevel MDPs, the next mental state depends on the current mental state and the computation executed; but it additionally depends on the previous action and the *new* world state. The former captures the fact that the agent may understand how their actions affect

the world. The latter captures the agent's ability to perceive the changing world state.

In the interleaved case, the decision about how much to think becomes more nuanced. Specifically, one must not only decide *how much* to think but also *when* to think. In some cases, it may make sense to do all of one's thinking up front, as the standard metalevel MDP assumes. This strategy makes sense because it ensures that every action you take is informed by all the computation you do. However, there are at least three reasons one might want to begin acting before having a complete plan. First, one may be able to continue thinking while executing the early part of the plan, for example considering which way one will turn while walking down a long corridor (O'Ceallaigh & Ruml, 2015). Second, if the world dynamics are stochastic, knowing how those transitions unfold will allow the agent to focus their planning on the the situations they actually encounter. Third, forming a complete plan may impose representational costs that could be avoided by only constructing a concrete plan for the immediate future, perhaps having a more abstract plan for the more distant future (Ho et al., 2020).

6.2.4 Optimizing the architecture

A third assumption of the framework is that only the policy, and not the metalevel MDP itself, is optimized. This assumption is almost always made in standard applications of MDPs, as the MDP represents the external environment, which the agent has no direct control over. In contrast, because a metalevel MDP represents an agent's internal computational environment, it is likely that the metalevel MDP itself is adapted to the structure of the external problems the agent has to solve.

There are two timescales at which adaptation of an organism's cognitive architecture could occur (in animals): evolutionary and developmental. Clearly, evolution has a major role in shaping the total amount of biological resources allocated to cognition (e.g., brain volume), as well as the ease with which certain types of computations can be learned and executed. However, at the level of abstraction that we have posed metalevel MDPs, the developmental timescale is likely to be more relevant. Indeed, it is natural to view many types of learning as a process of developing new possible mental states and computational actions. For example, as a consequence of reading this dissertation, you have (hopefully) acquired a computation along the lines of "identify the computational actions in this cognitive model."

In some cases, these acquired computational actions may be composed of simpler operations, as in hierarchical reinforcement learning (Sutton et al., 1999; Dietterich, 2000; Bacon et al., 2016). However, a key assumption of hierarchical RL is that all abstract actions are ultimately grounded out in concrete actions. Although it is possible to design systems

that learn to do complex reasoning using truly primitive operations (Piantadosi, 2021), it is also plausible that the basic operations over which metalevel control operates are better explained as emerging from a sub-symbolic process. Taking this latter approach, Chang et al. (2019) present a metalevel MDP model that simultaneously learns a set of computations (represented as small neural networks) and a policy for applying them, showing that the model can outperform standard deep learning approaches in problems requiring combinatorial generalization.

6.3 Parting words: On frameworks

Psychologists often make the distinction between *topic people* and *methods people*. The topic people adopt some specific cognitive domain (e.g., memory), and they apply different techniques to develop a deeper understanding of it. In contrast, the methods people have some preferred technical tool (e.g., Bayesian inference), and they look for areas where it could be productively applied. I have never liked this distinction. This is, in part, because I would clearly fall into the latter camp, which is generally considered to be the less noble of the two. But beyond that, I think it is leaving out an important class: *framework people*. In contrast to the tool-obsessed methods people, framework people are obsessed with conceptual minimalism. Certainly, frameworks often have associated methods; but these are just means to an end. The real goal is a unifying explanation of the mind in terms of a simple set of conceptual primitives.

I have been drawn to frameworks since I first discovered science. First, it was natural selection; enraptured by Dawkin's *The Selfish Gene*, I tried to explain everything as the result of evolution (something that will quickly get you into trouble in Psychology). In college, it was "Vector Symbolic Architectures" (Kanerva, 1988; Plate et al., 1995), which I was convinced would provide a unifying link between symbolic and neural representations.[4] And then in graduate school, I found metalevel MDPs.

Going the distance with one framework, however, I began to doubt what I once viewed as a core part of my scientific identity. At first, it was just *other* frameworks, whose applications often seemed dogmatic, artificial, or forced. Then, I began to see it in my own work. I found myself carefully designing experiments to make sure I could compute the optimal policy, abandoning interesting projects when I couldn't. The problem with frameworks is that they can constrain your thinking (restricting the sets of computations and mental states, if you will). Adapting on an over-used metaphor, if all you have is a hammer, you will only look

[4] https://fredcallaway.com/pdfs/callaway-undergrad-thesis.pdf

for nails.

But the existence of screws does not invalidate the hammer. As I argued in the introduction, constraints can be a guide rather than an impediment, providing structure in the vast space of possible cognitive theories. When the structure of the constraint is well-aligned with the aspect of cognition one wishes to understand, the framework can be enormously useful. The fundamental constraint of the proposed framework is that cognitive processes be characterized as optimal solutions to sequential decision problems. I believe (and hopefully have convinced the reader) that this constraint is quite well-aligned with many aspects of cognition—but certainly not all aspects. Taking a page from Gigerenzer's book, metalevel MDPs should only be one tool in a cognitive modeler's "adaptive toolbox."

There is an important difference, however, between frameworks and hammers. Like the cognitive processes that they model, frameworks should be flexible, adapting to the problems that they are applied to. Indeed, much of the value I have found in the metalevel MDP framework is not in its *application*, but rather in its *refinement*. In the course of writing this dissertation, I had to make many changes to the formalism to satisfactorily capture all three models in Chapters 3-5. And as discussed above, there are many ways the framework can continue to grow. On the other hand, if one freely contorts a framework to accommodate the idiosyncrasies of each application, then the framework ceases to provide meaningful constraints, and—by the same token— ceases to have any value.

In the end, I think frameworks can be very useful, as long as we maintain a healthy degree of skepticism and humility about them. The human mind is extraordinarily complex, and I have come to doubt that this complexity can be captured under any single framework—at least not one developed by the human mind itself. Thus, we should not view frameworks as goals in and of themselves, as possible candidates for a grand unifying theory, but rather as means to an end, as instruments to be applied in the ongoing project of understanding the human mind. Perhaps then, frameworks really are just tools. And if that makes me a methods person, so be it.

A

Supplementary information for Chapter 3

A.1 Task descriptions

This datasets for binary and trinary choice were initially reported in Krajbich et al. (2010) and Krajbich & Rangel (2011), respectively. For the convenience of the reader, we include the task description from the original papers.

A.1.1 Binary choice

The experiment consisted of 39 Caltech students. Only subjects who self-reported regularly eating the snack foods (for example, potato chips and candy bars) used in the experiment and not being on a diet were allowed to participate. These steps were taken to ensure that the food items we used would be motivationally relevant. This would not have been the case if the subjects did not like junk food. Subjects were asked to refrain from eating for 3 h before the start of the experiment. After the experiment they were required to stay in the room with the experimenter for 30 min while eating the food item that they chose in a randomly selected trial (see below). Subjects were not allowed to eat anything else during this time.

In an initial rating phase subjects entered liking ratings for 70 different foods using an on-screen slider bar ("how much would you like to eat this at the end of the experiment?", scale -10 to 10). The initial location of the slider was randomized to reduce anchoring effects. This rating screen had a free response time. The food was kept in the room with the subjects dur-

ing the experimental session to assure them that all the items were available. Furthermore, subjects briefly saw all the items at this point so that they could effectively use the rating scale.

In the choice phase, subjects made their choices by pressing the left or right arrow keys on the keyboard. The choice screen had a free response time. Food items that received a negative rating in the rating phase of the experiment were excluded from the choice phase. The items shown in each trial were chosen pseudo-randomly according to the following rules: (i) no item was used in more than 6 trials; (ii) the difference in liking ratings between the two items was constrained to be 5 or less; (iii) if at some point in the experiment (i) and (ii) could no longer both be satisfied, then the difference in allowable liking ratings was expanded to 7, but these trials occurred for only 5 subjects and so were discarded from the analyses. The spatial location of the items was randomized. After subjects indicated their choice, a yellow box was drawn around the chosen item (with the other item still on-screen) and displayed for 1 s, followed by a fixation screen before the beginning of the next trial.

Subjects' fixation patterns were recorded at 50 Hz using a Tobii desktop-mounted eye-tracker. Before each choice trial, subjects were required to maintain a fixation at the center of the screen for 2 s before the items would appear, ensuring that subjects began every choice fixating on the same location.

A.1.2 Trinary choice

Thirty Caltech students participated in the experiment. The screening, pre-experimental instructions, eye-tracking and liking rating phase were identical to those used in the binary choice task described in the previous section.

In the choice phase, subjects made their choices using the keyboard. The choice screen had a free response time. The items shown in each trial were randomly chosen. In all trials the three items were displayed in a triangular formation with the left and right items at the same vertical position, and the center item at the opposite vertical position. In half of the trials the center item was on the top half of the screen, and in the other half it was on the bottom half of the screen. Subjects indicated their choice by pressing the left, down, or right arrow keys for the left, center, and right items, respectively. After subjects indicated their choice, a yellow box was drawn around the chosen item (with the other item still on the screen) and displayed for 1 s, followed by a fixation screen, before the beginning of the next trial.

A.2 UCB policy optimization

To identify a set of 80 near-optimal policies, we used a method based on upper-confidence bound bandit algorithms (Auer et al., 2002).

Each "bandit" corresponds to a weight vector for the BMPS policy (Equation 3.8). We sample 8000 such weight vectors to roughly uniformly tile the space of possible weights. Concretely, we divide a three-dimensional hypercube into $800 = 20^3$ equal-size boxes and sample a point uniformly from each box. The first two dimensions are bounded in $(0, 1)$ and are used to produce $w_{1:3}$ using the following trick: Let x_1 and x_2 be the lower and higher of the two sampled values. We then define $w_{1:3} = [x_1, x_2 - x_1, 1 - x_2]$. If x_1 and x_2 are uniformly sampled from $(0, 1)$, and indeed they are, then this produces $w_{1:3}$ uniformly sampled from the 3-simplex. The third dimension produces the future cost weight; we set $w_4 = x_3 \cdot \text{maxcost}$ where maxcost is the lowest cost such that the policy would always terminate in the initial mental state.

Next, we simulate 100 decision trials for each of the resulting policies, providing a baseline level of performance. Using these simulations, we compute an upper confidence bound of each policy's performance equal to $\hat{\mu}_i + 3\hat{\sigma}_i$, where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the empirical mean and standard deviations of the metalevel returns sampled for policy i. A standard UCB algorithm would then simulate from the policy maximizing this value. However, because we are interested in identifying a set of policies, we instead select the top 80 (i.e. 1% of) policies and simulate 10 additional trials for each, updating $\hat{\mu}_i$ and $\hat{\sigma}_i$ for each one. We iterate this step 5000 times.

Finally, we select the 80 policies with the highest expected performance as our characterization of optimal behavior in the metalevel MDP. To eliminate the possibility of fitting noise in the optimization procedure, we use one set of policies to compute the likelihood on the training data and re-optimize a new set of policies to generate plots and compute the likelihood of the test data. Note that we use the box sampling method described in the previous paragraph rather than a deterministic low discrepancy sampling strategy (Sobol, 1967) so that the set of policies considered are not exactly the same in the fitting and evaluation stages.

A.3 Individual fits

We have focused on group-level fits because we are especially interested in the ability of the model to predict differences between binary and trinary decisions. However, it is important to verify that the qualitative effects that we emphasize also hold in individual data, and

are not aggregation artifacts. It is also interesting to see to what extent the model can account for individual variability in fixation and choice behavior. To address both of these concerns, we present versions of each plot shown in the main text with separate panels for each participant. The model was fit to each participant's data following the same fitting procedure as for the group-level fit (using the same precomputed likelihood histograms). Finally, because many of the behavioral patterns are quite noisy with only 50 trials, we aditionally plot Bayesian linear model fits for both the human and model-simulated data (using logistic regression for binary dependent variables). These predictions were generated using the rstanarm package (Goodrich et al., 2020). The plots can be found at https://doi.org/10.1371/journal.pcbi.1008863.s002.

In brief, we found that most behavioral patterns shown in the main text figures were consistently demonstrated by a majority of participants. However, although most effects were consistently present and in the correct direction, the strength often varied considerably across individuals. In many cases, the model showed only a modest ability to capture this variability. This reflects the strong *a priori* assumptions of the model, in particular, the assumption that attention is allocated optimally.

A.4 Parameter recovery

To validate our model fitting approach, we conducted a parameter recovery exercise. We began by sampling 1024 "true" parameter configurations from the promising region of the parameter space that we considered when fitting human data (see main text *Methods*). We sampled these values using the 5-dimensional Sobol sequence (Sobol, 1967) to ensure good coverage of the space. For each parameter configuration, we computed two sets of 80 near-optimal policies (one for binary choice and one for trinary choice) using the UCB-based method described in the main text. Then, for each set, we simulated the even trials of the corresponding dataset. We simulated each trial only once (to match the amount of data when fitting participants), cycling between the 80 near-optimal policies. We then applied the full approximate maximum likelihood estimation procedure described in the main text for each dataset.[1] For each configuration, the maximum likelihood estimate of each parameter was its mean in the 30 configurations with highest likelihood (following our reporting approach for the fits to human data).

The results, shown in Figure A.1, suggest that we were able to recover parameters with fairly high accuracy. For all parameters besides the softmax temperature, the Pearson corre-

[1] We reused the likelihood histograms that we computed when fitting participant data. Critically, however, the policies used to generate these histograms were not the same ones used to generate the simulated data.

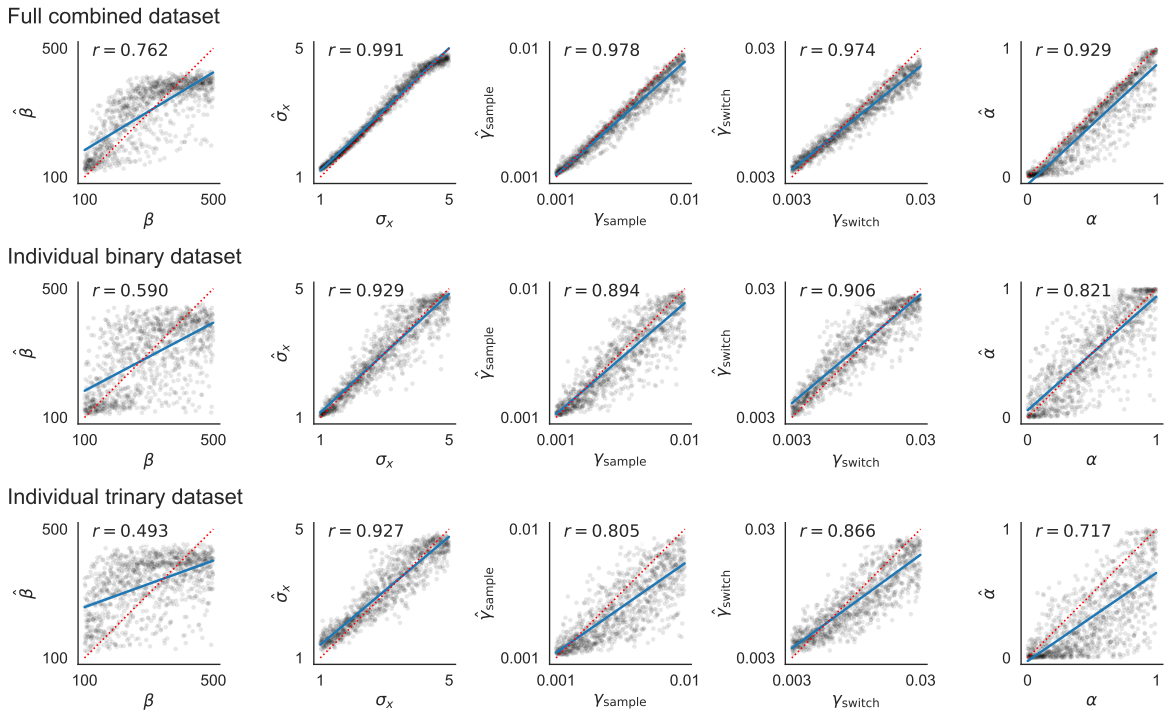

Figure A.1: Parameter recovery. Each panel plots the estimated parameter value as a function of the true parameter value. Each black dot corresponds to one simulated dataset. The dotted red line shows equality (i.e., perfect recovery) and the solid blue line shows the linear trend. The top row shows results when simulating the full joint dataset. The middle row shows results when simulating 50 trials (the amount of fitting data one individual produces) of binary choice. The bottom row shows the same for trinary choice.

lation was over 0.9. Importantly, we found only slight bias in the estimation procedure, with the best fitting linear regression line falling close to the equality line for all parameters. The largest bias was for the prior bias parameter, α, for which the recovered parameter was on average 0.095 less than the true parameter.

To validate our approach when fitting individual subjects, we repeated the steps above, except using only 50 simulated trials (the number of fitting trials for each subject). Unsurprisingly, we find that the estimates become less reliable; however the correlations are still fairly strong. In the trinary case, we see substantial bias for both γ_{sample} and α. Thus, care must be taken when interpreting the individual fitting results.

A.5 Implementation and validation of the aDDM

In order to compare our model to the predictions of the aDDM (Krajbich et al., 2010; Krajbich & Rangel, 2011), we reimplemented it based on code provided from the first author.

We made one change to the simulation procedure. In the original papers, the model predictions were generated by simulating an equal number of trials for all possible combinations of item ratings. In contrast, we have simulated each trial in the dataset a fixed number of times. That is, our simulations follow the empirical distribution of the item ratings. To verify the correctness of our implementation, we have replicated four key plots from the original binary and trinary papers, shown in Figures A.3 and A.4 respectively. Note that for these plots, we use the original approach of simulating each possible combination a fixed number of times.

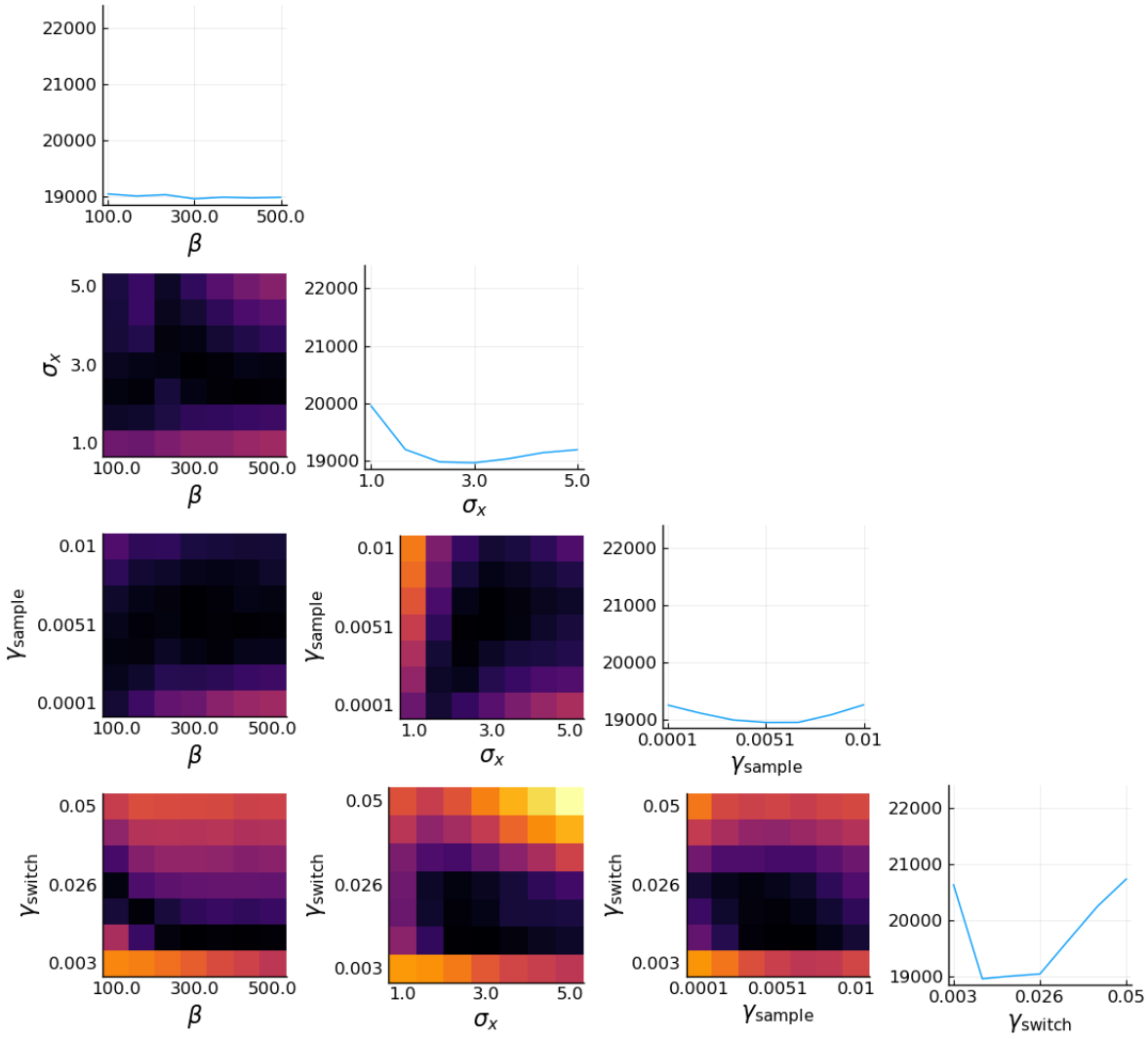

Figure A.2: Grid search on model parameters. Each panel shows the best likelihood achieved for each value of one parameter (diagonal) or combination of values for two parameters (off-diagonal), i.e., minimizing over all the non-plotted variables.

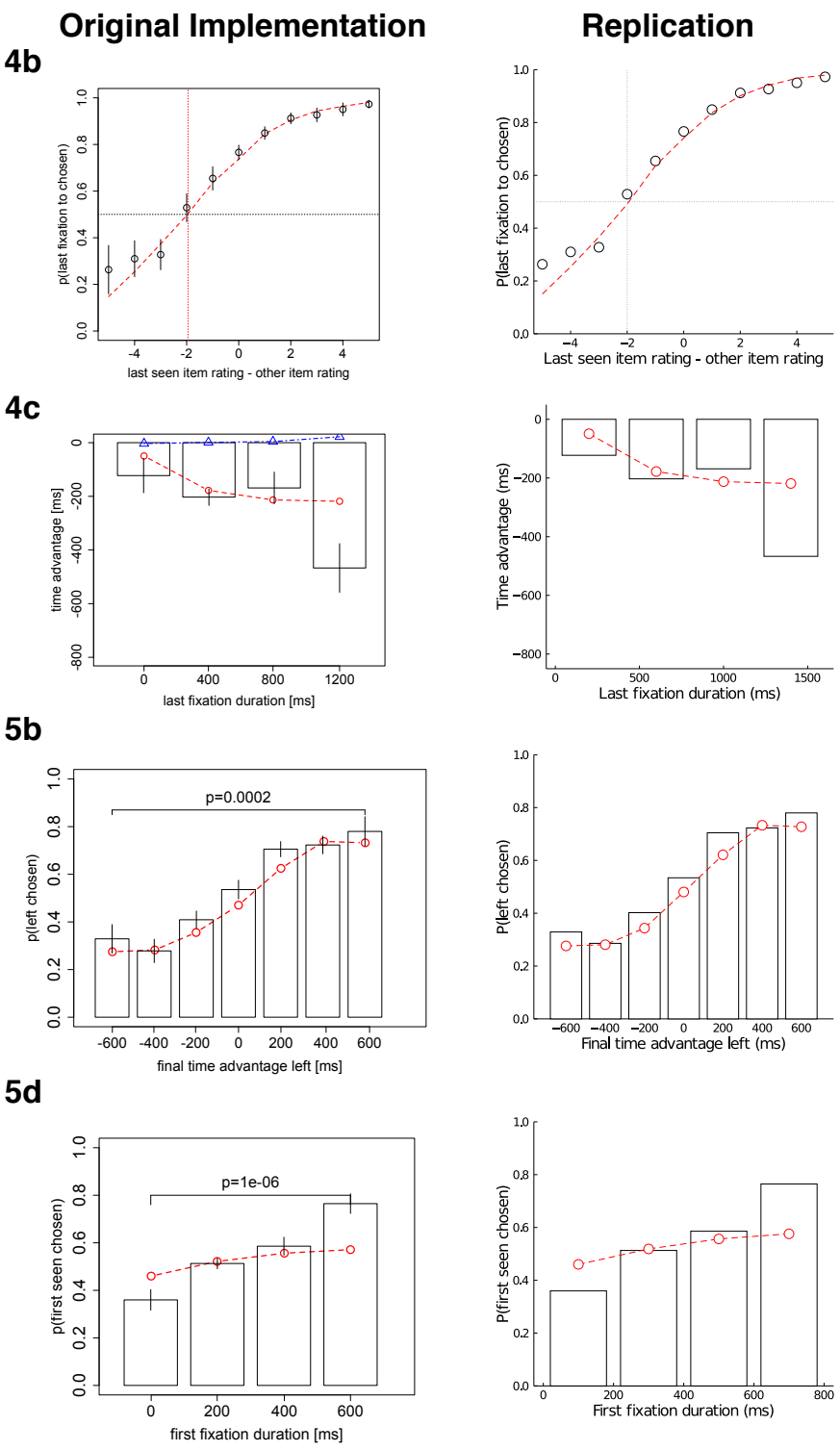

Figure A.3: Replication of Krajbich et al. (2010). Note that x axis labels in the orginal plots sometimes reflected the left tail of the bin; in these cases, we adjusted the tick locations accordingly.

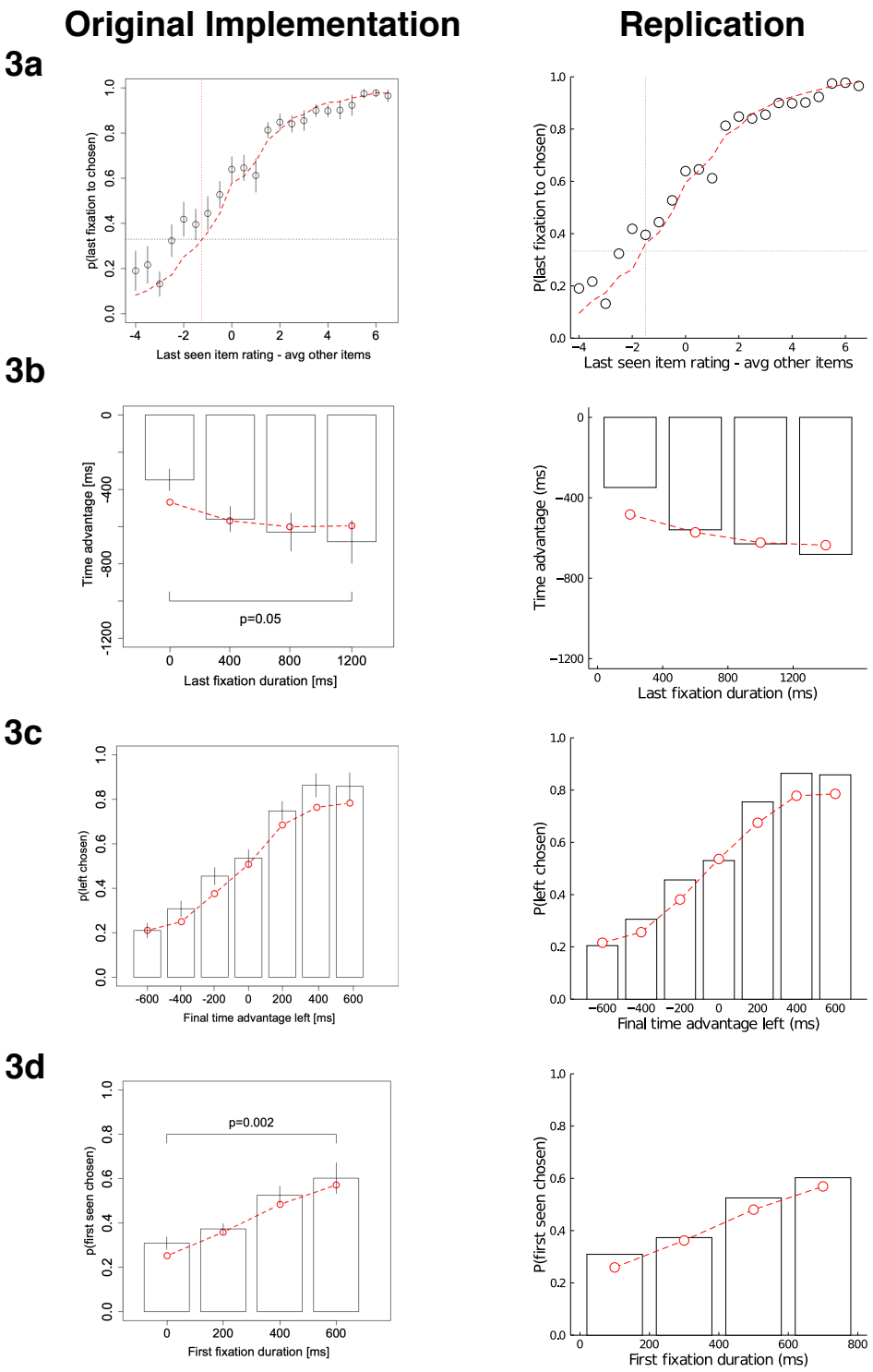

Figure A.4: Replication of Krajbich & Rangel (2011). Note that there are slight deviations in model predictions due to noise in the simulations; the orginal plots are based on 2000 simulated trials.

A.6 Derivations for VOI features

Here we present derivations of the value of information (VOI) features used by the policy approximation method (Callaway et al., 2018) applied to the current metalevel MDP model.

A.6.1 Myopic Value of Information

The myopic value of information is the value of the information acquired by a single computation, that is, the expected increase in decision quality from executing a single computation and then deciding rather than making a decision immediately. Formally,

$$\mathrm{VOI}_{\mathrm{myopic}}(m_t, c) = \mathop{\mathrm{E}}_{m_{t+1}|m_t,c}\left[R(m_{t+1}, \perp)\right]. \tag{A.1}$$

In our model, this is equal to the expected value of the item that will be chosen after taking an additional sample. That is,

$$\mathrm{VOI}_{\mathrm{myopic}}(m_t, c) = \mathop{\mathrm{E}}_{\mu_{t+1}|\mu_t,\lambda_t}\left[\max_i \mu_{t+1}^{(i)}\right]. \tag{A.2}$$

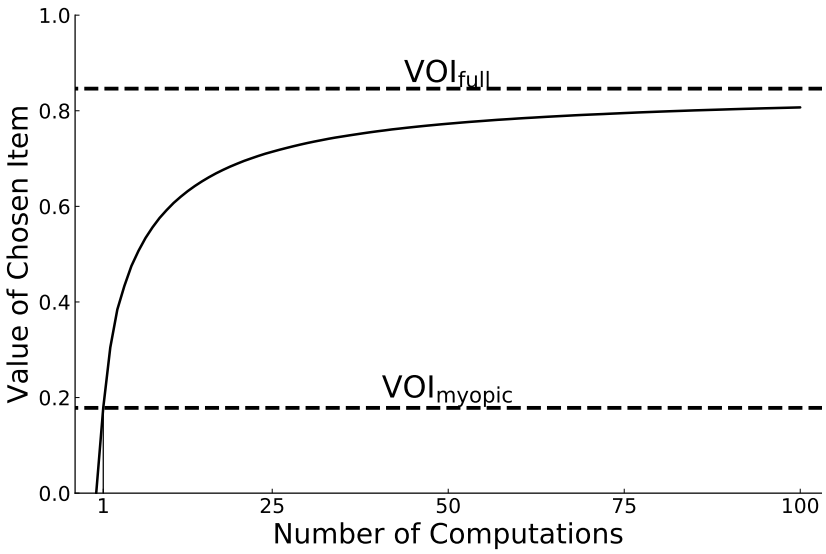

Figure A.5: Illustration of the value of information features. The solid line shows the average value of the item chosen after different numbers of computations selected by a near-optimal policy assuming no computational costs. The dashed lines show values for two of the VOI features in the initial mental state: $\mathrm{VOI}_{\mathrm{myopic}}$ is the value after one computation and $\mathrm{VOI}_{\mathrm{full}}$ is the asymptotitc value after infinite computations.

Because μ_{t+1} differs from μ_t only for item c, we can rewrite the expectation as

$$\mathop{\mathrm{E}}_{\mu_{t+1}^{(c)} | \mu_t^{(c)}, \lambda_t^{(c)}} \left[\max \left\{ \mu_{t+1}^{(c)}, \max_{i \neq c} \mu_t^{(i)} \right\} \right]. \tag{A.3}$$

Thus, the term inside the expectation is the maximum of a constant, $\max_{i \neq c} \mu_t^{(i)}$, and a univariate random variable, $\mu_{t+1}^{(c)} \mid \mu_t^{(c)}, \lambda_t^{(c)}$. To simplify notation, we suppress the conditioning variables in the following derivation.

To derive an analytic expression for Equation A.3, we first derive the distribution of $\mu_{t+1}^{(c)}$, that is, the distribution over the posterior mean after taking a sample. Applying the transition dynamics given in Equation 3.3, we have

$$\mu_{t+1}^{(c)} = \frac{\sigma_x^{-2} x_t + \lambda_t^{(c)} \mu_t^{(c)}}{\lambda_t^{(c)} + \sigma_x^{-2}}. \tag{A.4}$$

Since $x_t \mid u^{(c)} \sim \text{Normal}(u^{(c)}, \sigma_x^2)$ and $\mu_{t+1}^{(c)}$ is a linear transformation of x_t, it follows that $\mu_{t+1}^{(c)}$ is a Gaussian random variable. Additionally, because the mental state is a distribution over the true utility, we have $u^{(c)} \mid \mu_t^{(c)}, \lambda_t^{(c)} \sim \text{Normal}(\mu_t^{(c)}, 1/\lambda_t^{(c)})$. Combining these two statements, we see that $x_t \mid \mu_t^{(c)}, \lambda_t^{(c)}$ is a Gaussian whose mean is itself a Gaussian. Applying the fact that $\text{Normal}(\mu, \sigma^2) = \mu + \text{Normal}(0, \sigma^2)$, we can then derive that $x_t \mid \mu_t^{(c)}, \lambda_t^{(c)} \sim \text{Normal}(\mu_t^{(c)}, 1/\lambda_t^{(c)}) + \text{Normal}(0, \sigma_x)$, which reduces to $\text{Normal}(\mu_t^{(c)}, 1/\lambda_t^{(c)} + \sigma_x^2)$. Finally, applying the linear transformation of x_t given by Equation A.4, we have

$$\mu_{t+1}^{(c)} \sim \text{Normal}(\mu_\mu, \sigma_\mu^2) \tag{A.5}$$

where

$$\mu_\mu = \frac{\sigma_x^{-2}}{\lambda_{t+1}^{(c)}} \mu_t^{(c)} + \frac{\lambda_t^{(c)} \mu_t^{(c)}}{\lambda_{t+1}^{(c)}} = \mu_t^{(c)} \tag{A.6}$$

and

$$\sigma_\mu^2 = \left(\frac{\sigma_x^{-2}}{\lambda_{t+1}^{(c)}} \right)^2 \left(\frac{1}{\lambda_t^{(c)}} + \sigma_x^2 \right). \tag{A.7}$$

Having derived the distribution of $\mu_{t+1}^{(c)}$, we now turn to the expected maximum in Equation A.3. From basic probability theory we know that for any constant z and random variable X,

$$\mathrm{E}[\max\{X, z\}] = \Pr[X \leq z] \cdot z + (1 - \Pr[X \leq z]) \cdot \mathrm{E}[X \mid X > z]. \tag{A.8}$$

Substituting $\max_{i \neq c} \mu_t^{(i)}$ for z and $\mu_{t+1}^{(c)}$ for X, we can use this formula to derive an analytical solution for the myopic value of information. First, we have

$$\Pr\left[\mu_{t+1}^{(c)} \leq \max_{i \neq c} \mu_t^{(i)}\right] = \Phi\left(\beta\right), \tag{A.9}$$

where Φ is the cumulative density function (CDF) of a standard Gaussian , and

$$\beta = \frac{\max_{i \neq c} \mu_t^{(i)} - \mu_{\mu \cdot}}{\sigma_\mu} \tag{A.10}$$

Next, we apply the standard formula for the expectation of a truncated Gaussian, giving us

$$E\left[\mu_{t+1}^{(c)} \mid \mu_{t+1}^{(c)} > \max_{i \neq c} \mu_t^{(i)}\right] = \mu_\mu + \frac{\varphi\left(\beta\right)}{1 - \Phi\left(\beta\right)}\sigma_\mu, \tag{A.11}$$

where φ is the standard normal probability density function. Finally, putting this together we find that $\text{VOI}_{\text{myopic}}(b, c)$ is equal to

$$\Phi(\beta) \cdot \max_{i \neq c} \mu_t^{(i)} + (1 - \Phi(\beta)) \cdot \left(\mu_\mu + \frac{\varphi\left(\beta\right)}{1 - \Phi\left(\beta\right)}\sigma_\mu\right). \tag{A.12}$$

A.6.2 Value of Perfect Information About One Item

Whereas $\text{VOI}_{\text{myopic}}$ captures the information value of a single sample, VOI_{item} captures the information value of an infinite number of samples for one item, that is, the value of knowing the exact value of one item. Formally,

$$\text{VOI}_{\text{item}}(m_t, c) = \mathop{\text{E}}_{u^{(c)} \mid \mu_t^{(c)}, \lambda_t^{(c)}}\left[\max\left\{u^{(c)}, \max_{i \neq c} \mu_t^{(i)}\right\}\right]. \tag{A.13}$$

The derivation is similar to that of $\text{VOI}_{\text{myopic}}$, but instead of taking the expectation over the posterior mean after one computation, $\mu_{t+1}^{(c)}$, we take the expectation over the true utility, $u^{(c)} \mid \mu_t^{(c)}, \lambda_t^{(c)} \sim \text{Normal}(\mu_t^{(c)}, 1/\lambda_t^{(c)})$. Thus, we apply the same steps beginning with Equation A.8, but replacing $\mu_{t+1}^{(c)}$ with $u^{(c)} \mid \mu_t^{(c)}, \lambda_t^{(c)}$. This results in $\text{VOI}_{\text{item}}(b, c)$ equal to

$$\Phi(\beta') \cdot \max_{i \neq c} \mu_t^{(i)} + \left(1 - \Phi(\beta')\right) \cdot \left(\mu_t^{(c)} + \frac{\varphi\left(\beta'\right)}{1 - \Phi\left(\beta'\right)}\sqrt{1/\lambda_t^{(c)}}\right), \tag{A.14}$$

where

$$\beta' = \frac{\max_{i \neq c} \mu_t^{(i)} - \mu_t^{(c)}}{\sqrt{1/\lambda_t^{(c)}}}. \tag{A.15}$$

A.6.3 Value of Perfect Information About All Items

VOI_{full} captures the information value of learning the exact value of every item in the choice set, that is acquiring full information. In this case, the DM will make an exactly optimal choice, gaining the utility of the item that is in fact best. Formally,

$$\text{VOI}_{\text{full}}(b) = \mathop{\text{E}}_{u|\mu_t^{(c)},\lambda_t^{(c)}} \left[\max_i \left\{ u^{(i)} \right\} \right]. \tag{A.16}$$

For the case of n items, the conditional expectation term is given by the integral

$$\int \dots \int \left[\max_i u^{(i)} \prod_{i=1}^{k} \text{Normal}(u^{(i)}; \mu_t^{(i)}, 1/\lambda_t^{(i)}) \right] du^{(1)} \dots du^{(n)}. \tag{A.17}$$

Unfortunately, there is no analytic solution to this integral. However, we can substantially reduce our computational burden by reducing to a piecewise one-dimensional integral.

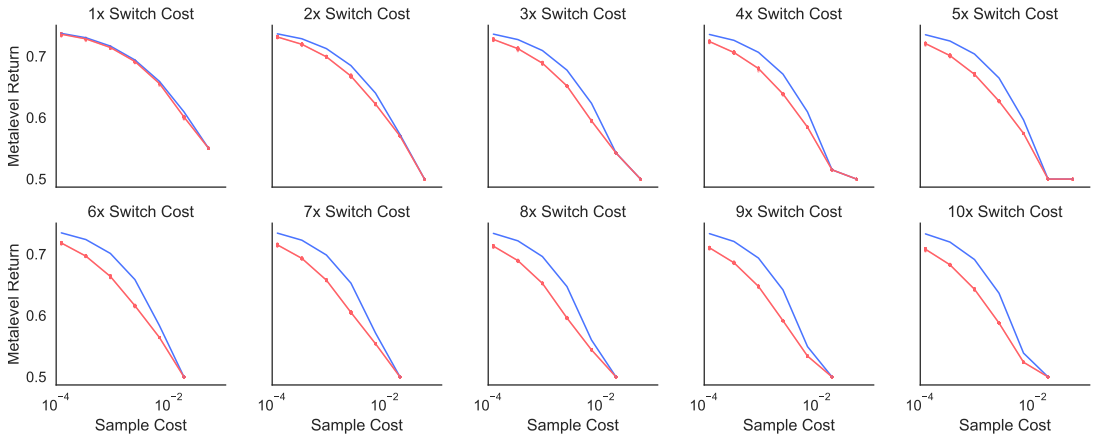

Figure A.6: Performance of BMPS on the Bernoulli model with switching costs. Each line shows the average metalevel return achieved by the BMPS policy (red) or the true optimal policy (blue). The red line shows mean performance from the top 80 policies identified by the UCB algorithm. Additionally, each individual policy's performance is plotted as an individual point, but performance is so consistent that the points are not visually distinct.

First, we can express the expectation of any random variable as a piecewise integral,

$$\mathrm{E}[X] = -\int_{-\infty}^{0} F_X(x)\,dx + \int_{0}^{\infty} (1 - F_X(x))\,dx, \tag{A.18}$$

where F_X is the CDF of X. Next, we can express the CDF of the maximum of a set of random variables as the product of the CDF for each variable alone,

$$F_{\max \mathcal{X}}(x) = \prod_{X \in \mathcal{X}} F_X(x), \tag{A.19}$$

because the maximum of a set is less than x if and only if each element in the set is less than x. In our case, the set χ contains the posterior distributions for each item. Letting M denote $\max \mathcal{X}$, we can define its CDF as

$$F_M(m) = \prod_{i=1}^{n} \Phi\left(\sqrt{\lambda_t^{(i)}}\left(m - \mu_t^{(i)}\right)\right). \tag{A.20}$$

Combining Equations Equation A.16, Equation A.18, and Equation A.19, we arrive at the following expression for $\mathrm{VOI}_{\mathrm{full}}(b)$:

$$-\int_{-\infty}^{0} F_M(x)\,dx + \int_{0}^{\infty} (1 - F_M(x))\,dx. \tag{A.21}$$

We evaluate these two integrals numerically to a minimum precision of 10^{-5} by the adaptive Gauss-Kronrod quadrature method implemented in the QuadGK Julia package.

Despite the dimensionality reduction, we found that evaluating these integrals was still the primary computational bottleneck for simulating the model. Thus, in order to reduce computation time, we only compute $\mathrm{VOI}_{\mathrm{full}}$ when it is necessary to determine which computation the policy will execute. As detailed below, this is often unnecessary because the other features already determine which feature has maximal Q^{bmps}.

Critically, the modification that we describe here has no effect on the behavior of the policy or the predictions of the models; we have verified this assertion through simulation.

This computational trick is based on three insights. First, note that $\mathrm{VOI}_{\mathrm{full}}$ helps to decide whether or not to take another sample, but not which item to sample from. Thus, we can determine which computation the policy would take, conditional on taking a sample at all, based only on the $\mathrm{VOI}_{\mathrm{myopic}}$ and $\mathrm{VOI}_{\mathrm{item}}$ features. Given that these two features have an analytical solution, as derived above, we can quickly identity the best item to sample from,

which is given by

$$c^* = \underset{c \neq \perp}{\operatorname{argmax}} \left\{ w_1 \cdot \mathrm{VOI}_{\mathrm{myopic}}(b, c) + w_2 \cdot \mathrm{VOI}_{\mathrm{item}}(b, c) - \mathrm{cost}(c) + w_4 \right\}. \tag{A.22}$$

Second, since $Q(b, \perp) = R(b, \perp)$, it follows that if $Q^{\mathrm{bmps}}(b, c^*) > R(b, \perp)$, the policy should sample from item c^*, and otherwise it should stop sampling. In general, evaluating this inequality requires evaluating $\mathrm{VOI}_{\mathrm{full}}(b)$. However, in some cases it can be determined without knowing $\mathrm{VOI}_{\mathrm{full}}(b)$. In particular, we can take advantage of the fact that $\mathrm{VOI}_{\mathrm{item}}(b, c) \leq \mathrm{VOI}_{\mathrm{full}}(b)$ for all b, c. We can thus compute a lower bound on $Q^{\mathrm{bmps}}(b, c)$ by replacing $\mathrm{VOI}_{\mathrm{full}}(b)$ with $\mathrm{VOI}_{\mathrm{item}}(b, c)$ in Equation A.22. If this lower bound is positive, then we know the full approximation would also be positive, and thus the optimal choice is to sample from item c^*. Otherwise, we compute $\mathrm{VOI}_{\mathrm{full}}(b)$ and identify the optimal computation using all of the features.

Third, at first sight this approach might seem to be incompatible with the soft-maximizing policy, where computation c is selected with probability proportional to $\exp \beta Q^{\mathrm{bmps}}(b, c)$. In particular, the standard method for sampling from this distribution requires fully evaluating $Q^{\mathrm{bmps}}(b, c)$. However, we can circumvent this issue using the Gumbel-max trick (Yellott Jr, 1977), which provides a way to sample from a Boltzman (softmax) distribution by taking the argmax of the unexponentiated values corrupted by Gumbel noise. Formally,

$$\Pr\left[\underset{i}{\operatorname{argmax}} \{x_i + \varepsilon_i\} = j \right] = \frac{\exp x_j}{\sum_i \exp x_i}. \tag{A.23}$$

As a result, we can rewrite the soft-max policy as

$$\pi(b; \mathbf{w}, \beta) = \underset{c}{\operatorname{argmax}} \left\{ \beta Q^{\mathrm{bmps}}(b, c) + \varepsilon_c \right\}, \tag{A.24}$$

where $\varepsilon_c \sim \mathrm{Gumbel}(0, 1)$. We can then implement steps 1 and 2 of the short-cut, adding ε_c to the right hand side of Equation A.22, and comparing the lower-bound of $Q^{\mathrm{bmps}}(b, c)$ to $R(b, \perp) + \varepsilon_\perp$.

A.7 Quality of the approximation method in Bernoulli model

The approximation method used here has previously been shown to learn policies with near-optimal performance on a metalevel MDP similar to the one in the present model, but with Bernoulli-distributed samples and *no* switching costs (Callaway et al., 2018). The logic of the

problem is identical: A DM wants to select the best item and informs her decision by drawing noisy samples with an expected value equal to the items' true utility. However, in the simpler Bernoulli case that has been previously studied, true utilities take values between 0 and 1, samples from item c are drawn from Bernoulli$(u^{(c)})$, and the uniform distribution over all possible utilities, Beta$(1, 1)$, provides a conjugate prior. Thus, posterior beliefs take the form Beta$(1+\alpha, 1+\beta)$, where α and β are respectively the number of times 1 and 0 have been sampled for the given item. Critically, the resulting belief space is discrete because α and β are integers. This allows the computation of the exact optimal policy by dynamic programming, if an upper bound on the number of samples that can be taken is assumed.

Callaway et al. (2018) take advantage of this fact to show that the policy approximation method used here provides a highly-accurate approximation of the optimal policy. However, their model does not have switching costs, which could potentially make the approximation perform much worse. Here, we investigate this issue by adding switching costs to the Bernoulli model, and measuring their impact on the method's performance. Ideally we would be able to directly assess the performance of our method in the full model with Gaussian samples, but an optimal solution for this case is not available and deriving one is beyond the scope of the study.

To aid interpretation, we re-parameterized the switching cost as $\gamma_{\text{switch}} = (k-1)\gamma_{\text{sample}}$ such that k can be interpreted as a multiplier on the base sample cost. For example, $k = 1$, indicated by "1x" in the figure, corresponds to no switching cost. We considered a grid of cost parameters with $\gamma_{\text{sample}} \in \{e^{-9}, e^{-8}, \ldots, e^{-3}\}$ and $k \in 1, 2 \ldots, 10$. We set an upper bound of 75 samples. As shown in Figure A.6, we replicated previous results that the approximated policy is nearly optimal when there is no switch cost. As the figure shows, relative performance degrades somewhat when switch costs are added, but the approximation still achieves 92% of the optimal metalevel reward in the worst case explored.

B

Supplementary information for Chapter 4

B.1 Deviations from pre-registration

After pre-registering and running the experiments, we discovered a flaw in the implementation of the lesioned models without meta-level control. In our first implementation, stopping times and fixation durations were sampled directly from the empirical distribution in the human data (discretized into 100ms time steps). We reasoned that this would give the model the best chance of capturing the human behavior without a metacognitive component. However, we later realized that this approach effectively prevents the model from utilizing non-decision time, as any non-decision time would be added to a distribution that already perfectly matches the human data.

After realizing this, we implemented the more flexible lesioned model with stopping times and fixation durations drawn from arbitrary Gamma distributions, as reported in the main text. Note that this model has more free parameters in the two-item case, thus we also decided to fit the lesioned model to data in Experiment 2 (contrary to our intention to use the fitted parameters from Experiment 1). The predictions of the original model for both experiments are shown in Figures B.2-B.5.

This analysis also revealed that two of the effects which we had believed to be evidence of meta-level control could in fact be produced by the model without meta-level control. The first such effect is shown in Figure B.1B. The original logic of this plot was to look at how strength affects the decision to skip a trial over time, after factoring out the effect of strength

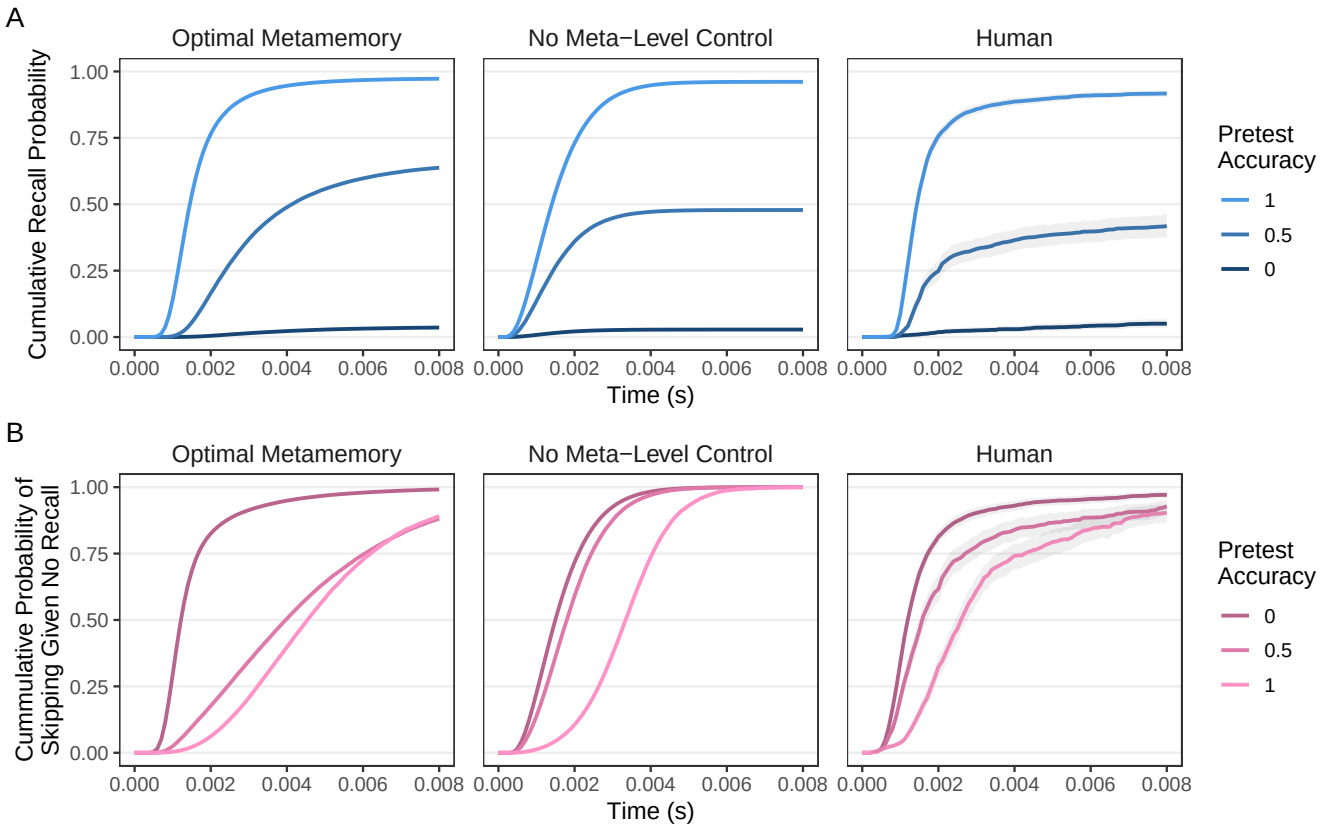

Figure B.1: Timecourse of recall and stopping. (A) For each time point (in 100ms steps), the proportion of trials on which the target was recalled before that point, grouping trials by the accuracy rate of the presented cue in the pretest phase. (B) For each time point, the proportion of trials that were skipped before that point, conditioning on the fact that the target had not already been recalled. *Note*: Lines show means of participants means and ribbons show 95% bootstrapped confidence intervals over participant means.

on recall (which in turn prevents skipping). With a low or moderate non-decision time distribution, this effect cannot be captured by a model without meta-level control. However, when non-decision time is a sufficiently large portion of total response time, the effect can be produced through mechanistic means. Given this, we could not provide a useful interpretation of the effect, and thus removed it from the main text.

The second effect is the prediction that final fixations are longer than non-final fixations (Figure 4.10A). We elected to keep this effect in the main text because its interpretation as an indication of rational metamemory was further supported by the additional analysis of the distribution of pretest accuracy for long fixations (Figure 4.10B).

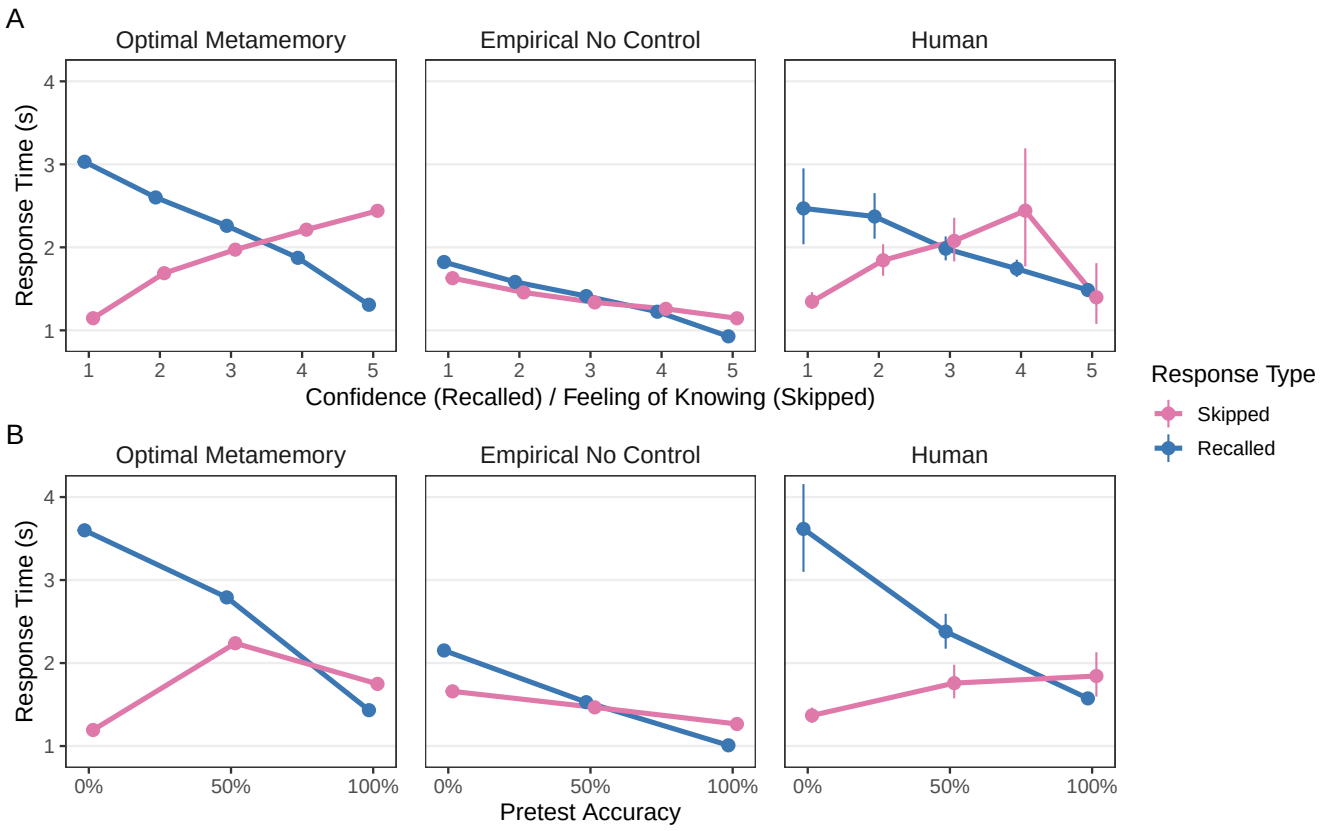

Figure B.2: Alternative version of Figure 4.5 where the lesioned model draws stopping times from the empirical distribution.

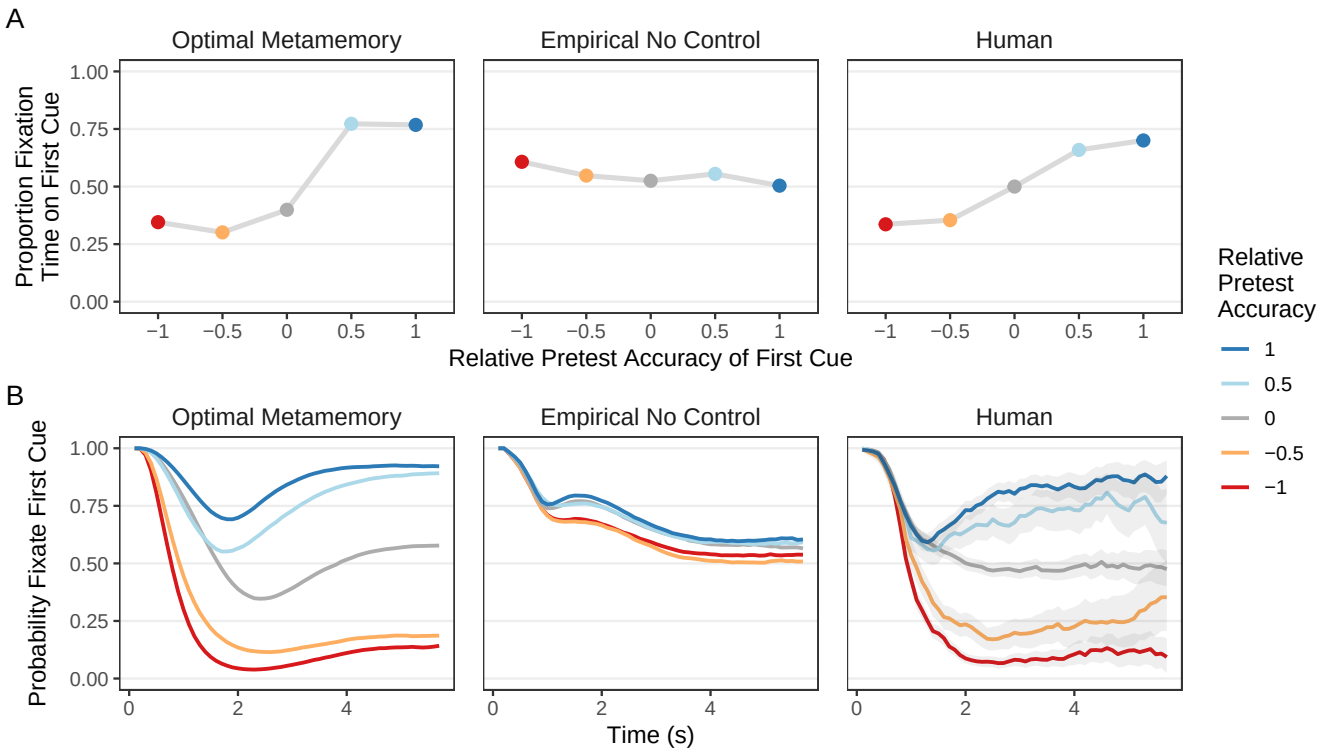

Figure B.3: Alternative version of Figure 4.8 where the lesioned model draws stopping and switching times from the empirical distribution.

B.2 Previously pre-registered experiments

Before running the experiments presented in the main text, we ran another set of large experiments using the exact same experimental design. These were also pre-registered. However, the results from these studies were not conclusive (for different reasons), so we reran them.

For Experiment 1, we initially planned to run 125 participants, and to z-score response times within participants before regressing them on pretest accuracy. This analysis yielded a marginally significant effect of pretest accuracy on response time for skip trials ($p = .062$). The original pre-registration is available at `https://aspredicted.org/ss8x3.pdf`. While analyzing the data, we also discovered that the z-scoring step actually reduced statistical power, as it selectively diminished the contribution of participants who showed the effect most robustly (and thus had more variable response times). We thus eliminated the z-scoring step, pre-registered the slightly modifide analysis, and reran the experiment with a target of 500 participants. Note that the mixed effects analysis still accounts for individual

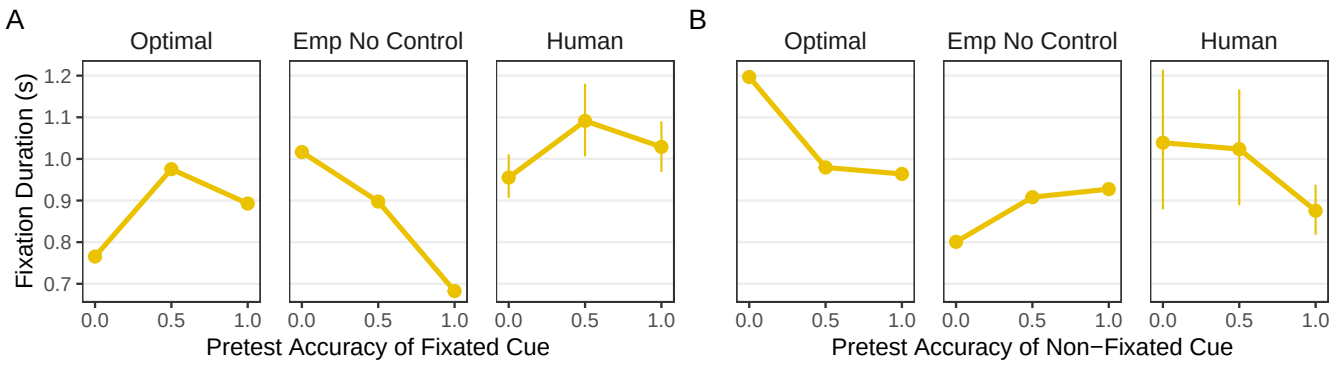

Figure B.4: Alternative version of Figure 4.9 where lesioned model draws stopping and switching times from the empirical distribution.

variability in both average response times and sensitivity to pretest accuracy.

for Experiment 2 (which was actually run first), we had planned to use a definition of memory strength that combined response time and accuracy on the pretest trials. The logic was that a fast response indicated a stronger memory; thus, highers response time in the pretest should predict less fixation time in the critical trials. However, we then discovered that response time on the pretest was actually *positively* correlated with fixation durations in the critical trials (contrary to our predictions). This may be due to factors other than memory strength, such as perceptual encoding, which would uniformly increase the amount of time spent looking at an item. Additionally, we had planned a different set of analyses, focusing on the duration of individual fixations as a function of relative strength of the two cues, rather than breaking down the effect of the fixated and non-fixated cues, as we do now. The full pre-registration is available at `https://aspredicted.org/ss8x3.pdf`. Given the extent of the changes we wished to make to the analysis, we pre-registered the new analysis plan and reran the experiment.

As shown in the sections below, all of the results reported in the main text are also significant in this previous sample. We reproduce all the figures from the main text in Figures B.6-B.9.

B.2.1 Experiment 1

We recruited 124 participants and excluded 14 (11%) participants who did not provide a response on more than 90% of critical trials. This yielded 110 participants in our final analysis.

Participants were faster to correctly recall targets that they reported greater confidence in ($B = -0.203$, 95% CI [-0.289, -0.117], $t(119.2) = -4.62$, $p < .001$), but slower to

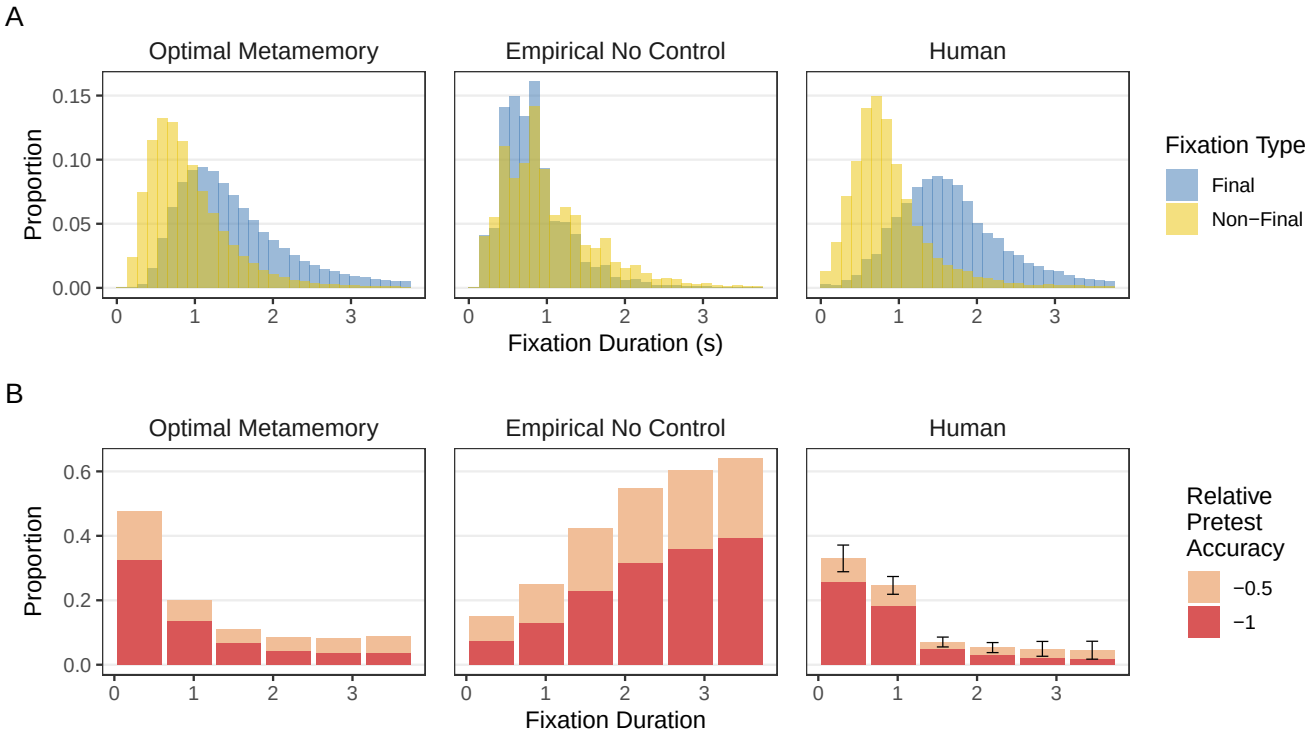

Figure B.5: Alternative version of Figure 4.10 where lesioned model draws stopping and switching times from the empirical distribution.

skip targets that they reported higher feeling-of-knowing for ($B = 0.474$, 95% CI [0.302, 0.646], $t(35.4) = 5.41$, $p < .001$). People were likewise faster to recall targets that they had previously recalled correctly ($B = -1.160$, 95% CI [-1.724, -0.597], $t(41.2) = -4.03$, $p < .001$). More importantly, they were also slower to skip such targets ($B = 0.478$, 95% CI [0.256, 0.700], $t(151.2) = 4.22$, $p < .001$).

B.2.2 Experiment 2

We recruited 459 participants and excluded 65 (14%) participants who failed to correctly recall a target on more than 50% of critical trials. This yielded 394 participants in our final analysis.

Participants spent substantially more time looking at cues that were stronger than the other available cue ($B = 0.191$, 95% CI [0.182, 0.199], $t(572.0) = 42.40$, $p < .001$). Their non-final fixations increased with the pretest accuracy of the fixated cue ($B = 0.106$, 95% CI [0.075, 0.137], $t(230.7) = 6.65$, $p < .001$) and decreased with the pretest accuracy of the non-fixated cue ($B = -0.456$, 95% CI [-0.578, -0.334], $t(59.7) = -7.32$, $p < .001$; first

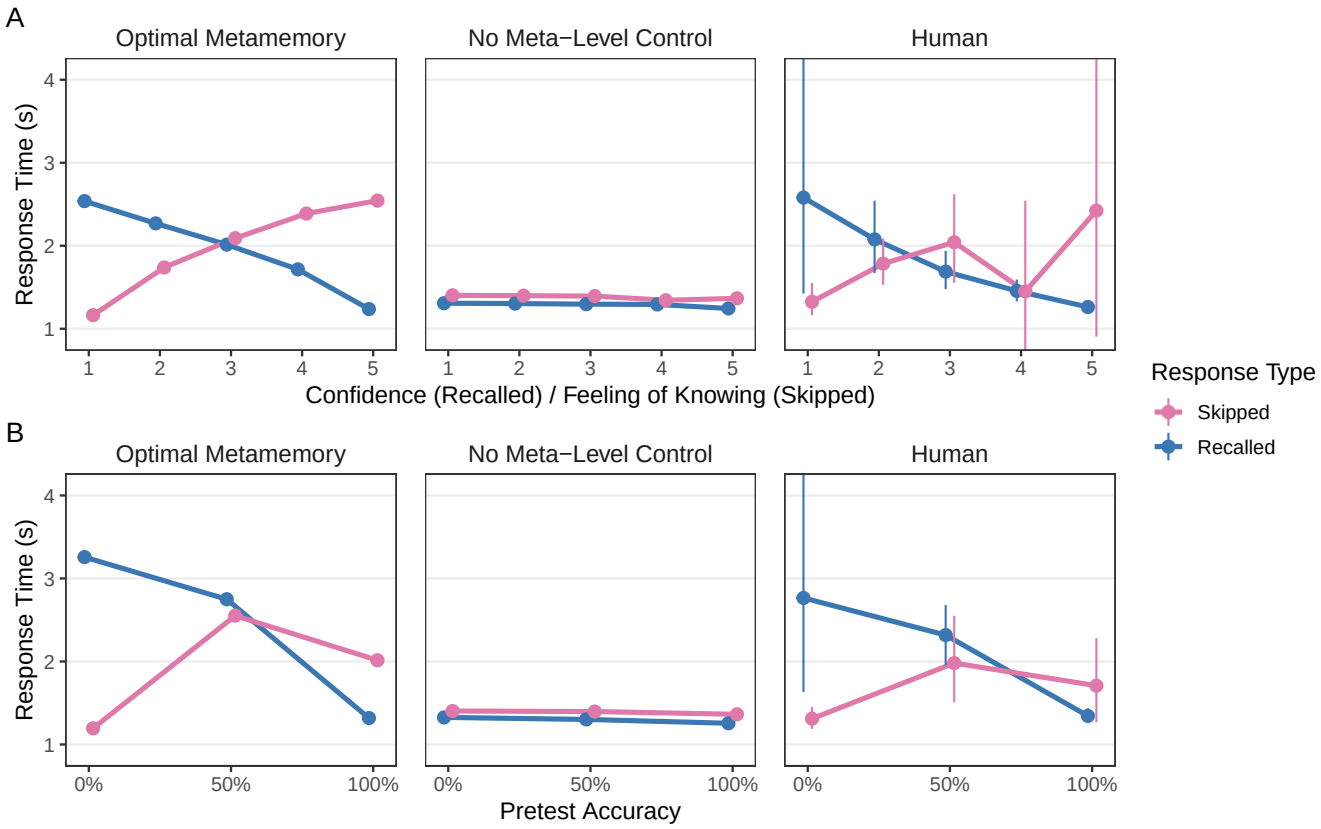

Figure B.6: Figure 4.5 with previous experimental data. The models are fit to the data shown in the plot. We use the same axis limits as in the main text to facilitate comparison (the error bars extend beyond the plotted range).

fixations excluded). Final fixations were longer than their non-final fixations ($B = 0.872$, 95% CI [0.813, 0.930], $t(435.8) = 29.29$, $p < .001$). The probability of fixating the weaker memory significantly decreased with fixation duration ($B = -1.482$, 95% CI [-1.651, -1.313], $z = -17.22$, $p < .001$; logistic regression, excluding trials where the cues had equal pretest accuracy).

The final result is notable because it confirms the exploratory rational commitment analysis we developed using the new dataset. Thus, although the result in the main text was not pre-registered, the result presented above effectively is; we had finalized the analysis before examining the results it yielded with the old dataset.

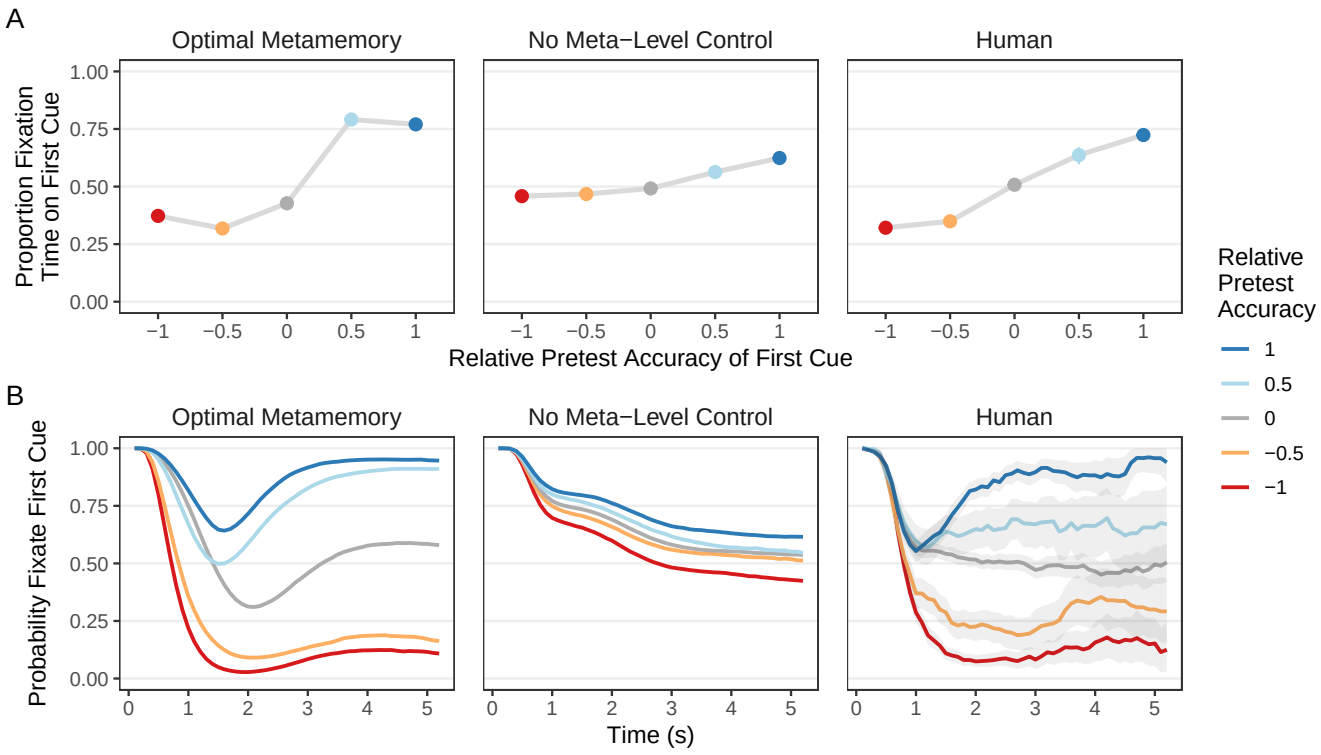

Figure B.7: Figure 4.8 with previous experimental data.

B.3 Optimizing the lesioned model to predict effects

To more conclusively determine which effects are inconsistent with the lesioned model, we conducted a thorough search of the parameter space to see if the model could produce each qualitative effect under any parameter setting. For each experiment, we sampled 100,000 parameter configurations and simulated 100,000 trials for each. From these, we excluded simulations with accuracy below 1% or above 99%, as these yield unreliable estimates for effects that condition on accuracy. For similar reasons, in Experiment 2, we excluded configurations for which fewer than 1% of trials had at least two fixations. For each un-excluded simulation, we then performed a standard linear regression corresponding to the regression we reported in the main text. Next, we selected the 100 configurations who produced the largest effect, as estimated by the lower bound of a 95% confidence interval (this prevents selecting for models that simply produce highly variable estimates). If the lower confidence bound for any of these was larger than a "minimal interesting difference", then we concluded that the lesioned model could produce the effect.

Note that this analysis occurred to us after running the experiment, and was thus not pre-

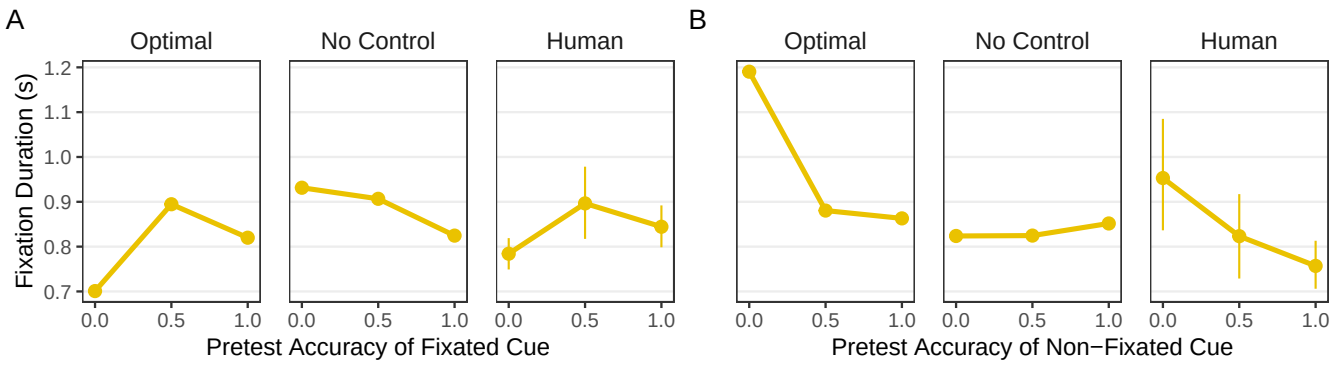

Figure B.8: Figure 4.9 with the old previous experimental data.

registered.

B.3.1 Experiment 1

As expected, we found that the lesioned model could predict the observed negative effect both judgments ($B = -0.762$; 95% CI [-0.765, -0.759]) and pretest accuracy ($B = -6.073$; 95% CI [-6.109, -6.037]) on response time for the correct trials. Note that these are the largest effect sizes we found; they are substantially larger than those seen in the data. Surprisingly, we also found that the lesioned model could predict a substantial positive relationship between judgment and response time on the skip trials ($B = 0.615$; 95% CI [0.605, 0.626]). However, this model failed to predict the negative relationship on correct trials ($B = 0.052$; 95% CI [0.050, 0.054]; note that B should be negative). No parameter configuration was able to predict the crossover pattern, with a negative relationship between judgment and response time on correct trials but a positive relationship on skip trials. Furthermore, no configuration was able to predict the positive relationship between pretest accuracy and response time on skip trials (even allowing the relationship for correct trials to be positive).

B.3.2 Experiment 2

Consistent with Figure 4.8A and 4.10A, the lesioned model was able to capture the overall-proportion ($B = 0.315$; 95% CI [0.309, 0.321]) and the long-final-fixation ($B = 2.370$; 95% CI [2.368, 2.373]) effects. Perhaps surprisingly, we found that the lesioned model was also capable of capturing the both fixation duration effects (fixated: $B = 0.113$; 95% CI [0.107, 0.120]; non-fixated: $B = 0.095$; 95% CI [0.087, 0.102]). However this configuration

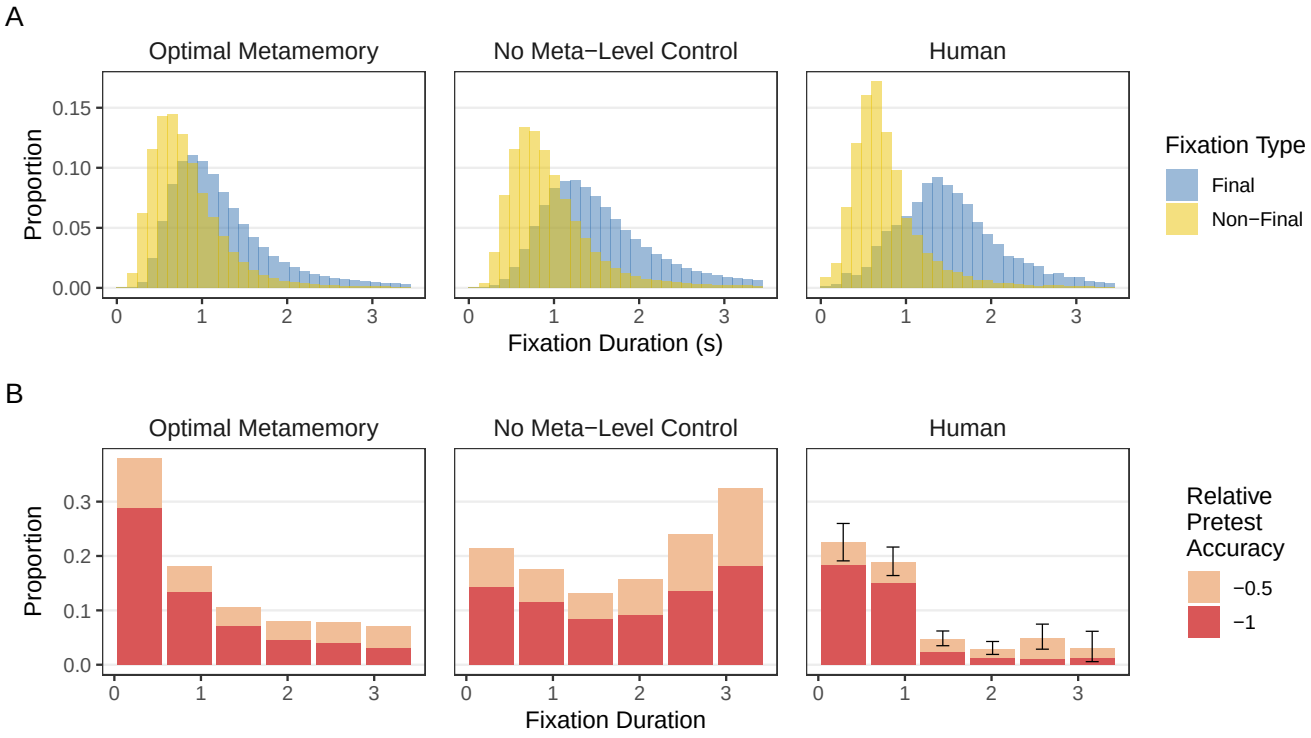

Figure B.9: Figure 4.10 with previous experimental data.

(which, not coincidentally, maximized both effects) achieved an accuracy of only 7.3%. The model was able to capture the effect through a selection effect. It only provided a correct response on the small percentage of trials in which it happened to sample long fixations on the strongest cues. Running the analysis again with the requirement that the model achieve at least 65% accuracy (compared to 84% in the human data), we found that no configuration could capture the non-final fixation duration effects.

C

Supplementary information for Chapter 5

C.1 Deviations from pre-registration

Here we document all deviations our pre-registered analysis plans.

For several experiments we recruited slightly more participants than originally intended as a result of participants completing the experiment after being flagged as incomplete by Prolific.

In Experiments 1 and 3, we pre-registered proportion tests for best-first search and the forward search bias. We switched to Wilcoxon tests over participants means as this test correctly respects the group structure of the experiment. The qualitative conclusions are the same with either approach. The original results were $z = 72.1,\ p < .001$ for best-first search and $z = 51.1,\ p < .001$ for expansion.

Similarly, in Experiments 1 and 4, we initially performed fixed effects logistic regressions, but decided that mixed-effects regressions were more principled. The qualitative conclusions are the same with either approach, although the coefficients are substantially larger with the mixed-effects regression. For Experiment 1, the fixed-effects regression coefficient for best path value is $\beta = 0.579$ (95% CI [0.521, 0.637], $z = 19.6, p < .001$); for best vs. next, $\beta = 1.111$ (95% CI [1.059, 1.163], $z = 41.9, p < .001$. This is compared to $\beta = 0.539$ and $\beta = 1.900$ in the optimal model. For Experiment 4, the fixed-effects regression coefficient for value on violation of forward search is $\beta = 0.579$, 95% CI [0.465, 0.694], $z = 9.9$, $p < .001$.

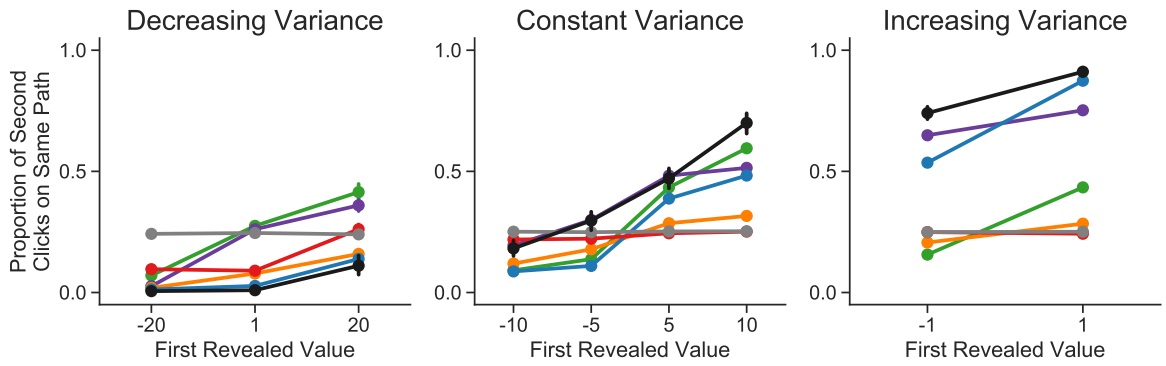

Figure C.1: Alternative version of Figure 5.4D. Here, the predictions of all alternative models are genreated with the best-fitting set of heuristic mechanisms. As in Figure 5.4D, points show means and error bars show bootstrapped 95% confidence intervals, both computed across participants.

In Experiment 1, we pre-registered that we would report the difference in likelihood between the optimal model and the best model without best vs. next. We ultimately decided that this comparison was not of special interest. The likelihood difference is $\Delta_{\text{LL}} = 1911$ in favor of the optimal model.

For Experiment 2, we pre-registered that we would report the difference in likelihood between the best-fitting model and the next-best-fitting model. However, we later decided that the difference from the optimal model was a more useful comparison in the constant variance case, where the optimal model did not fit best. The best-fitting model in that case was the best-first model with best vs. next and depth limits. The next best model (excluding other best-first variants) was depth-first with satisficing, depth limits, and pruning ($\Delta_{\text{LL}} = 885$).

For Figure 5.44d, we pre-registered that we would plot the predictions of the same models shown in main-text Figure 5.44d. However, as illustrated in Figure C.1, the pruning mechanism allows depth-first search to closely mimic best-first search on the second click (although not necessarily on later clicks, as the model comparison reveals). We ultimately decided that it was more important to convey the qualitative difference between the different search orders than to show how the heuristic mechanisms can improve the fit to data. For this reason, we switched to plotting the predictions of the best-, breadth-, and depth-first models without any heuristic mechanisms.

For Experiment 3, we neglected to mention that we would not consider pruning and depth limits in the model comparison. As stated in the main text, it is not clear how to generalize these mechanisms to the case where there is not a forward search constraint. Furthermore, the most obvious generalizations that effectively treat each node independently

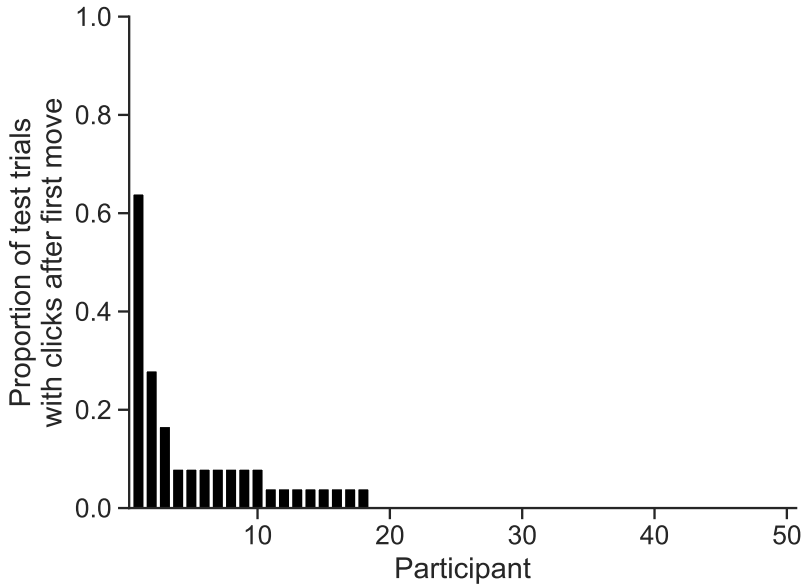

Figure C.2: Experiment 5 results. The proportion of test trials in which each participant made at least one click after moving the spider.

(rather than operating on branches of a decision tree) make computing the likelihood excessively computationally intensive due to the need to marginalize over all possible pruning decisions (see Methods). Importantly, the initial omission was simply an error in the preparation of the pre-registration document. We decided to omit these mechanisms long before running the experiment, when we discovered that we could not apply the existing implementation of pruning and depth limits to pilot data.

C.2 Experiment 5: Interleaved planning and action

In Experiments 1-3, we constrained participants to do all their planning (clicking) before taking their first action (moving the spider). We constrained the task in this way for two reasons: First, the optimal policy always does all its planning before taking any action; this is because moving cannot inform future planning but further planning could make one regret moves one has already made (in which case, one should have done that planning earlier). Second, allowing participants to violate this principle would require modifying the model to account for this possibility; this would complicate the implementation substantially. However, one could argue that by constraining participants to do all of their planning upfront, we are forcing them to follow the optimal strategy in this regard. It is thus important to know whether people would violate the principle if given the opportunity.

To address this question, we ran an experiment that was exactly like Experiment 1 except that we allowed participants to click at any time (even after moving). This was visually indicated by highlighting the clickable frontier nodes after each movement, with the frontier expanding to include states adjacent to the spider's new location. The results are illustrated in Figure C.2. We found that participants very rarely chose to click after moving the spider. Although 36 out of 50 participants (72%) revealed a reward after moving the spider on at least one practice trial (suggesting that they understood it was possible to do so), only 3 participants (6%) clicked after moving on more than 2 out of the 25 test trials. Overall, participants clicked after moving on 3.9% of test trials. This perhaps reflects a sensitivity to the informational asymmetry of moving and planning described above.

C.3 Probability-based stopping rules

In the main text, we considered two heuristic stopping rules based on the expected values of the best and second-best paths, satisficing and best vs. next. These stopping rules are truly heuristic, in the sense that they are very easy to compute. However, by reducing paths to their expected value, they potentially throw away useful information about the *distribution* of possible values the path could take. Thus, we also considered more sophisticated variants of each stopping rule which are based on probabilities rather than expected values. These stopping rules involve extensive computation (concretely, marginalizing over joint distributions over possible path values) and are thus not truly "heuristic". However, they can potentially provide insight into which aspects of our partcipants' reasoning the heuristic models fail to capture.

C.3.1 Model

First, some notation. Let V_i be a random variable describing the distribution of possible values path i could have and let b be the path with highest expected value (we address ties below).

The probabilistic satisficing term is defined $\Pr(V_b \geq \alpha)$ where V_b is a the true value of the best path and α is a threshold. If multiple paths have maximal expected value, we compute the term for all paths and use the maximum. We refer to this component of the stopping rule as "prob better" as it gives the probability that the best path is better than some value.

The probabilistic best vs. next term could be defined in several ways, the primary decision being whether to choose the competing value based on the current expected values or instead based on hypothetical true values. We chose the latter as it produces an intuitively

appealing quantity: the probability that the path with best expected value in fact has maximal value. Thus, abusing the max operator slightly, we have $\Pr(V_b \geq \max_{i \neq b} V_i)$ where $\max_{i \neq b} V_i$ should be interpreted as a random variable describing the maximum value of any path (excluding b although this constraint is irrelevant in this case). As with satisficing, in the case of ties, we use the maximal value. We refer to this component of the stoping rule as "prob best".

The extended heuristic model adds these two new terms (with accompanying β weights) to Equation 5 in the main text. Rewriting the original satisficing and best vs. next rules with the new notation, we have

$$\begin{aligned} f_{\text{stop}}(s) = &\beta_{\text{satisfice}} \cdot \mathrm{E}[V_b] + \\ &\beta_{\text{bestnext}} \cdot \left(\mathrm{E}[V_b] - \max_{i \neq b} \mathrm{E}[V_i] \right) + \\ &\beta_{\text{prob-better}} \cdot \Pr(V_b \geq \alpha) + \\ &\beta_{\text{prob-best}} \cdot \Pr\left(V_b \geq \max_{i \neq b} V_i \right) + \theta_{\text{stop}}. \end{aligned} \tag{C.1}$$

C.3.2 Results

In Experiment 1, adding the two new terms to the heuristic models substantially improved fit ($\Delta_{\text{LL}} = 713$). This improvement was driven entirely by the "prob best" rule; the "prob-better" term did not improve overall fit either alone or in addition to the "prob-best" rule. Although the optimal model still performed better in terms of total likelihood, 49 participants were best fit by the one of the best-first models vs. 37 by the optimal model (compare to 41 vs. 45 without the new terms). However, because this metric selects which heuristic mechanisms to include based on performance on the test set, it is an overestimate of the true predictive performance of the best-first search model. Looking instead at individual models (i.e. combinations of stopping rules), no model fit more than half of participants better than the optimal model in a head-to-head contest (45 vs. 50 at most). Thus, there is not good evidence that the augmented heuristic models out-perform the optimal model in terms of number of participants fit.

As shown in Table C.2, the remaining experiments paint a similar picture. In general, the "prob best" term improves fits, sometimes dramatically. However, this improvement does not push the heuristic models' performance above the optimal model's with the exception of the constant variance condition of Experiment 3 where it gives the heuristic model a 53 point lead. The "prob better" term provides a smaller boost, rarely improving fit when "prob

best" is already included; one exception is the constant variance condition of Experiment 2 where including "prob better" improves log, likelihood by 87 points.

Model Class	1 Constant	2 Decreasing	2 Constant	2 Increasing	3 Decreasing	3 Constant	3 Increasing	4 Constant
Best	23272	20037	28269	22056	28039	27420	41049	6457
Depth	26073	20073	29241	14888	31219	32583	37832	6535
Breadth	26003	19591	31473	24114	29916	32419	42901	6480
Optimal	22022	18128	30419	14283	24978	26911	34171	5740
Myopic	27832	19404	35219	25418	26771	30726	36589	6035
Random	32868	25957	35988	28490	33905	35463	45468	6562

Table C.1: Model comparison for different different search orders across all experiments. Each column shows one condition of one experiment. Each number is the minimum negative log, likelihood achieved by any model with the search order specified in the left column. Likelihoods for each model and participant, as well as fitted parameters are available at https://osf.io/6venh/.

Model Class	1 Constant	2 Decreasing	2 Constant	2 Increasing	3 Decreasing	3 Constant	3 Increasing	4 Constant
Basic -Satisfice	23426	19617	28269	14888	28337	27569	38785	6457
Basic -BestNext	23933	19700	28555	14892	29927	28568	37809	6500
Basic -Forward					33113	38897	39531	6986
Basic	23272	19591	28269	14888	28039	27420	37809	6457
Basic +ProbBetter	23272	19487	28090	14888	28039	27420	37759	6449
Basic +ProbBest	22560	19336	27838	14870	27359	26858	37313	6005
Basic +Both	22560	19336	27751	14870	27359	26858	37313	5999
Optimal	22022	18128	30419	14283	27946	32994	35909	7826
Optimal +Forward					24978	26911	34171	5740

Table C.2: Model comparison for different combinations of heuristic mechanisms across all experiments. Each column shows one condition of one experiment. Each number is the minimum negative log, likelihood achieved by any heuristic model with the set of mechanisms specified in the left column. “Basic” corresponds to the five mechanisms we consider in the main text: satisficing, best vs. next, depth limits, pruning, and forward-planning bias. “ProbBetter” and “ProbBest” are the two additional terms in the stopping rule described above.

References

Acharya, A., Chen, X., Lewis, R. L., & Howes, A. (2017). Human Visual Search as a Deep Reinforcement Learning Solution to a POMDP. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 51–56).

Agrawal, M., Mattar, M. G., Cohen, J. D., & Daw, N. D. (2022). The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. *Psychological Review*, 129(3), 564–585.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, (pp. 503–546).

Anderson, B. A. (2016). The attention habit: How reward learning shapes attentional selection. *Annals of the New York Academy of Sciences*, 1369(1), 24–39.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, 85(4), 249–249.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Psychology Press.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409–429.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, 51(4), 355–365.

Anderson, J. R. & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4), 703–719.

Armel, K. C., Beaumel, A., & Rangel, A. (2008). Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making*, 3(5), 396–403.

Armel, K. C. & Rangel, A. (2008). Neuroeconomic models of economic decision making: The impact of computation time and experience on decision values. *American Economic Review*, 98(2), 163–168.

Ashby, F. G. & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, 39(2), 216–233.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3), 183–193.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3), 235–256.

Bacon, P.-L., Harb, J., & Precup, D. (2016). The Option-Critic Architecture.

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H. R., Reid, A., Verfaellie, M., Shadlen, M. N., & Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *eLife*, 8, undefined–undefined.

Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp.0). The MIT Press.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, 19(2), 11.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327–18332.

Bennett, S. T., Benjamin, A. S., & Steyvers, M. (2017). A Bayesian model of knowledge and metacognitive control: Applications to opt-in tasks. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp.6).

Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb), 281–305.

Berkowitsch, N. A., Scheibehenne, B., & Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, 143(3), 1331–1348.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, 59(1), 65–98.

Bhui, R. & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, 125(6), 985–1001.

Biderman, N., Bakkour, A., & Shohamy, D. (2020). What Are Memories For? The Hippocampus Bridges Past Experience with Future Decisions. *Trends in Cognitive Sciences*, 24(7), 542–556.

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological Review*.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: Drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, 8(February), 1–17.

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.

Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, 2019(1), niz004.

Botvinick, M. M. & Cohen, J. D. (2014). The Computational and Neural Basis of Cognitive Control: Charted Territory and New Frontiers. *Cognitive Science*, 38(6), 1249–1285.

Brown, R. & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5(4), 325–337.

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and Neural Bases of Multi-Attribute, Multi-Alternative, Value-based Decisions. *Trends in Cognitive Sciences*, 23(3), 251–263.

Busemeyer, J. R. & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychol Rev*, 100(3), 432–459.

Butko, N. J. & Movellan, J. R. (2008). I-POMDP: An infomax model of eye movement. In *2008 7th IEEE International Conference on Development and Learning* (pp. 139–144).

Callaway, F., Gul, S., Krueger, P., Griffiths, T. L., & Lieder, F. (2018). Learning to select computations. In *Uncertainty in Artificial Intelligence: Proceedings of the Thirty-Fourth Conference*.

Callaway, F., Jain, Y. R., van Opheusden, B., Das, P., Iwama, G., Gul, S., Krueger, P. M., Becker, F., Griffiths, T. L., & Lieder, F. (2022a). Leveraging artificial intelligence to improve people's planning strategies. *Proceedings of the National Academy of Sciences*, 119(12), e2117432119.

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLOS Computational Biology*, 17(3), e1008863.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022b). Rational use of cognitive resources in human planning. *Nat Hum Behav*, 6(8), 1112–1125.

Camerer, C. & Ho, T. (2004). A Cognitive Hierarchy Model of Games*. *Quarterly Journal of Economics*, 119(3), 861–898.

Caplin, A. & Dean, M. (2013). Behavioral Implications of Rational Inattention with Shannon Entropy. *NBER Working Paper*, (August), 1–40.

Cassey, T. C., Evens, D. R., Bogacz, R., Marshall, J. A. R., & Ludwig, C. J. H. (2013). Adaptive sampling of information in perceptual decision-making. *PLoS ONE*, 8(11).

Castel, A. D. (2007). The Adaptive and Strategic Use of Memory By Older Adults: Evaluative Processing and Value-Directed Remembering. In A. S. Benjamin & B. H. Ross (Eds.), *Psychology of Learning and Motivation*, volume 48 of *Skill and Strategy in Memory Use* (pp. 225–270). Academic Press.

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, 143(4), 1476–1488.

Chang, M. B., Gupta, A., Levine, S., & Griffiths, T. L. (2019). Automatically Composing Representation Transformations as a Means for Generalization. In *Proceedings of the International Conference on Learning Representations*: arXiv.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, 9(2), 129–136.

Chase, W. G. (1978). Elementary information processes. In *Handbook of Learning & Cognitive Processes: V. Human Information* (pp. 19–90). Oxford, England: Lawrence Erlbaum.

Chase, W. G. & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1), 55–81.

Chen, H., Chang, H. J., & Howes, A. (2021). Apparently Irrational Choice as Optimal Sequential Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1), 792–800.

Chen, X., Starke, S. D., Baber, C., & Howes, A. (2017). A Cognitive Model of How People Make Decisions Through Interaction with Visual Displays. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17 (pp. 1205–1216). New York, NY, USA: Association for Computing Machinery.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(1), 142–150.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7), 410–418.

Cushman, F. & Morris, A. (2015). Habitual control of goal selection in humans. *PNAS*, 112(45), 13817–13822.

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.

Dayan, P. & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective and Behavioral Neuroscience*, 8(4), 429–453.

De Groot, A. D. (1965). *Thought and Choice in Chess*. The Hague: De Gruyter Mouton.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, 16(1), 105–110.

Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in Multisensory Perception. *Trends in Cognitive Sciences*, 20(10), 736–747.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychol Sci*, 29(5), 761–778.

Dietterich, T. G. (2000). Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition. *Journal of Artificial Intelligence Research*, 13, 227–303.

Ditterich, J. (2006). Stochastic models of decisions about motion direction: Behavior and physiology. *Neural Networks*, 19(8), 981–1012.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The Cost of Accumulating Evidence in Perceptual Decision Making. *Journal of Neuroscience*, 32(11), 3612–3628.

Dunlosky, J. & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In *Metacognition in Educational Theory and Practice*, The Educational Psychology Series (pp. 249–275). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Dunlosky, J. & Lipko, A. R. (2007). Metacomprehension: A Brief History and How to Improve Its Accuracy. *Curr Dir Psychol Sci*, 16(4), 228–232.

Dunlosky, J. & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, 20(4), 374–380.

Eakin, D. K. (2005). Illusions of knowing: Metamemory and memory under conditions of retroactive interference. *Journal of Memory and Language*, 52(4), 526–534.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, 11(5), 1–36.

Eliaz, K. & Schotter, A. (2007). Experimental Testing of Intrinsic Preferences for NonInstrumental Information. *American Economic Review*, 97(2), 166–169.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics*, (pp. 643–669).

Erev, I. & Barron, G. (2005). On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. *Psychological Review*, 112(4), 912–931.

Fisher, G. (2017). An attentional drift diffusion model over binary-attribute choice. *Cognition*, 168, 34–45.

Fleming, S. M. & Daw, N. D. (2017). Self-Evaluation of Decision-Making: A General Bayesian Framework for Metacognitive Computation. *Psychol Rev*, 124(1), 91–114.

Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1), 1–8.

Ford, J. K., Schmitt, N., Schechtman, S. L., Hults, B. M., & Doherty, M. L. (1989). Process tracing methods: Contributions, problems, and neglected research questions. *Organizational behavior and human decision processes*, 43(1), 75–117.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599), 2213–2223.

Frömer, R., Dean Wolf, C. K., & Shenhav, A. (2019). Goal congruency dominates reward value in accounting for behavioral and neural correlates of value-based decision-making. *Nature Communications*, 10(1), 1–11.

Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, 12(1), 1030.

Fudenberg, D., Strack, P., & Strzalecki, T. (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12), 3651–3684.

Gabaix, X. & Laibson, D. (2005). 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. *American Economic Review*, 96 (4)(4), 1043–1068.

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 204, 104394.

Gershman, S. J. (2021). The rational analysis of memory. *Oxford handbook of human memory.*

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

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on psychological science*, 3(1), 20–29.

Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, 62(1), 451–482.

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological review*, 103(4), 650.

Gigerenzer, G. & Todd, P. M. (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press, USA.

Glaholt, M. G. & Reingold, E. M. (2009). Stimulus exposure and gaze bias: A further test of the gaze cascade model. *Attention, Perception & Psychophysics*, 71(3), 445–450.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proc Natl Acad Sci U S A*, 108 Suppl 3, 15647–15654.

Gluth, S., Kern, N., Kortmann, M., & Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, 4(6), 634–645.

Gluth, S., Spektor, M. S., & Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *eLife*, 7, e39659.

Gold, J. I. & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36(2), 299–308.

Goldstein, D. G. & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109(1), 75–90.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1), 21–54.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1.

Gopnik, A. (1998). Explanation as Orgasm*. *Minds and Machines*, 8(1), 101–118.

Gottlieb, J. & Oudeyer, P. Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12), 758–770.

Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11).

Grahek, I., Musslick, S., & Shenhav, A. (2020). A computational perspective on the roles of affect in cognitive control. *International Journal of Psychophysiology*, 151, 25–34.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29, 24–30.

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.

Gruneberg, M. M. & Monks, J. (1974). 'Feeling of knowing' and cued recall. *Acta Psychologica*, 38(4), 257–265.

Gruneberg, M. M., Monks, J., & Sykes, R. N. (1977). Some methodological problems with feeling of knowing studies. *Acta Psychologica*, 41(5), 365–371.

Gul, S., Krueger, P. M., Callaway, F., Griffiths, T. L., & Lieder, F. (2018). Discovering rational heuristics for risky choice. In *The 14th Biannual Conference of the German Society for Cognitive Science, GK*.

Gureckis, T. M. & Markant, D. B. (2012). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspect Psychol Sci*, 7(5), 464–481.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of educational psychology*, 56(4), 208.

Hay, N. (2016). Principles of Metalevel Control.

Hay, N., Russell, S., Tolpin, D., & Shimony, S. E. (2012). Selecting computations: Theory and applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12 (pp. 346–355). Arlington, Virginia, USA: AUAI Press.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: Making working memory work. *Neuroscience*, 139(1), 105–118.

He, R. & Lieder, F. (2022). Where do adaptive planning strategies come from?

Hébert, B. & Woodford, M. (2017). Rational inattention with sequential information sampling. *Working Paper*, (pp. 1–141).

Hebert, B. & Woodford, M. (2019). Rational inattention when decisions take time. *Journal of Chemical Information and Modeling*, 53(9), 1689–1699.

Hemmer, P. & Steyvers, M. (2009). A bayesian account of reconstructive memory. *Topics in Cognitive Science*, 1(1), 189–202.

Ho, M. K., Abel, D., Cohen, J., Littman, M., & Griffiths, T. (2020). The efficiency of human cognition reflects planned information processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 (pp. 1300–1307).

Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The Piecewise Linear Ballistic Accumulator model. *Cogn Psychol*, 85, 1–29.

Hoppe, D. & Rothkopf, C. A. (2019). Multi-step planning of eye movements in visual search. *Scientific Reports*, 9(1), 144.

Horvitz, E. (1987). Reasoning about beliefs and actions under computational resource constraints. In L. N. Kanal, T. S. Levitt, & J. F. Lemmer (Eds.), *UAI '87: Proceedings of the Third Annual Conference on Uncertainty in Artificial Intelligence, Seattle, WA, USA, July 10-12, 1987* (pp. 301–324).: Elsevier.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1), 22–26.

Howes, A., Duggan, G. B., Kalidindi, K., Tseng, Y.-C., & Lewis, R. L. (2016). Predicting Short-Term Remembering as Boundedly Optimal Strategy Choice. *Cognitive Science*, 40(5), 1192–1223.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational Adaptation Under Task and Processing Constraints: Implications for Testing Theories of Cognition and Action. *Psychological Review*, 116(4), 717–751.

Hu, X. (2021). A Bayesian inference model for metamemory. *Psychological Review*, 128(5), 824.

Hu, X., Luo, L., & Fleming, S. M. (2019). A role for metamemory in cognitive offloading. *Cognition*, 193, 104012.

Hunt, L. T., Daw, N. D., Kaanders, P., MacIver, M. A., Mugan, U., Procyk, E., Redish, A. D., Russo, E., Scholl, J., Stachenfeld, K., Wilson, C. R. E., & Kolling, N. (2021). Formalizing planning and information search in naturalistic decision-making. *Nat Neurosci*, 24(8), 1051–1064.

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F., & Behrens, T. E. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience*, 15(3), 470–476.

Hunt, L. T., Rutledge, R. B., Malalasekera, W. M. N., Kennerley, S. W., & Dolan, R. J. (2016). Approach-Induced Biases in Human Information Sampling. *PLoS Biol*, 14(11), e2000638.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of experimental psychology: General*, 129(2), 220.

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.

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences of the United States of America*, 112(10), 3098–103.

Itti, L. & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10), 1295–1306.

Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10), 1489–1506.

Jain, Y. R., Gupta, S., Rakesh, V., Dayan, P., Callaway, F., & Lieder, F. (2019). How do people learn how to plan? In *Proceedings of the Annual Conference on Cognitive Computational Neuroscience*.

Jameson, K. A., Narens, L., Goldfarb, K., & Nelson, T. O. (1990). The influence of near-threshold priming on metamemory and recall. *Acta Psychologica*, 73(1), 55–68.

Jang, A. I., Sharma, R., & Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *eLife*, 10, e63436.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychol Rev*, 119(1), 186–200.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8), 589–604.

Just, M. A. & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2), 99–134.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American psychologist*, 64(6), 515.

Kahneman, D. & Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292.

Kanerva, P. (1988). *Sparse Distributed Memory*. MIT press.

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.

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.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, 84(6), 1329–1342.

Kiani, R. & Shadlen, M. N. (2009). Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science*, 324(5928), 759–764.

Knill, D. C. & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press.

Kocsis, L. & Szepesvári, C. (2006). Bandit Based Monte-Carlo Planning. (pp. 282–293).

Kool, W. & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, 2(12), 899–908.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, 28(9), 1321–1333.

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

Koriat, A. (2000). Control processes in remembering. In *The Oxford Handbook of Memory* (pp. 333–346). New York, NY, US: Oxford University Press.

Krajbich, I. (2018). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, 29, 6–11.

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.

Krajbich, I., Lu, D., Camerer, C., & Rangel, A. (2012). The Attentional Drift-Diffusion Model Extends to Simple Purchasing Decisions. *Front Psychol*, 3.

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.

Krueger, P. M., Lieder, F., & Griffiths, T. L. (2017). Enhancing metacognitive reinforcement learning using reward structures and feedback. In *Proceedings of the Annual Meeting of the Cognitive Science Society*: Cognitive Science Society.

Krusche, M. J. F., Schulz, E., Guez, A., & Speekenbrink, M. (2018). Adaptive planning in human search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behav Brain Sci*, 36(6), 10.1017/S0140525X12003196.

Lachman, J. L., Lachman, R., & Thronesbery, C. (1979). Metamemory through the adult life span. *Developmental Psychology*, 15(5), 543–551.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1), 1–64.

Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in Soar: The anatomy of a general learning mechanism. *Mach Learn*, 1(1), 11–46.

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nat Neurosci*, 18(8), 1159–1167.

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.

Lieder, F. (2018). *Beyond Bounded Rationality: Reverse-engineering and Enhancing Human Intelligence*. PhD thesis, UC Berkeley.

Lieder, F. & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, 124(6), 762–794.

Lieder, F. & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, 14(4), 1–27.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, 122(2), 337–363.

Lohse, G. L. & Johnson, E. J. (1996). A Comparison of Two Process Tracing Methods for Choice Tasks. *Organizational Behavior and Human Decision Processes*, 68(1), 28–43.

Lu, Q., Hasson, U., & Norman, K. A. (2022). A neural network model of when to retrieve and encode episodic memories. *eLife*, 11, e74445.

Ludwig, C. J. H. & Evens, D. R. (2017). Information Foraging for Perceptual Decisions. *J Exp Psychol Hum Percept Perform*, 43(2), 245–264.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), 176–201.

Madan, C. R. (2021). Exploring word memorability: How well do different word properties explain item free-recall probability? *Psychonomic Bulletin & Review*, 28(2), 583–595.

Manohar, S. G. & Husain, M. (2013). Attention as foraging for information and value. *Frontiers in Human Neuroscience*, 7(November), 1–16.

Markant, D. & Gureckis, T. (2014). A preference for the unpredictable over the informative during self-directed learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36(36).

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-Directed Learning Favors Local, Rather Than Global, Uncertainty. *Cognitive Science*, 40(1), 100–120.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: WH Freeman.

Matêjka, F. & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1), 272–298.

Matheson, J. E. (1968). The Economic Value of Analysis and Computation. *IEEE Transactions on Systems Science and Cybernetics*, 4(3), 325–332.

Mattar, M. G. & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nat Neurosci*, 21(11), 1609–1617.

McMillen, T. & Holmes, P. (2006). The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology*, 50(1), 30–57.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychol Rev*, 100(1), 3–22.

Metcalfe, J. (2009). Metacognitive Judgments and Control of Study. *Curr Dir Psychol Sci*, 18(3), 159–163.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(4), 851–861.

Milosavljevic, M., Malmaud, J., & Huth, A. (2010). The Drift Diffusion Model can account for the accuracy and reaction time of value-based choices under high and low time pressure. 5(6), 437–449.

Miner, A. C. & Reder, L. M. (1994). A new look at feeling of knowing: Its metacognitive role in regulating question answering. *Metacognition: Knowing about knowing*, (pp. 47–70).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.

Moreno-Bote, R. (2010). Decision Confidence and Uncertainty in Diffusion Models with Partially Correlated Neuronal Integrators. *Neural Computation*, 22(7), 1786–1811.

Moreno-Bote, R., Ramírez-Ruiz, J., Drugowitsch, J., & Hayden, B. Y. (2020). Heuristics and optimal solutions to the breadth–depth dilemma. *PNAS*, 117(33), 19799–19808.

Murphy, K. P. (2007). *Conjugate Bayesian Analysis of the Gaussian Distribution*. Technical report, University of British Columbia.

Musslick, S. & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, 0(0).

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015). A computational model of control allocation based on the expected value of control. In *The 2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making*.

Najemnik, J. & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031), 387–391.

Narens, L., Jameson, K. A., & Lee, V. A. (1994). Subthreshold priming and memory monitoring. *Metacognition: Knowing about knowing*, (pp. 71–92).

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat Neurosci*, 15(7), 1040–1046.

Nelder, J. A. & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4), 308–313.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3), 402–407.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1), 109–133.

Nelson, T. O. & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. In G. H. Bower (Ed.), *Psychology of Learning and Motivation*, volume 26 (pp. 125–173). Academic Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, 65(3), 151–166.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program. In *IFIP Congress*, volume 256 (pp.64).: Pittsburgh, PA.

Newell, A. & Simon, H. (1956). The logic theory machine–A complex information processing system. *IRE Transactions on Information Theory*, 2(3), 61–79.

Newell, A., Simon, H. A., et al. (1972). *Human Problem Solving*, volume 104. Prentice-hall Englewood Cliffs, NJ.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99 (pp. 278–287).

Nhouyvanisvong, A. & Reder, L. M. (1998). Rapid feeling-of-knowing: A strategy selection mechanism. In *Metacognition: Cognitive and Social Dimensions* (pp. 35–52). Thousand Oaks, CA, US: Sage Publications, Inc.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3), 139–154.

Noguchi, T. & Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological Review*, 125(4), 512–544.

Norris, D. & Cutler, A. (2021). More why, less how: What we need from models of cognition. *Cognition*, 213, 104688.

Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4), 608–631.

O'Ceallaigh, D. & Ruml, W. (2015). Metareasoning in Real-Time Heuristic Search. *Eighth Annual Symposium on Combinatorial Search*, (pp. 87–95).

O'Donoghue, T. & Rabin, M. (1999). Doing It Now or Later. *American Economic Review*, 89(1), 103–124.

Ongchoco, J. D., Jara-Ettinger, J., & Knobe, J. (2019). Imagining the good: An offline tendency to simulate good options even when no decision has to be made. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 904–910).

O'Reilly, R. C. & Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, 18(2), 283–328.

Orquin, J. L. & Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, 144(1), 190–206.

Ortega, P. A. & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153), 20120683.

Oulasvirta, A., Jokinen, J. P. P., & Howes, A. (2022). Computational Rationality as a Theory of Interaction. In *CHI Conference on Human Factors in Computing Systems*, CHI '22 (pp. 1–14). New York, NY, USA: Association for Computing Machinery.

Parr, R. & Russell, S. (1998). Reinforcement learning with hierarchies of machines. *Advances in neural information processing …*, (pp. 1043–1049).

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, 16(2), 366–387.

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.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. The Adaptive Decision Maker. New York, NY, US: Cambridge University Press.

Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, 8(9), 095118.

Piantadosi, S. T. (2021). The Computational Origin of Representation. *Minds and Machines*, 31(1), 1–58.

Pirrone, A., Azab, H., Hayden, B. Y., Stafford, T., & Marshall, J. A. R. (2018). Evidence for the speed–value trade-off: Human and monkey decision making is magnitude sensitive. *Decision*, 5(2), 129–142.

Plate, T. et al. (1995). Holographic reduced representations. *Neural networks, IEEE transactions on*, 6(3), 623–641.

Pleskac, T. J. & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, 117(3), 864–901.

Polanía, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural Oscillations and Synchronization Differentially Support Evidence Accumulation in Perceptual and Value-Based Decision Making. *Neuron*, 82(3), 709–720.

Polanía, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nat Neurosci*, 22(1), 134–142.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A Context Maintenance and Retrieval Model of Organizational Processes in Free Recall. *Psychological Review*, 116(1), 129–156.

Posner, M. I. & McLeod, P. (1982). Information Processing Models-In Search of Elementary Operations. *Annual Review of Psychology*, 33(1), 477–514.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Raaijmakers, J. G. & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88(2), 93–134.

Rahnev, D. & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, 41, e223.

Ramírez-Ruiz, J. & Moreno-Bote, R. (2021). Optimal allocation of finite sampling capacity in accumulator models of multi-alternative decision making. *arXiv:2102.01597 [q-bio]*.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59–108.

Ratcliff, R. & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Comput*, 20(4), 873–922.

Ratcliff, R. & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychol Sci*, 9(5), 347–356.

Ratcliff, R. & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review*, 111(2), 333–367.

Ratcliff, R., Smith, P. L., Brown, S. D., & Mckoon, G. (2016). Diffusion Decision Model : Current Issues and History. *Trends in Cognitive Sciences*, 20(4), 260–281.

Ratcliff, R. & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaching to dealing with contaminant reaction and parameter variability. *Psychonomic Bulletin & Review*, 9(3), 438–481.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology*, 19(1), 90–138.

Reder, L. M. (1988). Strategic Control of Retrieval Strategies. In G. H. Bower (Ed.), *Psychology of Learning and Motivation*, volume 22 (pp. 227–259). Academic Press.

Reder, L. M. & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, memory, and cognition*, 18(3), 435.

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.

Ritz, H., Leng, X., & Shenhav, A. (2021). Cognitive control as a multivariate optimization problem. *arXiv:2110.00668 [q-bio]*.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108(2), 370–392.

Russell, S. & Wefald, E. (1991a). Principles of metareasoning. *Artificial Intelligence*, 49(1-3), 361–395.

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence*, 94(1), 57–77.

Russell, S. J. & Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall.

Russell, S. J. & Subramanian, D. (1995). Provably Bounded-Optimal Agents.

Russell, S. J. & Wefald, E. (1991b). *Do the Right Thing: Studies in Limited Rationality*. MIT press.

Russo, J. E. & Dosher, B. A. (1983). Strategies for multiattribute binary choice. *J Exp Psychol Learn Mem Cogn*, 9(4), 676–696.

Ryan, M. P., Petty, C. R., & Wenzlaff, R. M. (1982). Motivated remembering efforts during tip-of-the-tongue states. *Acta Psychologica*, 51(2), 137–147.

Savage, L. J. (1954). *The Foundations of Statistics*. The Foundations of Statistics. Oxford, England: John Wiley & Sons.

Schacter, D. L. & Worling, J. R. (1985). Attribute information and the feeling-of-knowing. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 39(3), 467–475.

Schulte-Mecklenbeck, M., Kuehberger, A., & Johnson, J. G. (2011). Visiting the decision factory: Observing cognition with MouselabWEB and other information acquisition methods. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods for Decision Research* (pp. 37–58). Psychology Press.

Schwartz, B. L. (2001). The relation of tip-of-the-tongue states and retrieval time. *Memory & Cognition*, 29(1), 117–126.

Schwartz, B. L. & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5), 1074–1083.

Schwartz, B. L. & Metcalfe, J. (2017). Metamemory: An Update of Critical Findings. In *Learning and Memory: A Comprehensive Reference* (pp. 423–432). Elsevier.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological review*, 115(4), 893.

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *eLife*, 9, e60705.

Sezener, C. E., Dezfouli, A., & Keramati, M. (2019). Optimizing the depth and the direction of prospective planning using information values. *PLOS Computational Biology*, 15(3), 1–21.

Shenhav, A., Botvinick, M., & Cohen, J. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2), 217–240.

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. (March), 99–124.

Shi, S. W., Wedel, M., & Pieters, F. (2013). Information acquisition during online decision making: A model-based exploration using eye-tracking data. *Management Science*, 59(5), 1009–1026.

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12), 1317–1322.

Shrager, J. & Siegler, R. S. (1998). SCADS: A Model of Children's Strategy Choices and Strategy Discoveries. *Psychol Sci*, 9(5), 405–410.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1), 99–118.

Simon, H. A. (1979). Information Processing Models of Cognition. *Annual Review of Psychology*, 30(1), 363–396.

Simon, H. A. (1990a). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and Probability*, The New Palgrave (pp. 15–18). London: Palgrave Macmillan UK.

Simon, H. A. (1990b). Invariants of Human Behavior. *Annual Review of Psychology*, 41(1), 1–20.

Simoncelli, E. P. & Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 24(1), 1193–1216.

Sims, C. A. (1998). Stickiness. *Carnegie-Rochester Conference Series on Public Policy*, 49, 317–356.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, 152, 181–198.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, 119(4), 807–830.

Smith, S. M. & Krajbich, I. (2018). Attention and choice across domains. *J Exp Psychol Gen*, 147(12), 1810–1826.

Smith, S. M. & Krajbich, I. (2019). Gaze Amplifies Value in Decision Making. *Psychological Science*, 30(1), 116–128.

Snider, J., Lee, D., Poizner, H., & Gepshtein, S. (2015). Prospective Optimization with Limited Resources. *PLOS Computational Biology*, 11(9), e1004501.

Sobol, I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4), 784–802.

Solway, A. & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, 112(37), 11708–11713.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal Behavioral Hierarchy. *PLOS Computational Biology*, 10(8), e1003779.

Song, M., Wang, X., Zhang, H., & Li, J. (2019). Proactive information sampling in value-based decision-making: Deciding when and where to saccade. *Frontiers in Human Neuroscience*, 13(February), 1–10.

Stahl, D. O. & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization*, 25(3), 309–327.

Stewart, E. E. M., Ludwig, C. J. H., & Schütz, A. C. (2022). Humans represent the precision and utility of information acquired across fixations. *Sci Rep*, 12(1), 2411.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, 53(1), 1–26.

Stocker, A. A. & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci*, 9(4), 578–585.

Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J., & Speekenbrink, M. (2020). Uncertainty in learning, choice, and visual fixation. *Proceedings of the National Academy of Sciences of the United States of America*, 117(6), 3291–3300.

Suchow, J. W. & Griffiths, T. L. (2016). Deciding to Remember : Memory Maintenance as a Markov Decision Process. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2063–2068).

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. *PLOS Computational Biology*, 9(1), e1002803.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning* (pp. 216–224).

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. 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), 181–211.

Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature Neuroscience*, 22(9), 1503–1511.

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nat Commun*, 7(1), 12400.

Tavares, G., Perona, P., & Rangel, A. (2017). The attentional Drift Diffusion Model of simple perceptual decision-making. *Frontiers in Neuroscience*, 11(AUG), 1–16.

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, 24(4), 629–640.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.

Teodorescu, A. R. & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychol Rev*, 120(1), 1–38.

Thomas, A. W., Molter, F., Krajbich, I., Heekeren, H. R., & Mohr, P. N. C. (2019). Gaze bias differences capture individual choice behaviour. *Nature Human Behaviour*, 3(6), 625–635.

Todd, M., Niv, Y., & Cohen, J. D. (2008). Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. In *Advances in Neural Information Processing Systems*, volume 21: Curran Associates, Inc.

Todd, P. M. & Gigerenzer, G. (2003). Bounding rationality to the world. *Journal of economic psychology*, 24(2), 143–165.

Todd, P. M. & Gigerenzer, G. E. (2012). *Ecological Rationality: Intelligence in the World.* Oxford University Press.

Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proceedings of the National Academy of Sciences*, 110(40), E3858–E3867.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychol Rev*, 121(2), 179–205.

Turner, B. M. & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin and Review*, 21(2), 227–250.

Usher, M. & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3), 550–592.

Usher, M. & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychol Rev*, 111(3), 757–769.

van den Berg, R. & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *eLife*, 7, e34963.

van Opheusden, B., Acerbi, L., & Ma, W. J. (2020). Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *PLOS Computational Biology*, 16(12), e1008483.

Van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Vesonder, G. T. & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language*, 24(3), 363–376.

Vickers, D. (1970). Evidence for an Accumulator Model of Psychophysical Discrimination. *Ergonomics*, 13(1), 37–58.

Von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior.* Princeton, NJ, US: Princeton University Press.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? Optimal decisions from very few samples. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, volume 1 (pp. 66–72).

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2), 117–186.

Wald, A. & Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, 19(3), 326–339.

Wang, S., Feng, S. F., & Bornstein, A. M. (2022). Mixing memory and desire: How memory reactivation supports deliberative decision-making. *WIREs Cognitive Science*, 13(2), e1581.

Westbrook, A., van den Bosch, R., Määttä, J. I., Hofmans, L., Papadopetraki, D., Cools, R., & Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, 367(6484), 1362–1366.

Williams, M., Hong, S. W., Kang, M.-S., Carlisle, N. B., & Woodman, G. F. (2013). The benefit of forgetting. *Psychon Bull Rev*, 20(2), 348–355.

Williams, M. D. & Hollan, J. D. (1981). The process of retrieval from very long-term memory. *Cognitive Science*, 5(2), 87–119.

Woodford, M. (2014). Stochastic choice: An optimizing neuroeconomic model. *American Economic Review*, 104(5), 495–500.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3485–3492).

Yang, L. C., 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.

Yellott Jr, J. I. (1977). The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2), 109–144.

Yeung, N. & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philos Trans R Soc Lond B Biol Sci*, 367(1594), 1310–1321.

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, 20(11), 1178–1194.

Yoo, A. H., Klyszejko, Z., Curtis, C. E., & Ma, W. J. (2018). Strategic allocation of working memory resource. *Biorxiv*, (pp. 1–14).

Zhang, Q., Griffiths, T. L., & Norman, K. A. (2022). Optimal policies for free recall. *Psychol Rev*.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 12 point Minion Pro. A template that can be used to format a PhD dissertation with a similar 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. The source code for this dissertation can be found at `https://github.com/fredcallaway/dissertation`.