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

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

Contents

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

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

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 (e.g., 1982) and John Anderson (e.g., 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.

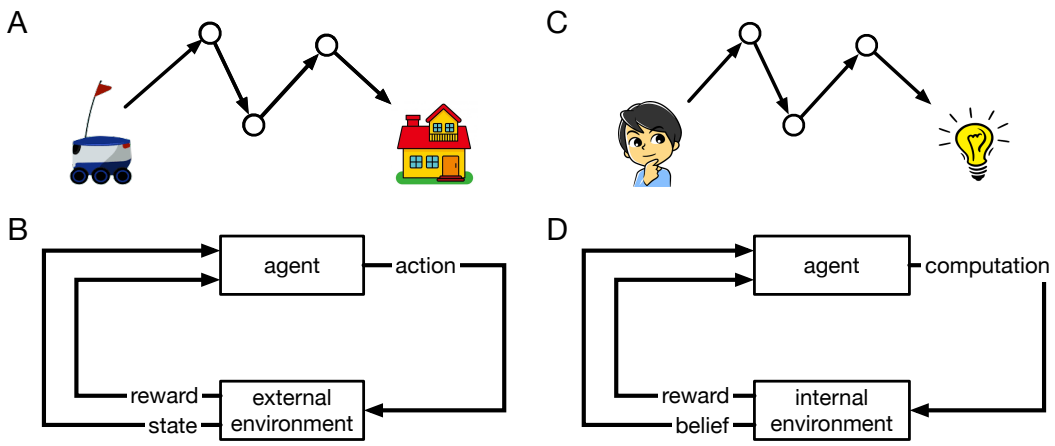

Figure 1.1: Sequential decision problems posed by external and internal environments.

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. Specificially, 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 and 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 and 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 cogntive 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 cogntive 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.

To make things concrete, consider the problem facing a delivery robot, illustrated in Figure 1.1A. Completing this task will require visiting multiple locations in sequence before arriving at the final destination. And at each location, the robot will need to decide where to go next. Thus, it is 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 receving the reward and new state, the agent selects another action and the cycle continues.

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.*

*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

Table 1.1: Classification of cognitive models.

Dynamic	Bounded	Optimized	Examples
✓			Evidence accumulation, Cogntive architectures
	✓		Heuristics and Biases
		✓	Rational analysis, Econ
✓	✓		Ecological rationality
✓		✓	?? Dynamic and optimized ??
	✓	✓	Expected value of control, Information-theoretic bounded rationality
✓	✓	~	Sampling,
✓	✓	✓	This dissertation

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.

1.2 Relation to previous work

The proposed approach builds on a long history of cognitive modeling. To understand how this approach is similar to and different from previous approaches, it is useful to highlight its three key assumptions: cognition is *dynamic* (sequential, occuring over time), *bounded* (subject to costs or constraints), and *optimized* (maximizing reward). As illustrated in Table 1.1, various combinations of these assumptions are made frequently in models of the mind. However, by capturing all three ideas at once, the current approach has advantages that cannot be achieved with any subset.

The fact that cognitive processes are dynamic is uncontroversial, perhaps even self-evident. Thus, this property is shared by most mechanistic models. One widely used class of dynamic models are *evidence accumulation* (or sequential sampling) models, such as the drift diffusion model (Ratcliff, 1978), the leaky competing accumulator (Usher and 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 evi-

it is associated with value. Nevertheless, this association may be deeply engrained, perhaps even genetically so. We return to this question in the conclusion.

dence for one choice is sufficiently greater than the evidence for the other(s). These models can thus explain not only the choices we make, but also how long it takes to make them.

Another important class of dynamic models, *cognitive architectures*, aims to capture a more diverse range of mental activities, beyond simply accumulating more evidence. These models, 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 long-term memory, and performing transformations on symbolic representations of hypothetical world states. However, applying these models in practice poses a substantial challenge. Because any number of cognitive operations could be executed at any time, one must also specify a strategy for how the operations are chosen; in practice this is often done in an ad hoc manner (c.f., Howes et al., 2009).

Optimization, on the other hand, is the defining feature of rational models. Concretely, a cognitive model is optimized if its structure or parameters are set in order to maximize some definition of performance. The prototypical example is expected utility theory (Von Neumann and Morgenstern, 1944; Savage, 1954), which simply assumes that people will take actions that maximize their total expected utility. This model is so abstract that it tells us little about cognition itself. However, applying the optimization principle in more constrained domains has yielded important insights about many areas of cognition, including perception (Marr, 1982; Knill and Richards, 1996; Najemnik and Geisler, 2005) categorization (Anderson, 1991; Ashby and Alfonso-Reese, 1995), memory (Anderson and Milson, 1989), and language (Goldwater et al., 2009).

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 and 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 and Daw, 2018) for closely related efforts, which we discuss further below).

The use of optimization in cognitive models is more controversial.

The basic premise of the approach is that the mind should be well adapted to its environ-

ment, through some combination of learning and evolution (Anderson, 1990). Optimization simply takes this idea to the logical extreme, assuming that the mind is adapted *as well as possible* to the environment.

sampling information theoretic rational analysis ecological rationality EVC cognitive architectures

In comparison to non-rational dynamic models, the proposed approach - naturally captures metacognition and control - c.f. aDDM which assumes random attention - c.f. metamemory which ignores control - avoids combinatorial search

Compared with non-sequential rational models,

In comparison to non-dynamic rational models - captures the process/mechanism (c.f. information-theoretic accounts) - more optimal - captures sequential dependence

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 and Griffiths, 2001), categorization (Anderson, 1991; Ashby and Alfonso-Reese, 1995), perception (Knill and Richards, 1996), and information-seeking (Oaksford and Chater, 1994; Gureckis and 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 and 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 and Daw, 2018) for closely related efforts, which we discuss further below).

1.3 Optimal cognitive processes as solutions to Markov decision processes

The proposed approach rests on a key intuition: the thoughts one has at any moment depend on the thoughts one had before. That is, our mental processes are sequentially depen-

dent. Furthermore, thoughts are only useful insofar as they influence our behavior, and this behavior often occurs well after the thought itself. That is, the benefits of thought are temporally delayed. These two properties, sequential dependence and delayed reward make sequential decision problems very challenging to solve. Fortunately, a long history of work in artificial intelligence—from Newell and Simon's pioneering proof-writing programs (Newell and Simon, 1956) to super-human Chess and Go engines (Silver et al., 2017)—has focused on solving just this sort of problem.

In artificial intelligence research, sequential decision problems are often formalized with the framework of **Markov decision processes (MDP)** (Puterman, 2014; Sutton and Barto, 2018). An MDP describes a dynamic interaction between an agent and an environment. It is defined by a set of possible states the environment can be in, a set of actions the agent can execute, a reward function that specifies the immediate utility associated with executing each action in each state, and a transition function that specifies how actions change the state. The agent's goal is to maximize the total cumulative reward received. This can be accomplished by choosing actions according to the optimal policy, which specifies the best action to execute in each state. See Box 1 for a technical definition of MDPs.

The fact that cognition—or more generally, computation—poses a sequential decision problem was recognized by researchers in the field of **rational metareasoning**, which aims to build AI systems that can adaptively allocate their limited computational resources. In particular, Hay and colleagues Hay et al., 2012; Hay, 2016 formalize this problem of "selecting computations" as a **metalevel MDP**. In a metalevel MDP, the states correspond to beliefs and the actions correspond to computations that refine those beliefs (according to the transition function). The reward function encodes both the costs and benefits of computation; it assigns a strictly negative reward for each computation executed, but a potentially positive reward for the utility of the external action that is ultimately chosen (based on the belief produced by computation).

Applying the metalevel MDP formalism to cognitive science provides a suite of theoretical and computational tools, both to formalize the problems that our brains must solve, and to identify near-optimal solutions to those problems. In a psychological context, the states of a metalevel MDP can be interpreted as mental states, and the actions as cognitive operations. Along with the transition function, these specify the environment "within the skin" that a cognitive process must interact with. By further specifying a reward function, we can quantify the tradeoff between cognitive cost and extrinsic utility . This in turn allows us to identify the optimal cognitive process—that is, the one achieves the best possible tradeoff between cost and utility—as the optimal policy for the metalevel MDP.

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 underlying the proposed framework is that cognitive processes (including, for example, decision making) are themselves solutions to sequential decision problems. Drawing on a subfield of artificial intelligence known as *rational metareasoning* (Matheson, 1968; Russell and Wefald, 1991), we formalize this insight using the framework of *meta-level Markov decision processes* (meta-level 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 beliefs about the world. At each moment, the agent must choose whether to continue deliberating, refining their beliefs but accruing computational cost, or to instead stop computing and make a decision. 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 optimal action given their current belief and receive a reward associated with the external utility of that action.

In this chapter, I provide a formal description of the framework, identify technical challenges that arise when applying the framework, and discuss various strategies for addressing those strategies.

2.1 Markov decision processes

The core mathematical object underlying our approach is the Markov decision process (MDP), illustrated in Figure 2.1A. MDPs are the standard formalism for modeling the sequential interaction between an agent and a stochastic environment. An MDP is defined by

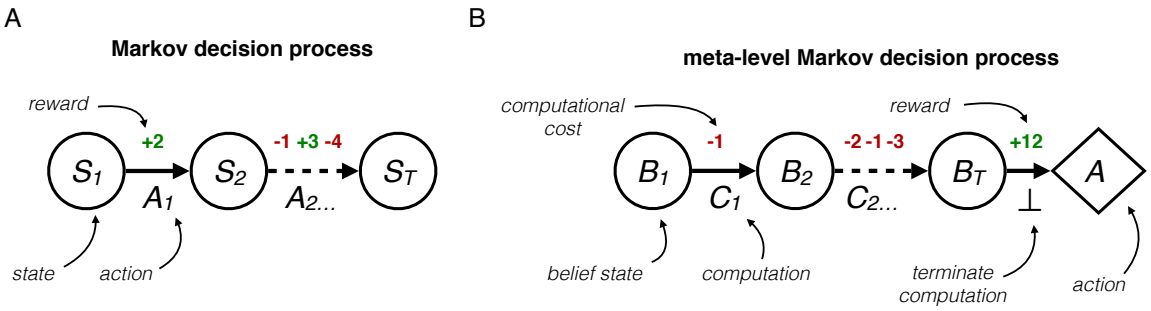

Figure 2.1: Formal framework: meta-level Markov decision processes. (A) A Markov decision process formalizes the problem of acting adaptively in a dynamic environment. The agent executes actions that change the state of the world and generate rewards, which the agent seeks to maximize. (B) A meta-level Markov decision process formalizes the problem of *deciding how to act* when computational resources are limited. The agent executes computations that update their belief state and incur computational cost. When the agent executes the termination operation $\perp$, they take an external action based on their current belief state.

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 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 specifies the reward or utility for executing a given action in a given state.

The standard goal in an MDP is to maximize the expected cumulative reward attained, that is, the *return*. Importantly, this may require incurring immediate losses (negative rewards) in order to get to a state from which a highly rewarding action can be executed. It is typically assumed that the agent selects their actions based on the current state; the mapping from state to action is called a policy, denoted π. Solving an MDP amounts to finding a policy that maximizes the expected return, that is, a mapping from states to actions that, when followed, maximizes the total reward one will receive on average.

In addition to their foundational role in artificial intelligence (Sutton and Barto, 2018), MDPs are widely used in models of human decision-making (Dayan and Daw, 2008). MDPs are the formal foundation for models of reinforcement learning (Niv, 2009) and model-based planning (Huys et al., 2015; Botvinick and Toussaint, 2012), as well as competition between the two systems (Daw et al., 2005; Keramati et al., 2011; Kool et al., 2017). They have also been used to study information-seeking (Gottlieb et al., 2013; Hunt et al., 2016), generalization (Tomov et al., 2021), and hierarchical abstraction (Solway et al., 2014; Botvinick et al., 2009). However, with a few notable exceptions (Dayan and Huys, 2008; Drugowitsch et al., 2012; Tajima et al., 2016), MDPs have primarily been used to model the sequential decision problems posed by the external world. In the following section, we show how this

framework can be applied to model the sequential decision problem posed by one's own cognitive architecture.

2.2 Meta-level Markov decision processes

Meta-level Markov decision processes (meta-level 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, there is a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, and a reward function r_{object} (we omit the transition function because we focus on one-shot decisions). These define the *object-level* problem: the external problem the agent must solve in the world. Additionally, the meta-level MDP defines a set of beliefs $\mathcal{B}$, a set of computations $\mathcal{C}$, and meta-level transition and reward functions, T_{meta} and r_{meta}. These define the *meta-level* problem: how to allocate limited computational resources in the service of solving the object-level problem.

As illustrated in Figure 2.1B, the meta-level problem is itself a sequential decision problem, analogous to one defined by a standard MDP. However, in the meta-level problem, the states are replaced by beliefs (mental states) and the actions are replaced by computations (cognitive operations). The meta-level transition function describes how computations update beliefs, and the meta-level reward function captures both computational cost and the object-level reward of the action that is ultimately executed. We provide a formal definition below.

We define a meta-level MDP as $(\mathcal{S}, \mathcal{A}, r_{\text{object}}, \mathcal{B}, \mathcal{C}, T_{\text{meta}}, r_{\text{meta}})$. The first three components define the object-level problem. They have the same interpretation as $\mathcal{S}$, $\mathcal{A}$, and r in a standard MDP. Note that we omit the transition function because we limit our attention to problems in which all computation must be performed before any action (sequential object-level problems can be accomodated by letting each element of $\mathcal{A}$ be a sequence of actions; see Chapter 5). This

The latter four components define the meta-level problem.

We now define these four components in turn.

Beliefs A belief state $b \in \mathcal{B}$ captures the agent's current knowledge about the relevant state of the world. Formally, a belief is a distribution states, $\mathcal{B} \subseteq \Delta(\mathcal{S})$. Note that $\Delta(\mathcal{S})$ denotes the set of all possible distributions over $\mathcal{S}$. Importantly, contrary to a standard rational treatment of beliefs, the belief states in a meta-level MDP do not include all the information that is available to the DM. Instead, the belief state only contains information that

is immediately accessible, excluding, for example, long-term memories and the number of calories in every box of cereal on a shelf.

Computations A computational operation $c \in \mathcal{C}$ is a primitive operation afforded by the computational architecture. Formally, it is a meta-level action that updates the belief in much the same way as a regular action changes state. All meta-level MDPs include the termination operation $\perp$, which denotes that computation should be terminated and an action should be selected based on the current belief state. We further explain belief updating and termination in the following two paragraphs.

Transition function The meta-level transition function $T_{\text{meta}} : \mathcal{B} \times \mathcal{C} \times \mathcal{S} \to \Delta(\mathcal{B})$ describes how computation updates beliefs. At each time step, the next belief is sampled from a distribution that depends on the current belief, the computational operation that was just executed, and the true state of the world, that is,

$$b_{t+1} \sim T_{\text{meta}}(b_t, c_t, s). \tag{2.1}$$

The transition function thus defines the core structure of the computational architecture. Following previous work (Matheson, 1968; Hay et al., 2012), we assume that the effect of computation is to generate or reveal information about the true state of the world, which is then integrated into the belief state. Thus, in expectation, computation has the effect of making one's beliefs more precise and accurate, although an individual computation may yield misleading information.

Reward function The meta-level reward function $r_{\text{meta}} : \mathcal{B} \times \mathcal{C} \times \mathcal{S} \to \mathbb{R}$ describes both the costs and benefits of computation. For the former, r_{meta} assigns a strictly negative reward for all non-terminating computational operations,

$$r_{\text{meta}}(b, c, s) = -\text{cost}(c) \text{ for } c \neq \perp. \tag{2.2}$$

The cost of computation may include multiple factors, such as energetic costs and opportunity costs.

Intuitively, the benefit of computation is that it allows one to make better decisions. This is captured by the meta-level reward for the termination operation $\perp$, defined as the true utility of the external action that the DM would execute given the current belief. We assume

that the action is selected optimally. Thus,

$$r_{\text{meta}}(b, \bot, s) = r_{\text{object}}(s, a^*(b)). \tag{2.3}$$

where

$$a^*(b) = \operatorname*{argmax}_a \mathrm{E}\left[r_{\text{object}}(s, a) \;\middle|\; s \sim b\right] \tag{2.4}$$

In English, the meta-level reward for termination is the *true* utility of the action* with maximal *estimated* utility.

Policy The solution to a meta-level MDP takes the form of a policy $\pi : \mathcal{B} \to \Delta(\mathcal{C})$ that (perhaps stochastically) selects which computation to perform in each possible belief state. The optimal policy is the one that maximizes expected meta-level return,

$$\pi^* = \operatorname*{argmax}_\pi \mathrm{E}\left[\sum_{t=1}^{T} r_{\text{meta}}(B_t, C_t, S) \;\middle|\; C_t \sim \pi\right]. \tag{2.5}$$

2.2.1 Value of computation

This suggests a strategy for selecting computations optimally. For each item, estimate how much one's decision would improve if one sampled from it (and then continued sampling optimally). Subtract from this number the cost of taking the sample (and also the estimated cost of the future samples). Now identify the item for which this value is maximal. If it is positive, it is optimal to take another sample for this item; otherwise, it is optimal to stop sampling and make a decision.

This basic logic is formalized in rational metareasoning as the *value of computation* (VOC; Russell and Wefald, 1991). Formally, $\mathrm{VOC}(b, c)$ is defined as the expected increase in total metalevel reward if one executes a single computation, c, and continues optimally rather than making a choice immediately (i.e., executing $\bot$):

$$\mathrm{VOC}(b_t, c) = R(b_t, c) + \mathrm{E}\left[\sum_{t'=t+1}^{T} R(b_{t'}, c_{t'}) \;\middle|\; c_{t'} \sim \pi^*(b_{t'})\right] - R(b, \bot).$$

*For notational clarity, we have assumed a single optimal action. When multiple actions have the same expected value, we assume that ties are broken randomly; thus, $a^*(b)$ is more precisely a uniform distribution over all optimal actions, and $r_{\text{meta}}(b, \bot, s)$ takes an expectation over them.

We can then define the optimal policy as selecting computations with maximal VOC:

$$\pi^*(b) \sim \text{Uniform}(\text{argmax}\, c\text{VOC}(b, c)).$$

For those familiar with reinforcement learning, this recursive joint definition of π^* and VOC is exactly analogous to the joint definition of the optimal policy with the state-action value function, Q (Sutton and Barto, 2018). Indeed, $\text{VOC}(b, c) = Q(b, c) - R(b, \bot)$.

2.2.2 Optimal metalevel policy

TODO: shorten, maybe move some material to metamdp section

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 belief state. Formally, $c_t \sim \pi(b_t)$. The optimal metalevel policy, π^*, is the one that maximizes expected total metalevel reward,

$$\pi^* = \text{argmax}\, \pi\, \text{E}\left[\sum_t^T R(b_t, c_t) \;\middle|\; c_t \sim \pi(b_t)\right].$$

Replacing R with its definition, we see that this requires striking a balance between the expected value of the chosen item and the computational cost of the samples that informed the choice,

$$\pi^* = \text{argmax}\, \pi\, \text{E}\left[\max_i \mu_T^{(i)} - \sum_t^{T-1} \text{cost}(b_t, c_t) \;\middle|\; c_t \sim \pi(b_t)\right].$$

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. This suggests a strategy for selecting computations optimally. For each item, estimate how much one's decision would improve if one sampled from it (and then continued sampling optimally). Subtract from this number the cost of taking the sample (and also the estimated cost of the future samples). Now identify the item for which this value is maximal. If it is positive, it is optimal to take another sample for this item; otherwise, it is optimal to stop sampling and make a decision.

This basic logic is formalized in rational metareasoning as the *value of computation* (VOC; Russell and Wefald, 1991). Formally, $\text{VOC}(b, c)$ is defined as the expected increase in total metalevel reward if one executes a single computation, c, and continues optimally rather

than making a choice immediately (i.e., executing $\perp$):

$$\mathrm{VOC}(b_t, c) = R(b_t, c) + \mathrm{E}\left[\sum_{t'=t+1}^{T} R(b_{t'}, c_{t'}) \;\middle|\; c_{t'} \sim \pi^*(b_{t'})\right] - R(b, \perp).$$

In our model, this can be rewritten

$$\mathrm{VOC}(b_t, c) = -\mathrm{cost}(b_t, c) + \mathrm{E}\left[\max_i \mu_T^{(i)} - \sum_{t'=t+1}^{T-1} \mathrm{cost}(b_{t'}, c_{t'}) \;\middle|\; c_{t'} \sim \pi^*(b_{t'})\right] - \max_i \mu_t^{(i)}.$$

That is, the VOC for sampling a given item in some belief state is the expected improvement in the value of the chosen item (rather than making a choice based on the current belief) minus the cost of sampling that item and the expected cost of all future samples.

We can then define the optimal policy as selecting computations with maximal VOC:

$$\pi^*(b) \sim \mathrm{Uniform}(\operatorname{argmax}_c \mathrm{VOC}(b, c)).$$

For those familiar with reinforcement learning, this recursive joint definition of π^* and VOC is exactly analogous to the joint definition of the optimal policy with the state-action value function, Q (Sutton and Barto, 2018). Indeed, $\mathrm{VOC}(b, c) = Q(b, c) - R(b, \perp)$.

Finally, by definition, $\mathrm{VOC}(b, \perp) = 0$ for all b. Thus, the optimal policy terminates sampling when no computation has a positive VOC.

2.2.3 Bayesian metalevel policy search

This method is based on an approximation of the VOC as a linear combination of features,

$$\widehat{\mathrm{VOC}}(b, c; \mathbf{w}) = w_1 \mathrm{VOI}_{\mathrm{myopic}}(b, c) + w_2 \mathrm{VOI}_{\mathrm{item}}(b, c) + w_3 \mathrm{VOI}_{\mathrm{full}}(b) - (\mathrm{cost}(c) + w_4), \tag{2.6}$$

for all $c \neq \perp$, with $\widehat{\mathrm{VOC}}(b, \perp; \mathbf{w}) = \mathrm{VOC}(b, \perp) = 0$.

We briefly define the features here, and provide full derivations in Appendix A.5. The VOI terms quantify the *value of information* (Howard, 1966) that might be gained by different additional computations. Note that the VOI is different from the VOC because the latter includes the costs of computation as well as its benefits. In general, the VOI is defined as the expected improvement in the utility of the action selected based on additional information rather than the current belief state: $E_{\tilde{b}|b}[R(\tilde{b}, \perp) - R(b, \perp)]$, where $\tilde{b}$ is a hypothetical future belief in which the information has been gained, the distribution of which depends on the

current belief.

$\text{VOI}_{\text{myopic}}(b, c)$ 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, b. $\text{VOI}_{\text{item}}(b, c)$ denotes the expected improvement from learning the true value of item c, and then choosing the best item based on that information. Finally, $\text{VOI}_{\text{full}}(b)$ denotes the improvement from learning the true value of every item and then making an optimal choice based on that complete information.

Together, these three features approximate the expected value of information that could be gained by the (unknown) sequence of future samples. Importantly, this true value of information always lies between the lower bound of $\text{VOI}_{\text{myopic}}$ and the upper bound of VOI_{full} (see Figure A.5), implying that the true VOI is a convex combination of these two terms. Note, however, that the weights on this combination are not constant across beliefs, as assumed in our approximation. Thus, including the VOI_{item} term, improves the accuracy of the approximation, by providing an intermediate value between the two extremes. Finally, the last two terms in Equation 2.6 approximate the cost of computation: $\text{cost}(c)$ is the cost of carrying out computation c and w_4 approximates the expected future costs incurred under the optimal policy.

Although maximizing $\widehat{\text{VOC}}(b, c; \mathbf{w})$ identifies the policy with the best performance, it is unlikely that humans make attentional decisions using such perfect and noiseless maximization. Thus, we assume that computations are chosen using a Boltzmann (softmax) distribution (McFadden, 2001) given by

$$\pi(c \mid b; \mathbf{w}, \beta) \propto \exp\left\{\beta \widehat{\text{VOC}}(b, c; \mathbf{w})\right\},$$

where the inverse temperature, β, is a free parameter that controls the degree of noise. Note that computation selection is fully random when $\beta = 0$ and becomes deterministic as $\beta \to \infty$.

To identify the weights used in the approximation, we first assume that $w_i \geq 0$ and $w_1 + w_2 + w_3 = 1$, since $w_{1:3}$ features form a convex combination and w_4 captures the non-negative future cost.

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 [person] is responsible for [their] choice and must accept the consequences, whatever they may be.

W. H. Auden

3

Attention

Fixation patterns in simple choice reflect optimal information sampling

Consider the problems 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 and McKoon, 2008; Ratcliff et al., 2016; Milosavljevic et al., 2010) and the leaky competing accumulator model (Usher and 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 and Reingold, 2009; Krajbich et al., 2010; Krajbich and Rangel, 2011; Cavanagh et al., 2014; Tavares et al., 2017; Smith and Krajbich, 2019), unless they are aversive, in which case they are chosen *less* frequently (Armel and 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 and Rangel, 2011; Smith and Krajbich, 2019;

Tavares et al., 2017). See Orquin and 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 and 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 and 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 and 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 and Geisler, 2005; Eckstein, 2011); in perceptual discrimination problems, people adapt their relative fixation time to the targets' noise levels (Cassey et al., 2013; Ludwig and 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 and 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 and 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 paper, 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 and Morgenstern, 1944) to additionally account for constraints imposed by limited cognitive resources (Lewis et al., 2014; Griffiths et al., 2015; Lieder and Griffiths, 2020; Gershman et al., 2015; Sims, 1998; Caplin and 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 and Wefald, 1991; 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 and 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 and Rangel, 2008; Armel et al., 2008; Krajbich et al., 2010; Krajbich and 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

3.1.1 Sequential sampling model

We consider simple choice problems in which a decision maker (DM) is presented with a set of items (e.g., snacks) and must choose one. Each item i is associated with some true but unknown value, $u^{(i)}$, the utility that the DM would gain by choosing it. Following previous work (Ratcliff and McKoon, 2008; Krajbich et al., 2010; Tajima et al., 2016, 2019; Fudenberg et al., 2018; Bogacz et al., 2006; Drugowitsch et al., 2012; Bitzer et al., 2014; Gold and Shadlen, 2002), we assume that the DM 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 DM 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 DM can only sample from one item at each time point, the item she is fixating on. This sets up a fundamental problem: How should she allocate fixations in order to make good decisions without incurring too much cost? Specifically, at each time point, the DM 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

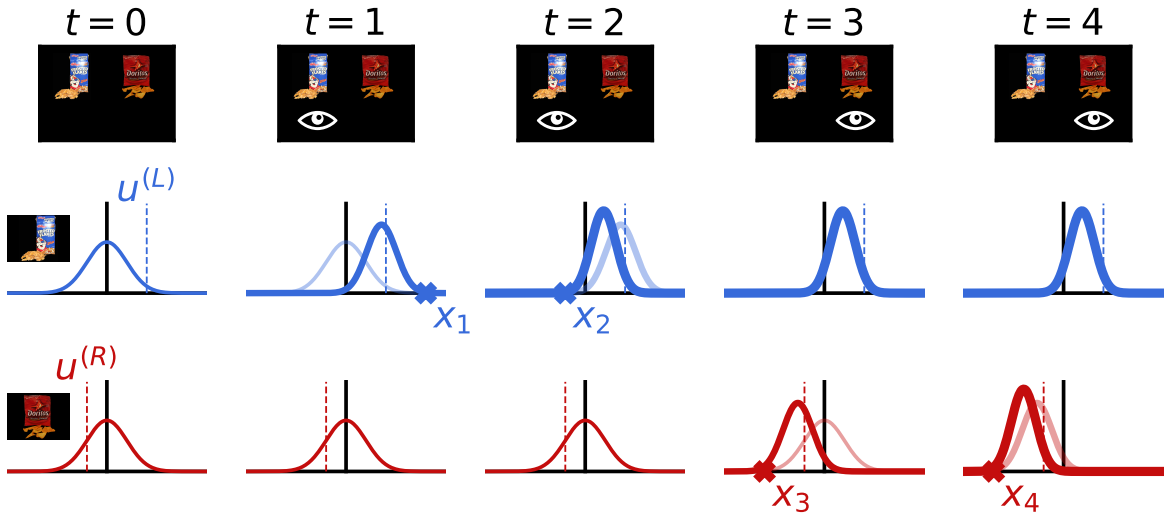

Figure 3.1: ***Sampling and belief updating in the binary choice task.*** The top row shows the experimental display, with the fixated item denoted by the eye symbol. The bottom two rows depict the first few steps of the sampling and belief updating process. The decision maker's beliefs about the value 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 belief 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.

which item to attend to based on her current value estimates and their uncertainty.

The DM's belief about the item values at time t is described by a set of Gaussians, one for each item, with means $\mu_t^{(i)}$ and precisions $\lambda_t^{(i)}$ (the precision is the inverse of the variance). These estimated value distributions are initialized to the DM's prior belief about the distribution of values in the environment. That is, she assumes that $u^{(i)} \sim \text{Normal}(\bar{\mu}, \bar{\sigma}^2)$ and consequently sets $\mu_0^{(i)} = \bar{\mu}$ and $\lambda_0^{(i)} = \bar{\sigma}^{-2}$ for all i. We further discuss the important role of the prior below.

We model the control of attention as the selection of cognitive operations, c_t, that specify either an item to sample, or the termination of sampling. If the DM wishes to sample from item c at time-step t, she selects $c_t = c$ and receives a signal

$$x_t \sim \text{Normal}(u^{(c)}, \sigma_x^2), \tag{3.1}$$

where $u^{(c)}$ is the *unknown* true value of the item being sampled, and σ_x^2 is a free parameter specifying the amount of noise in each signal. The posterior distribution over the sampled item's value is then updated in accordance with Bayesian inference (see Equation 3.4 below).

The cognitive cost of each step of sampling and updating is given by a free parameter, γ_{sample}. We additionally impose a switching cost, γ_{switch}, that the DM incurs whenever she samples from an item other than the one sampled on the last timestep (i.e., makes a saccade to a different item). Thus, the cost of sampling is

$$\text{cost}(c_t) = \gamma_{\text{sample}} + \mathbf{1}(c_t \neq c_{t-1})\, \gamma_{\text{switch}}. \tag{3.2}$$

Note that the model includes the special case in which there are no switching costs ($\gamma_{\text{switch}} = 0$).

In addition to choosing an item to sample, the DM can also decide to stop sampling and choose the item with the highest expected value. In this case, she selects $c_t = \perp$. It follows that if the choice is made at time step T (i.e., $c_T = \perp$) the chosen item is $i^* = \text{argmax}_i \mu_T^{(i)}$. The DM's total payoff on a single decision is given by:

$$\text{payoff} = \underbrace{u^{(i^*)}}_{\substack{\text{utility of} \\ \text{chosen item}}} - \underbrace{\sum_{t=1}^{T-1} \text{cost}(c_t)}_{\text{cognitive cost}}. \tag{3.3}$$

3.1.2 Metalevel Markov decision process

To characterize optimal attention allocation, we cast the model as a metalevel MDP in which the belief 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 four components of the metalevel MDP below.

Beliefs A belief state, $b \in \mathcal{B}$, corresponds to a set of posterior distributions over each item's value. Because the distributions are Gaussian, the belief can be represented by two vectors, $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, that specify the mean and precision of each distribution. That is

$$p(u^{(i)} \mid b) = \text{Normal}(u^{(i)}; \mu^{(i)}, 1/\lambda^{(i)})$$

To model the switching cost, the belief state must also encode the currently fixated item (taking a null value, $\oslash$, in the initial belief). Thus, a belief is a tuple $b_t = (\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, f_t)$. The dimensionality of the belief space is $2N + 1$ where N is the number of items.

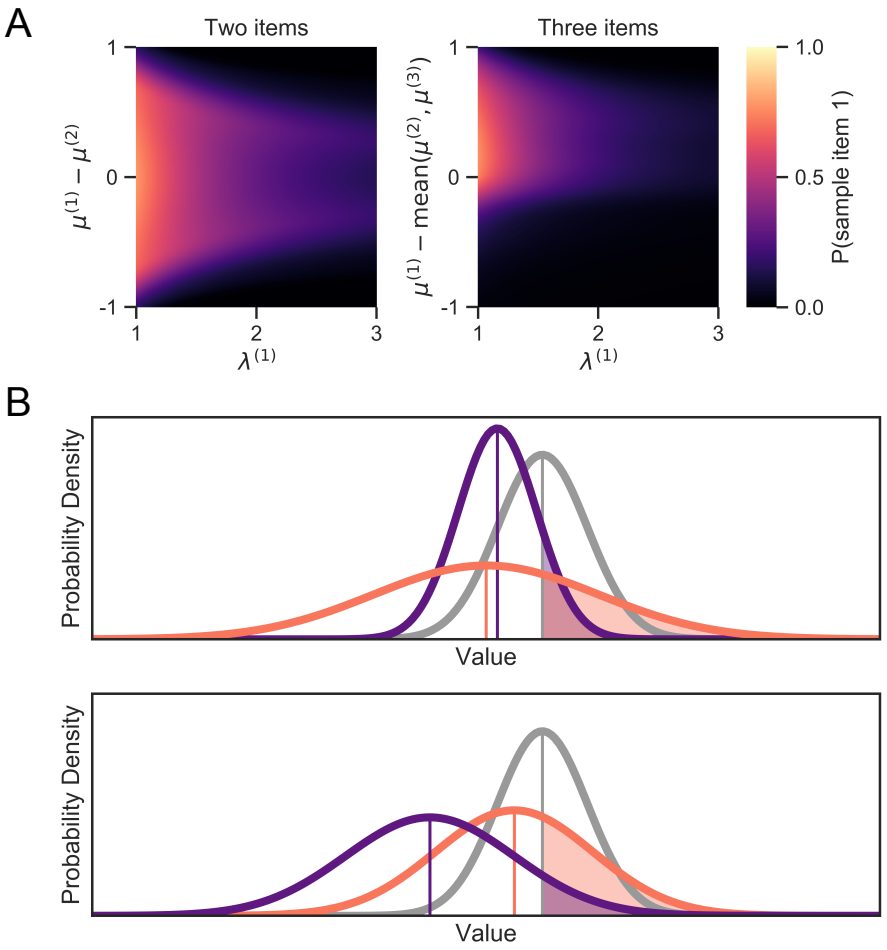

Figure 3.2: ***Optimal fixation policy.*** (A) Probability of fixating on item 1 as a function of the precision of its value estimate, $\lambda^{(1)}$, and the mean of its relative value estimate, $\mu^{(1)} - \text{mean}(\mu^{(2)}, \mu^{(3)})$. The heat map denotes the probability of fixating item 1 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 belief state for trinary choice. The curves depict the estimated beliefs for 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.

COMPUTATIONS A computation, $c \in \mathcal{C}$, corresponds to sampling an item's 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, $\perp$, that terminates the computation process (in our case, sampling) and selects an optimal external action given the current belief state (in our case, choosing the item with maximal posterior

mean).

TRANSITION FUNCTION The metalevel transition function describes how computations update beliefs. In our model, this corresponds to the sampling and Bayesian belief updating procedure described above. Given the current belief, $b_t = (\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, f_t)$, and computation, c, the next belief state, $b_{t+1} = (\boldsymbol{\mu}_{t+1}, \boldsymbol{\lambda}_{t+1}, f_{t+1})$, is sampled from the following generative process:

$$\begin{aligned} x_t &\sim \text{Normal}(u^{(c)}, \sigma_x^2) \\ \lambda_{t+1}^{(c)} &= \lambda_t^{(c)} + \sigma_x^{-2} \\ \mu_{t+1}^{(c)} &= \frac{\sigma_x^{-2} x_t + \lambda_t^{(c)} \mu_t^{(c)}}{\lambda_{t+1}^{(c)}} \\ \lambda_{t+1}^{(i)} &= \lambda_t^{(i)} \text{ and } \mu_{t+1}^{(i)} = \mu_t^{(i)} \text{ for } i \neq c. \\ f_{t+1} &= c \end{aligned} \tag{3.4}$$

REWARD FUNCTION Finally, the metalevel reward function incorporates both the cost of computation and the utility of the chosen action. The metalevel reward for sampling is defined

$$R(b_t, c_t) = -\text{cost}(b_t, c_t) = -(\gamma_{\text{sample}} + \mathbf{1}(f_t \neq \oslash \wedge c_t \neq f_t)\, \gamma_{\text{switch}}).$$

That is, the cost of sampling includes a fixed cost, γ_{sample}, as well as an additional switching cost, γ_{switch}, that is paid when sampling from a different item than that sampled on the last time step. We assume that this cost is not paid for the first fixation; however, this assumption has no effect on the optimal policy for reasonable parameter values.

The action utility is the true value of the chosen item, i.e., $u^{(i_t^*)}$ where $i_t^* = \text{argmax}_i\, \mu_t^{(i)}$. The metalevel reward for the termination computation, $\perp$, is the expectation of this value. Because we assume accurate priors and Bayesian belief updating, this expectation can be taken with respect to the agent's own beliefs (Hay et al., 2012), resulting in

$$R(b_t, \perp) = \text{E}\left[u^{(i_t^*)} \mid b_t\right] = \max_i \mu_t^{(i)}.$$

3.1.3 OPTIMAL POLICY

We assume that the decisions about where to fixate and when to stop sampling are made optimally, subject to the informational constraints described in the previous section. Formally, we assume that the c_t are selected by an *optimal policy*. A policy selects the next cognitive operation to execute, c_t, given the current belief state, (μ_t, λ_t); it is optimal if it selects c_t in

a way that maximizes the expectation of Equation 3.3. How can we identify such a policy? Problems of this kind have been explored in the artificial intelligence literature on rational metareasoning (Matheson, 1968; Russell and Wefald, 1991). Thus, we cast the model described above as a metalevel Markov decision process (Hay et al., 2012), and identify a near-optimal policy using a recently developed method that has been shown to achieve strong performance on a related problem (Callaway et al., 2018). In accordance with past work modeling people's choices (McFadden, 2001) and fixations (Gluth et al., 2020; Song et al., 2019), we assume that people follow a softmax policy in selecting each cognitive operation by sampling from a Boltzmann distribution based on their estimated values. Thus, their choices of cognitive operations are guided by the optimal policy, but subject to some noise. See Section 3.4.1 for details.

What does optimal attention allocation look like? In order to provide an intuitive understanding, we focus on two key properties of belief 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 *ceteris paribus* 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 for value-directed attention in a simplified non-dynamic case.

3.1.4 The prior distribution

Recall that the initial belief about each item's value is set to the DM'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 DM 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 DM 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 DM 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:

$$\begin{aligned} \bar{\mu} &= \alpha \cdot \text{mean}(\text{ratings}) \\ \bar{\sigma} &= \text{std}(\text{ratings}). \end{aligned} \tag{3.5}$$

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 DM 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 the DM encounters in her daily life.

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 and 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 sam-

ple takes 100ms* 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 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.2. We also describe a validation of our model fitting approach in Appendix A.3.

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.

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

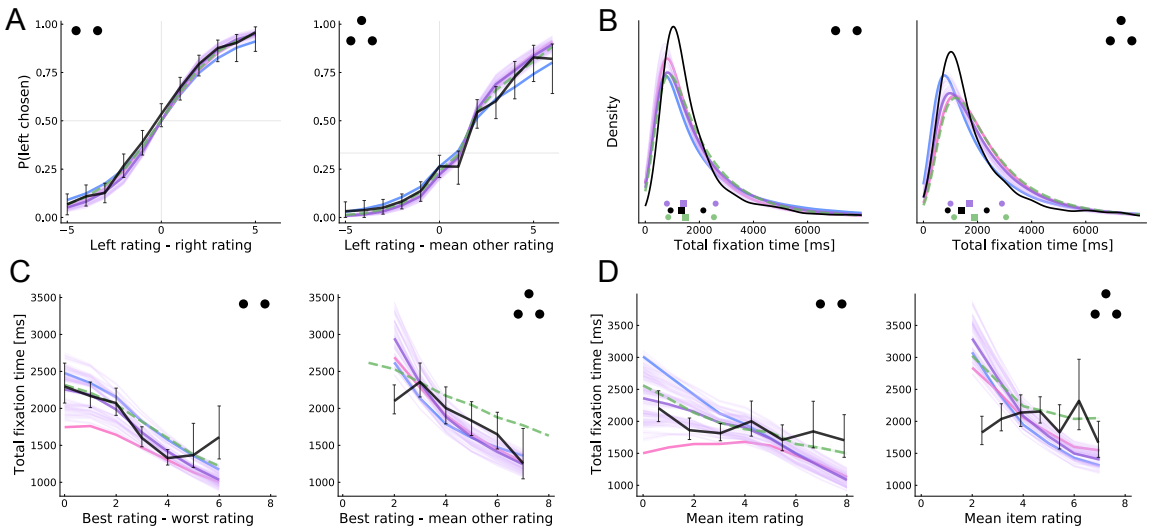

Figure 3.3: ***Basic psychometrics.*** Each panel compares human data (black) and model predictions for binary choice (left, two dots) and trinary choice (right, three dots). The main model predictions are shown in purple. The restricted model predictions for the case of a highly biased prior mean ($\alpha = 0$) 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).

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. However, 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 and 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 weight-

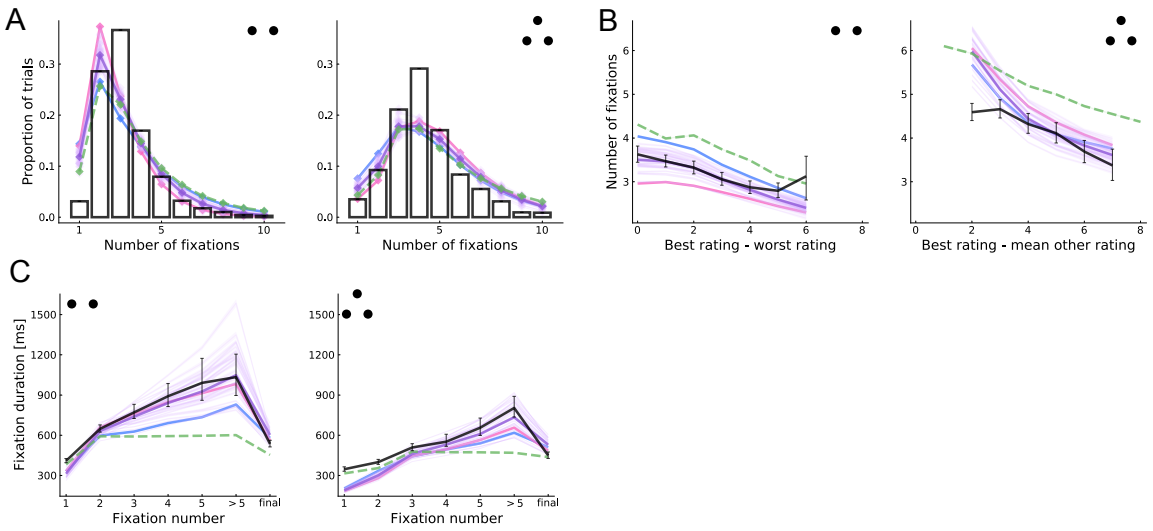

Figure 3.4: ***Basic fixation patterns.*** (A) Histogram of number of fixations in a trial. (B) Number of fixations as a function of decision difficulty, as measured by the relative rating of the best item. (C) Duration of fixation by fixation number. Final fixations are excluded from all but the last bin. See Figure 3.3 for more details.

ing (specifically, a hyperbolic function, c.f. Armel and Rangel, 2008), and thus our model also predicts this pattern. Surprisingly, we do not see strong evidence for the overall value effect in the datasets we consider, but we note that the effect has been found robustly in several other datasets (Smith and 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

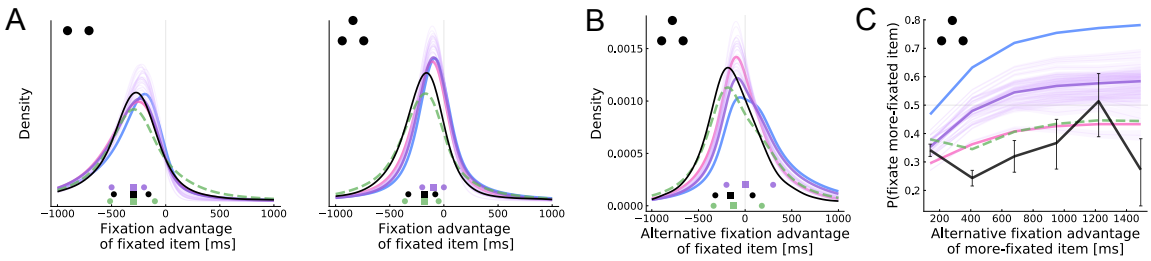

Figure 3.5: ***Uncertainty-directed attention.*** (A) Distribution of fixation advantage of the fixated item, computed at the beginning of each new fixation. Fixation advantage is defined as the cumulative fixation time to the item minus the mean cumulative fixation time to the other item(s). First fixations 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. See Figure 3.3 for more details.

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 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 there-

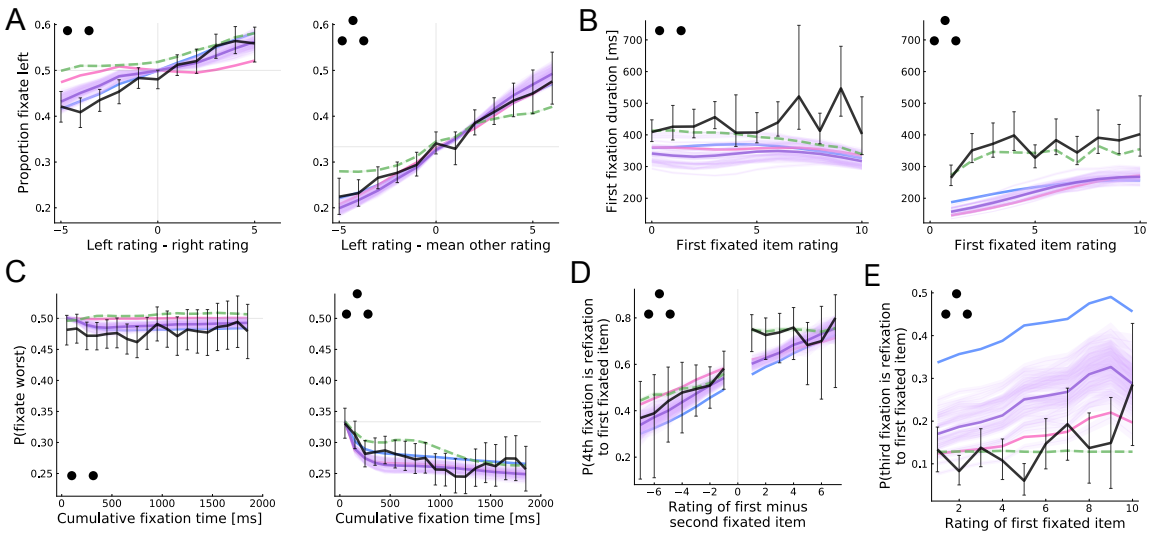

Figure 3.6: ***Value-directed attention.*** (A) Proportion of time fixating the left item as a function of its relative rating. (B) First fixation duration as a function of the rating of the first-fixated item. (C) Probability of fixating the lowest rated item as a function of the cumulative fixation 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. See Figure 3.3 for more details.

fore less likely to receive more fixations. However, we see that the predicted effect is much 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 (see Figure 3.7A 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.

Figure 3.6 shows that our main model provides a better prediction of some fixation patterns, whereas the aDDM provides a better fit of 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, literally, 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 and Rangel, 2008; Armel et al., 2008; Krajbich et al., 2010; Krajbich and 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

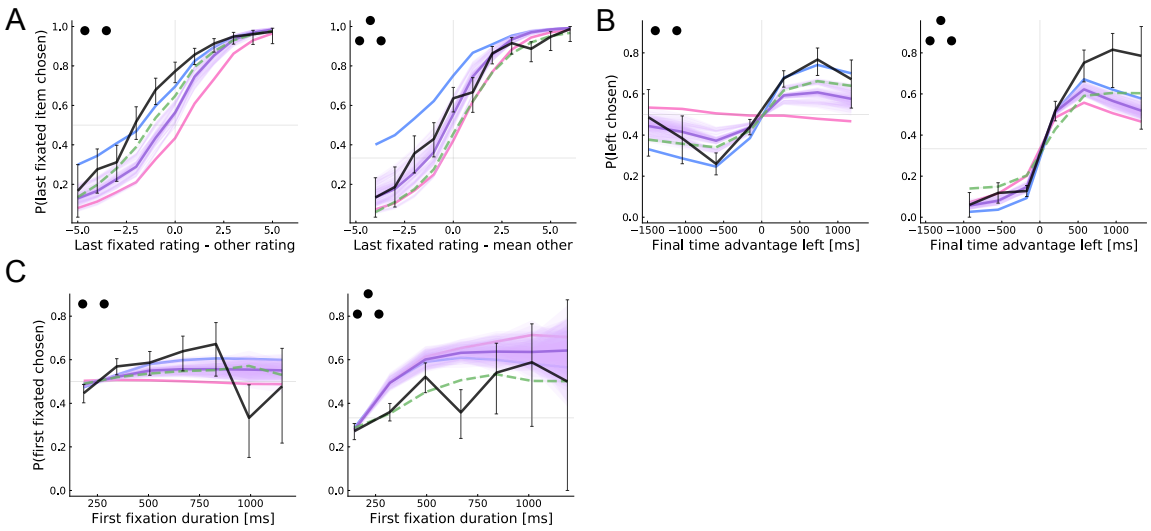

Figure 3.7: *Choice biases.* (A) Probability that the last fixated item is chosen as a function of its relative rating. (B) Probability that the left item is chosen as a function of its final fixation advantage, given by total fixation time to the left item minus the mean total fixation time to the other item(s). (C) Probability of choosing the first-seen item as a function of the first-fixation duration. See Figure 3.3 for more details.

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 and 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.4.

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 and McKoon, 2008; Teodorescu and Usher, 2013; Busemeyer and Townsend, 1993; Holmes et al., 2016) and that the process is modulated by visual attention (Krajbich et al., 2010; Krajbich and Rangel, 2011;

Gluth et al., 2018, 2020; Song et al., 2019; Smith and 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 and Wefald, 1991; 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 and 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 and Carpenter, 1976; Orquin and Mueller Loose, 2013): what a person is looking at is a good indicator of what they are thinking about. Im-

portantly, 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 and 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 and Krajbich, 2019). While the prior-bias mechanism predicts a multiplicative effect, the value-directed attention mechanism predicts that fixa-

tion 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 3.1). 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 and 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 and Stewart, 2018; Russo and Dosher, 1983; Trueblood et al., 2014; Usher and McClelland, 2004; Berkowitsch et al., 2014; Fisher, 2017; Krajbich et al., 2012; Westbrook et al., 2020; Shi et al., 2013; Manohar and 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. Extending the methods in this paper to that 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-choic`

3.4.1 Approximating the optimal policy

As described in Section 2.2, 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 belief state. The optimal metalevel policy, π^*, is the one that maximizes expected total metalevel

reward (Equation 2.5). In our model, we can write this as

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

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 **??**. In Appendix **??**, 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 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). We begin by sampling 8000 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 for which no computation has positive $\widehat{\text{VOC}}$ in the initial belief state. We then 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.

How good is the approximation method? In Appendix **??**, we show that this approach generates near-optimal policies on a related problem, with Bernoulli-distributed samples and no switching costs (Callaway et al., 2018). 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.6 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 belief 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 belief 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

$$u^{(i)} = \frac{r^{(i)} - \bar{\mu}}{\bar{\sigma}}, \tag{3.6}$$

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 belief 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 belief state, $b_0 = (\boldsymbol{\mu}_0 = 0, \boldsymbol{\lambda}_0 = 1, f_0 = \oslash)$. Note that $f_0 = \oslash$ indicates that no item is fixated at the onset of a trial.

The agent then selects an initial computation $c_0 \sim \pi(b_0)$ and the belief is updated according to the transition dynamics (Equation 3.4). Note that $\pi(c \mid b_0)$ assigns equal sampling probability to all of the items, since the subject starts with symmetrical beliefs. This process repeats until some time step, T, when the agent selects the termination action, $\perp$. The predicted choice is the item with maximal posterior value, $i^*_T = \operatorname{argmax}_i \mu^{(i)}_T$. In the event of a tie, the choice is sampled uniformly from the set of items with maximal expected value in the final belief state; in practice, this never happens with well-fitting parameter values.

To translate the sequence of computations into a fixation sequence, we assume that each sample takes 100 ms 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 DM'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 3.6.

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.3) 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 and 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 performance 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 and 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, $\hat{\theta} = \operatorname{argmax} \mathcal{L}(D \mid \theta)$. 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 and Bengio, 2012).

Note that the optimal policy does not depend on α because the DM 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 DM 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 $\beta = 330.0 \pm 97.9$.

Acknowledgements We thank Ian Krajbich for his help in simulating the aDDM and Bas van Opheusden for suggesting the method for efficiently computing VOI_{full}.

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 1890

4

Memory

Nulla facilisi. In vel sem. Morbi id urna in diam dignissim feugiat. Proin molestie tortor eu velit. Aliquam erat volutpat. Nullam ultrices, diam tempus vulputate egestas, eros pede varius leo.

Quoteauthor Lastname

5

Planning

Rational use of cognitive resources in human planning[*]

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 and Morgenstern, 1944; Stahl and Wilson, 1994; Camerer and Ho, 2004) and computer science (Newell and 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 and Simon, 1956; Newell et al., 1959, 1972). Even at this early stage, it was immediately recog-

[*]This chapter is based on the following paper:
Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Lieder, F., and Griffiths, T. L. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, pages 1–14.

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 and 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 and Griffiths, 2001), categorization (Anderson, 1991; Ashby and Alfonso-Reese, 1995), perception (Knill and Richards, 1996), and information-seeking (Oaksford and Chater, 1994; Gureckis and 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 and 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 and 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 and Wefald, 1991), 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

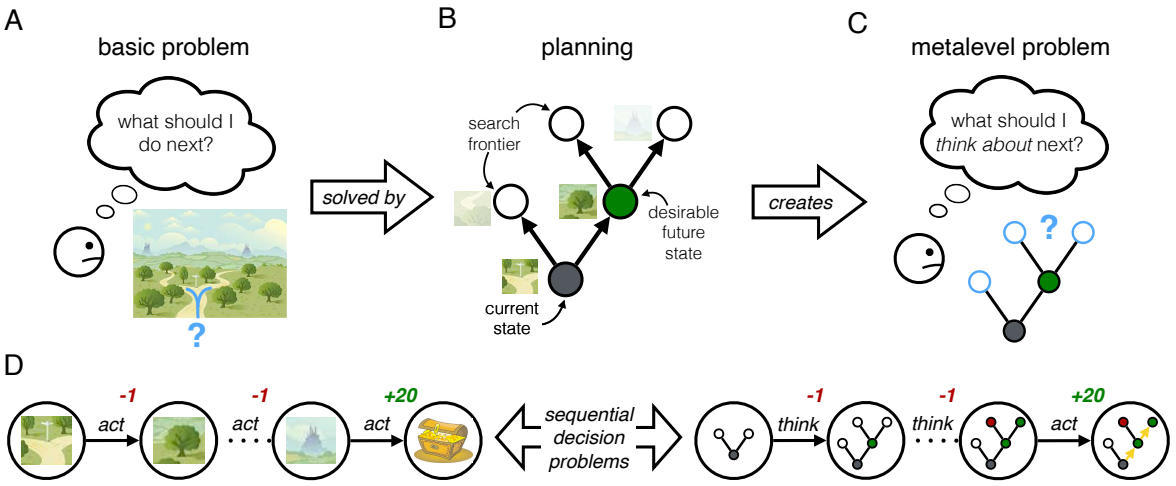

Figure 5.1: ***Formalizing planning under computational constraints.*** (A) The basic problem facing an intelligent agent is to take actions that maximize long-run reward. If the agent can predict the consequences of their actions, they can solve this problem by planning. (B) In one version of planning, the agent constructs a decision tree, where nodes (circles) represent possible future states of the world 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. But while the basic problem is defined by states of the world, physical actions, and external rewards, the metalevel problem is defined by decision trees and the mental operations that build them; the metalevel rewards capture both cognitive costs and also the external reward gained by executing the chosen plan. By formalizing planning in this way —concretely, as a Markov decision process (MDP)—we can use standard MDP-solving techniques to identify optimal planning strategies.

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 and 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 plan-

ning 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 belief states correspond to partially constructed decision trees and the computations correspond to expanding a node in the tree. We detail the four components of the metalevel MDP below.

Beliefs A belief state, $b \in \mathcal{B}$, 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. Thus, the largest possible decision tree has the same graphical structure as the environment itself. Let N be the size of this tree. We can then represent a decision tree (that is, a belief state) as a vector of length N where each position corresponds to a node in the tree (and a world state). The values, b_i, specify either the reward that can be attained at the world state i, or a special value, $\oslash$, indicating that the corresponding node has not been expanded yet. In the initial belief state, only the root node (the current world state) has been expanded, always having value 0; all other nodes have value $\oslash$.

Computations A computation, $c \in \mathcal{C}$, corresponds to expanding a node of the decision tree. This operation determines the cost or reward for visiting a 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

$$\text{frontier}(b) = \{c_i \mid b_i = \oslash \wedge b_{\text{parent}(i)} \neq \oslash\} \tag{5.1}$$

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

As in all metalevel MDPs, there is an additional operation, $\perp$, that terminates the computation process. Upon terminating planning, the agent executes an action sequence that has maximal expected value according to the decision tree it has built up until that point.

Transition function The metalevel transition function specifies the effect of node expansion on the decision tree. Executing c_i produces an updated b' that is identical to b except that b'_i is set to a reward which is sampled from a node-specific distribution, R_i. The node-dependent reward distribution is a key aspect of the environment that we will manipulate in Experiments 2 and 3.

Reward function Finally, the metalevel reward function captures both the cost of computation and the quality of the plan that is ultimately executed. Specifically, we assume that node expansion has a fixed cost, corresponding to the effort and time spent executing the operation. To capture plan quality, the reward for the termination action is the expected value of the external rewards one will attain while executing the chosen plan. The expected value of a plan is the sum of rewards up to and including the associated node, plus (for an incomplete plan) the expectation of the unknown future rewards. The chosen plan is the one that maximizes this expected value. Thus, the reward for the termination action is equal to the maximal expected value of any plan.

Formally, we define

$$R(b,c) = \begin{cases} \max_{p \in \mathcal{P}} V(b,p) & \text{if } c = \perp \\ -\lambda & \text{otherwise} \end{cases} \tag{5.2}$$

where λ is the cost of node expansion (a free parameter), p is a complete plan (i.e., a sequence of object-level states beginning with the current state and ending with a terminal state), $\mathcal{P}$ is the set of all such plans, and $V(b,p)$ is the expected value of executing a plan given the current belief state, that is

$$V(b,p) = \sum_{i \in p} \begin{cases} \mathrm{E}[R_i] & \text{if } b_i = \oslash \\ b_i & \text{otherwise.} \end{cases} \tag{5.3}$$

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 belief 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. In the terminology of MDPs, the optimal policy selects actions that maximize the optimal state-action value function, $Q(b, c)$. This function specifies the expected total reward an agent will receive (including both cost and decision quality) if she executes the node expansion action c in belief state b and continues selecting actions optimally until termination. Importantly, this function 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 error model used to fit human data; see Section 5.5.5).

Exactly computing Q for a large MDP is very computationally intensive. Early work in rational metareasoning proposed that this function can be approximated by a myopic one-step lookahead (Russell and Wefald, 1991). This 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 and 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

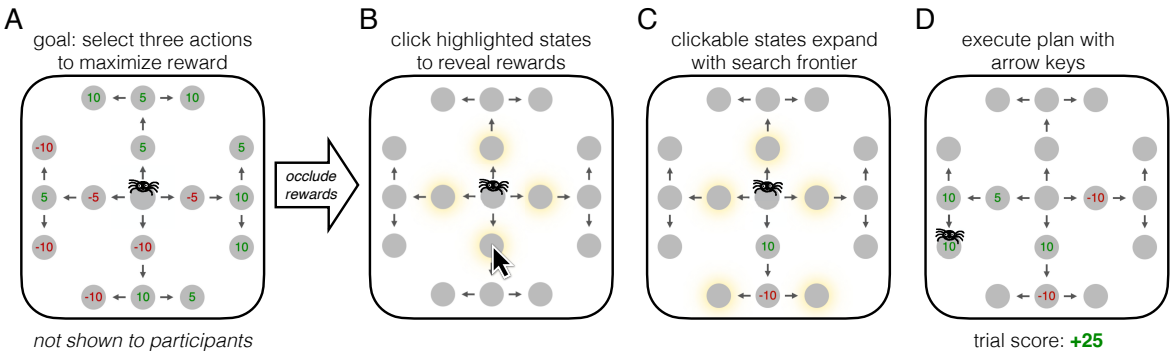

Figure 5.2: ***Experimental task.*** (A) Participants are presented with a sequential decision problem displayed as a graph. Gray circles indicate states, arrows indicate actions, and green and red numbers indicate rewards and punishments. (B) Rewards are initially occluded, but can be revealed by clicking 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.

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 and 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 limitb" 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 mechanismb". 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 5.5.5).

5.2 Task: Mouselab-MDP

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 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 and 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 and 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. Concretely, in the task illustrated in Figure 5.2 there are 8 possible choice sequences and over 2.7 billion node expansion sequences.

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 participantb' 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.

The task, "Mouselab-MDP", 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 and Simon, 1973) provide support for using this approach. We return to this point in the Discussion.

5.3 Results

5.3.1 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 participantb' 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 5.5.9).

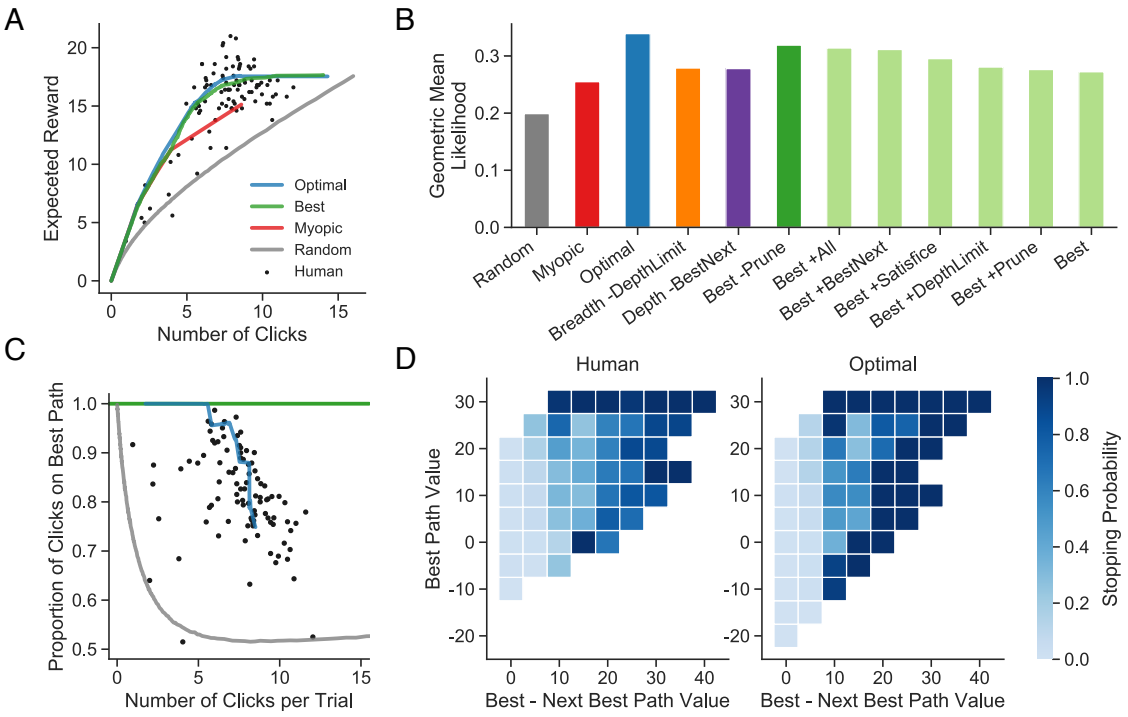

Figure 5.3: ***Experiment 1 results.*** (A) Pareto curves. Each point shows the average reward attained and number of clicks made by a participant (black dots) or model (colored lines). Note that with a small number of trials, it is possible to exceed the expected performance of the optimal model by 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.

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 participantb' 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 and Smith, 2004; Teodorescu and Usher, 2013)

and maximizing accuracy (McMillen and Holmes, 2006; Bogacz et al., 2006). However, although both the model's and our participantb' 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 $\Delta_{\text{LL}} > 3981$). In terms of total likelihood, it also outperformed best-first search (all $\Delta_{\text{LL}} > 1250$), 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.

5.3.2 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 vari-

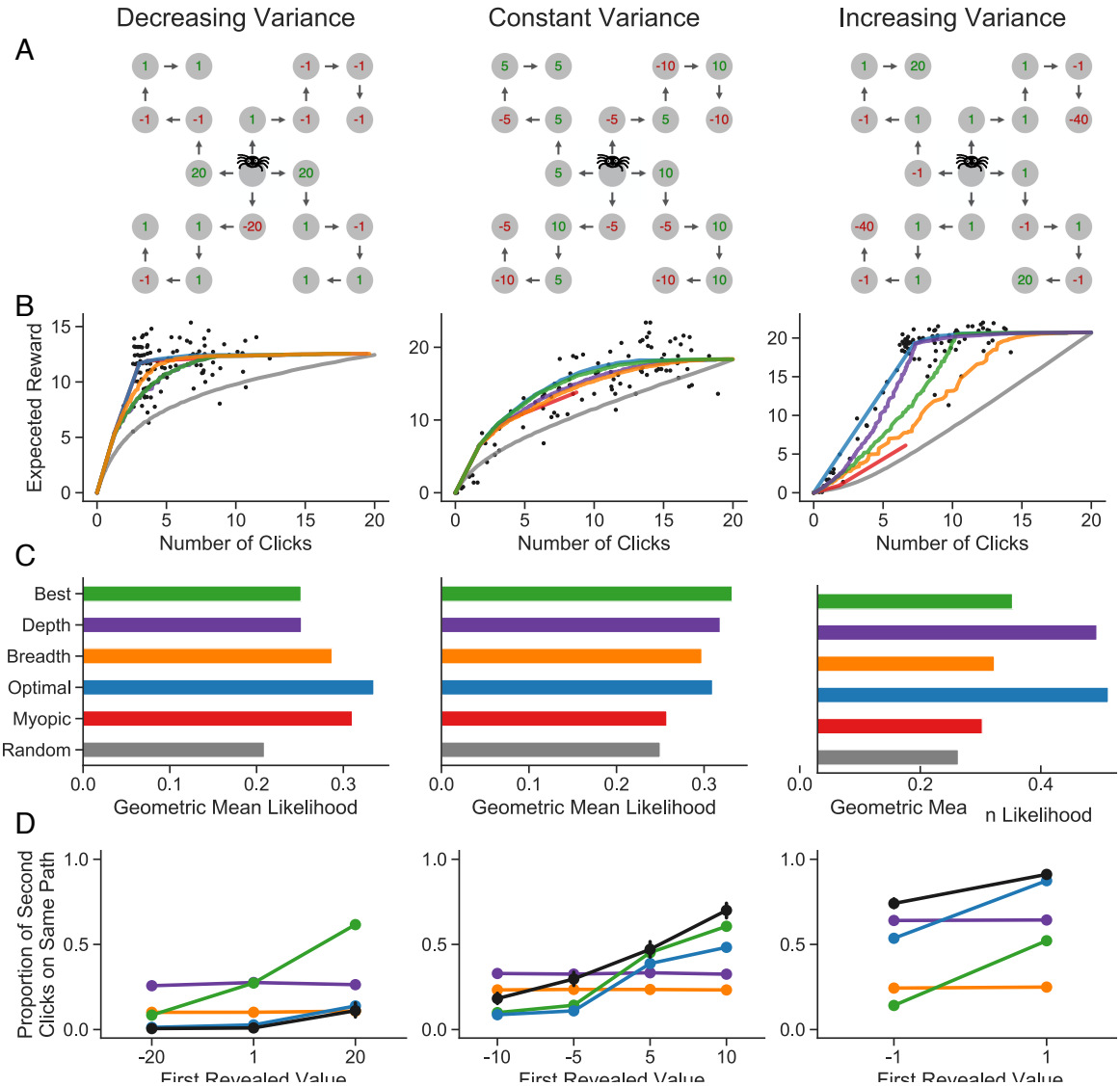

Figure 5.4: ***Experiment 2 results.*** Each column shows one experimental condition. (A) Example trials. Each condition is characterized by a different location-dependent reward distribution in which large values are found at the beginning of each path, at any location, or at the end of each path, respectively. (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). In each condition one classical algorithm achieves near optimal performance. (C) Model comparison. Best, Depth, and Breadth refer to the versions of the model that performed best in Experiment 1, as shown in Figure 5.3. Of these classical algorithms, the one that achieves the best reward-click trade off (shown in panel B) also best predicts human behavior. Depth limits are excluded because they allow the best-first and depth-first models to mimic breadth-first search. (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. Human data is in black and model colors match panel C. For human data, points show means and error bars show bootstrapped 95% confidence intervals, both computed across participants. In each condition one of the classical planning strategies captures the qualitative behavioral pattern, but only the optimal model captures the pattern in every condition. All heuristic mechanisms are excluded from this plot. See Figure B.1 for the same plot with the full models (including Myopic).

ance; 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 $\Delta_{\text{LL}} > 446$). 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 predicts human behavior in the increasing ($\Delta_{\text{LL}} = 606$) and decreasing ($\Delta_{\text{LL}} = 1276$) conditions; the best-first model with best vs. next fits best in the constant condition (compared to optimal: $\Delta_{\text{LL}} = 2150$).

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.

5.3.3 Experiment 3: Backwards planning

In the previous experiments, we constrained participantb' 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

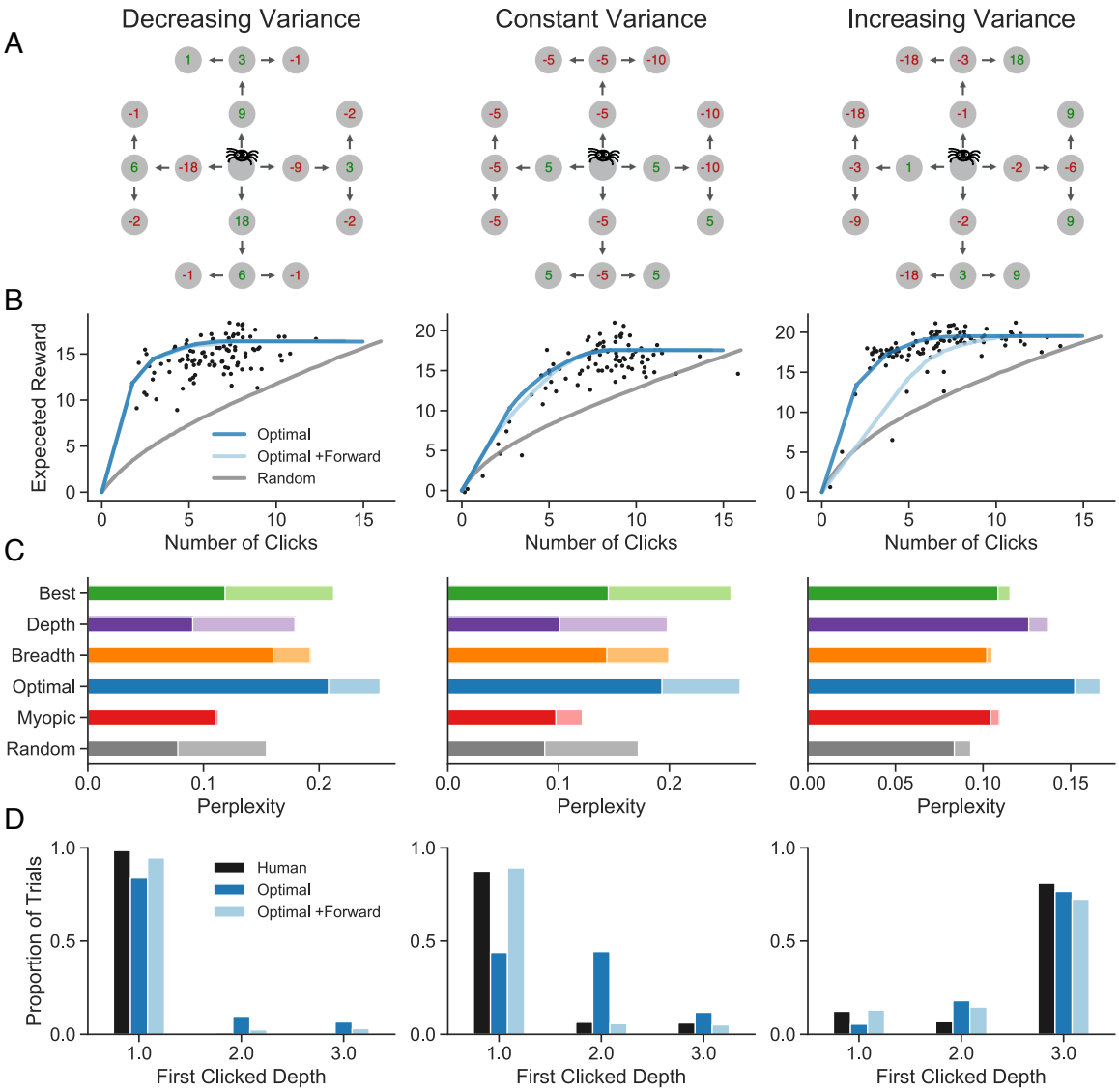

Figure 5.5: ***Experiment 3 results.*** (A) Example trials. Each condition is characterized by a different location-dependent reward distribution with standard deviation linearly increasing, decreasing, or remaining constant with depth. (B) Pareto curves. The light blue line shows the optimal model restricted 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).

(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 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.

5.3.4 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

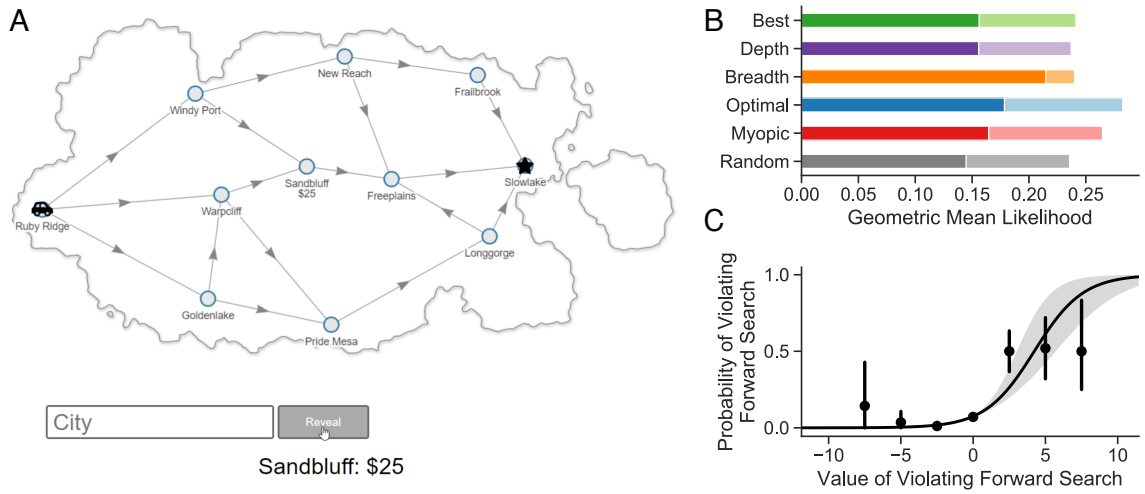

Figure 5.6: ***Experiment 4 results.*** (A) Task: participants acted as a travel agent, attempting to find a low-cost route from a start city to a goal city. They could reveal the price of passing through each city using a textual search interface. (B) Model comparison. Light bars show models augmented 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.

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 $[1.59, 3.37]$, $z = 5.48$, $p < .001$).

5.4 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 and Gaissmaier, 2011; Todd and Gigerenzer, 2003; Gigerenzer and Goldstein, 1996). However, while ecological rationality explicitly rejects the notion of optimality (Gigerenzer and 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 B.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 B.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

and 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 and 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 and Cutler, 2021). Using the approach of resource-rational analysis (Griffiths et al., 2015; Lieder and Griffiths, 2020), we can use the observed discrepancies to generate hypotheses about additional constraints (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 and Johnson, 1996). Nevertheless, there are at least five reasons to believe that the present results already reveal some-

thing 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 B.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. 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 and 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 and 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 and Chater, 1994; Gureckis and Markant, 2012; Tenenbaum and Griffiths, 2001; Anderson, 1991; Ashby and Alfonso-Reese, 1995; Savage, 1954). However, the constraints imposed by the external environment are insufficient to explain many key features of human cognition (Rahnev and Denison, 2018; Lieder and 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 and 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.

5.5 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 (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). We describe deviations from the pre-registered analysis plan in Appendix B.1.

5.5.1 Experiment 1

We recruited 104 participants (28.7 ± 8.2 years; 50 female, 5 not specified) 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 ± 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 re-

vealed), 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.

5.5.2 Experiment 2

All aspects of the design were identical to Experiment 1 except where noted otherwise. We recruited 313 participants (31.7 ± 11.1 years; 162 female, 12 not specified). 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 ± 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.

5.5.3 Experiment 3

All aspects of the design are identical to Experiment 2 except where noted otherwise. We recruited 319 participants (32.3 $\pm$ 11.8 years; 173 female, 20 not specified). 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 $\pm$ 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.

5.5.4 Experiment 4

We recruited 137 participants (33.4 $\pm$ 12.2 years; 55 female, 36 not specified) 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 $\pm$ 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.

5.5.5 Model specifications

Each of our candidate models corresponds to a parameterized family of metalevel policies. A policy is defined by a state-conditional distribution over actions, $\pi(a \mid s)$. For all models, this distribution 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 action, $\perp$. Second, if the frontier includes at least one node, then a random legal action is executed with some probability, ε. Otherwise (Step 3), the model executes $\perp$ with probability $p^M_{\text{stop}}(s)$; 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 $p^M_{\text{select}}(s, a)$ 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,

$$p^{H}_{\text{stop}}(s) = \frac{1}{1 + \exp\left\{-f_{\text{stop}}(s)\right\}}, \tag{5.4}$$

where

$$f_{\text{stop}}(s) = \beta_{\text{satisfice}} \cdot V_{\text{best}} + \beta_{\text{bestnext}} \cdot (V_{\text{best}} - V_{\text{next}}) + \theta_{\text{stop}}. \tag{5.5}$$

V_{best} and V_{next} are the expected values of the best and second-best paths given the current belief 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

$$p^{\text{SATISFICE}}_{\text{stop}}(s) = \frac{1}{1 + \exp\left\{-\beta_{\text{satisfice}}(V_{\text{best}} - \theta)\right\}}, \tag{5.6}$$

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

$$p^{H}_{\text{select}}(s, a) = \frac{1(a \in \text{frontier}(s)) \cdot \exp\left\{\beta_{\text{select}} \cdot f^{\text{ALG}}_{\text{select}}(s, a)\right\}}{\sum_{a' \in \text{frontier}(s)} \exp\left\{\beta_{\text{select}} \cdot f^{\text{ALG}}_{\text{select}}(s, a')\right\}}, \tag{5.7}$$

where $f^{\text{ALG}}_{\text{select}}(s, a)$ denotes a node-scoring function for each algorithm; specifically,

$$\begin{aligned} f^{\text{BEST}}_{\text{select}}(s, c_i) &= V(s, i) = \max_{p \in \{\mathcal{P} | i \in p\}} V(s, p) \\ f^{\text{DEPTH}}_{\text{select}}(s, c_i) &= \text{depth}(s, i) \\ f^{\text{BREADTH}}_{\text{select}}(s, c_i) &= -\text{depth}(s, i). \end{aligned} \tag{5.8}$$

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 $f^{\text{RAND}}_{\text{select}}(s, a) = 0$ and $f_{\text{stop}}(s) = \theta_{\text{stop}}$. 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_λ, of the metalevel MDP with computational cost λ. We computed the Q function using dynamic programming. The stopping rule is

$$p^O_{\text{stop}}(s) = \frac{\exp\left\{\beta_{\text{stop}} \cdot Q_\lambda(s, \perp)\right\}}{\sum_{a' \in \text{frontier}(s) \cup \{\perp\}} \exp\left\{\beta_{\text{stop}} \cdot Q_\lambda(s, a')\right\}} \tag{5.9}$$

and the selection rule is

$$p^O_{\text{select}}(s, a) = \frac{\exp\left\{\beta_{\text{select}} \cdot Q_\lambda(s, a)\right\}}{\sum_{a' \in \text{frontier}(s)} \exp\left\{\beta_{\text{select}} \cdot Q_\lambda(s, a')\right\}}. \tag{5.10}$$

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_λ function is replaced by a myopic one-step approximation (Russell and Wefald, 1991), which we denote $Q^{\text{myopic}}_\lambda$. For the termination action, this approximation is exact because the trial ends after this action is executed and thus $Q_\lambda(s, \perp) = Q^{\text{myopic}}_\lambda(s, \perp) = r(s, \perp)$. 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

$$Q^{\text{myopic}}_\lambda(s, a) = \mathop{\mathrm{E}}_{b' \sim T(\cdot|s,a)} [r(b', \perp)] - \lambda. \tag{5.11}$$

5.5.6 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

$$p_{\text{prune}}^{\text{VALUE}}(s,i) = \frac{1}{1+\exp\left\{-\beta_{\text{prune}}^{\text{VALUE}} \cdot (\theta_{\text{prune}}^{\text{VALUE}} - V(s,i))\right\}} \tag{5.12}$$

where $V(s,i)$ is the value of the best path that includes node i, defined in Equation 5.8. 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

$$p_{\text{prune}}^{\text{DEPTH}}(s,i) = \frac{1}{1+\exp\left\{-\beta_{\text{prune}}^{\text{DEPTH}} \cdot (\text{depth}(s,i) - \theta_{\text{prune}}^{\text{DEPTH}})\right\}}. \tag{5.13}$$

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

$$p_{\text{prune}}(s,i) = 1 - \left(1 - p_{\text{prune}}^{\text{VALUE}}(s,i)\right) \cdot \left(1 - p_{\text{prune}}^{\text{DEPTH}}(s,i)\right). \tag{5.14}$$

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 5.4 and 5.7), we marginalize over all possible frontiers that could result from different combinations of pruning decisions, weighing each by its probability according to Equation 5.14.

5.5.7 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: frontier(s) in Equations 5.7, 5.9, and 5.10 is replaced with

unexpanded$(s) = \{c_i \mid s_i = \oslash\}$. 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, $\beta_{\text{forward}} \cdot \mathbb{1}(a \in \text{frontier}(s))$. 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 5.10.

5.5.8 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.

5.5.9 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

- What is the reward function? (See footnote in introduction) - Dissociation between agent and environment is kind of homunculus like

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 and 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 during 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 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 (**?**). 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.3 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.* For each configuration, the maximum likelihood estimate of each parameter was its mean in the 30 configurations with highest likelihood (following our reporting

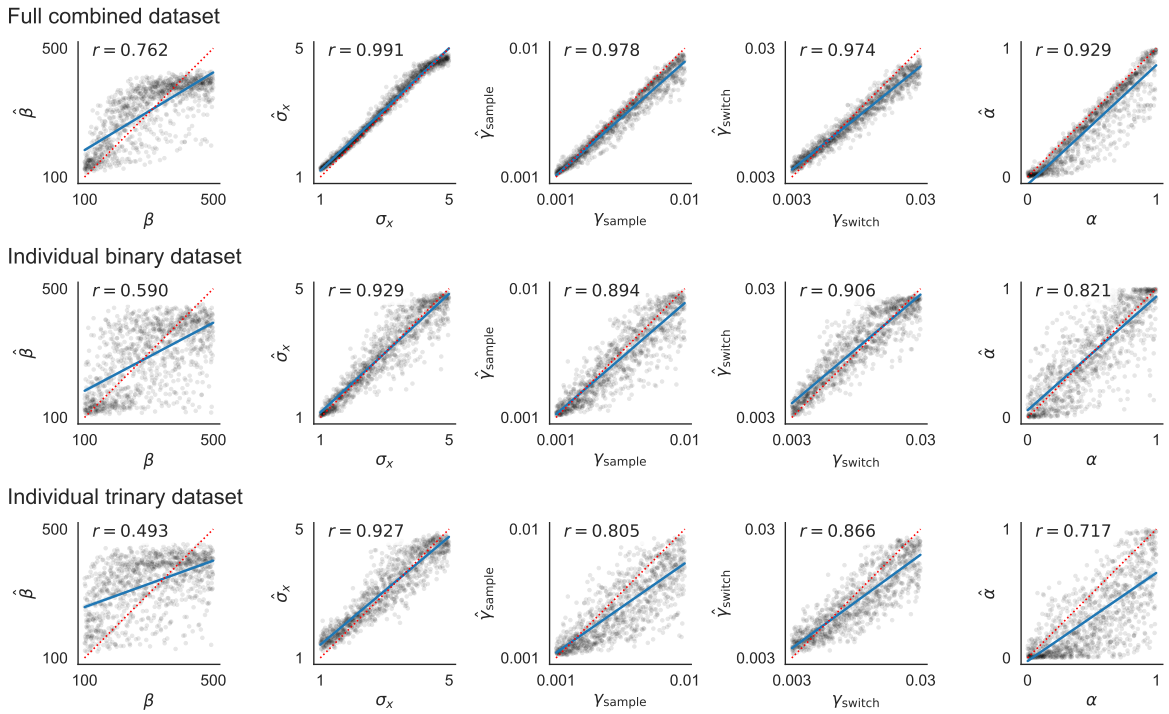

Figure A.1: ***Parameter recovery***. Each panel plots the estimated parameter value as a function of the true parameter value. Each black dot corresponds to one simulated dataset. The dotted red line shows equality (i.e., perfect recovery) and the solid blue line shows the linear trend. The top row 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.

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 correlation 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.

*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.

A.4 Implementation and validation of the aDDM

In order to compare our model to the predictions of the aDDM (Krajbich et al., 2010; Krajbich and 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

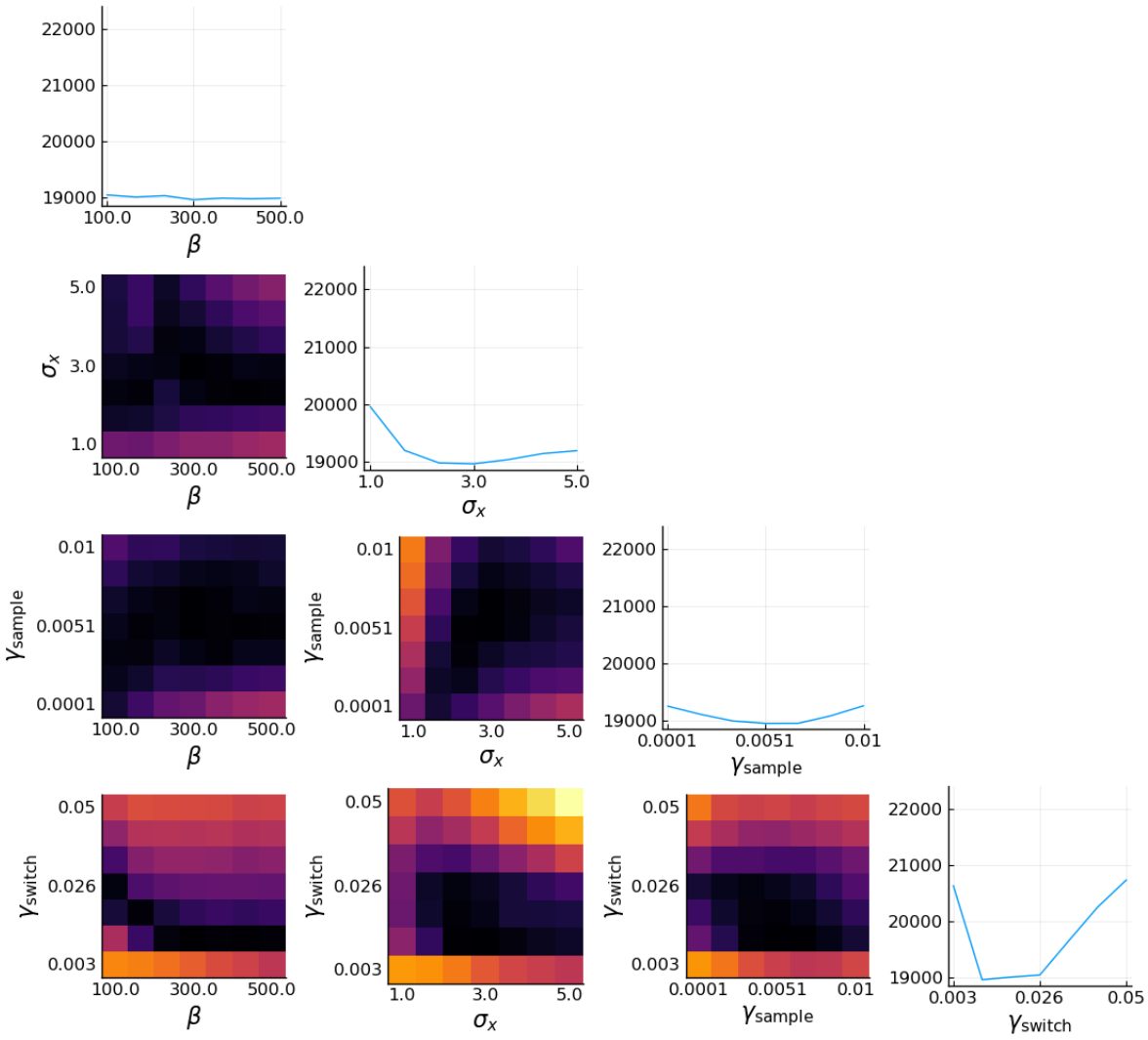

Figure A.2: ***Grid search on model parameters.*** Each panel shows the best likelihood achieved for each value of one parameter (diagonal) or combination of values for two parameters (off-diagonal), i.e., minimizing over all the non-plotted variables.

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.

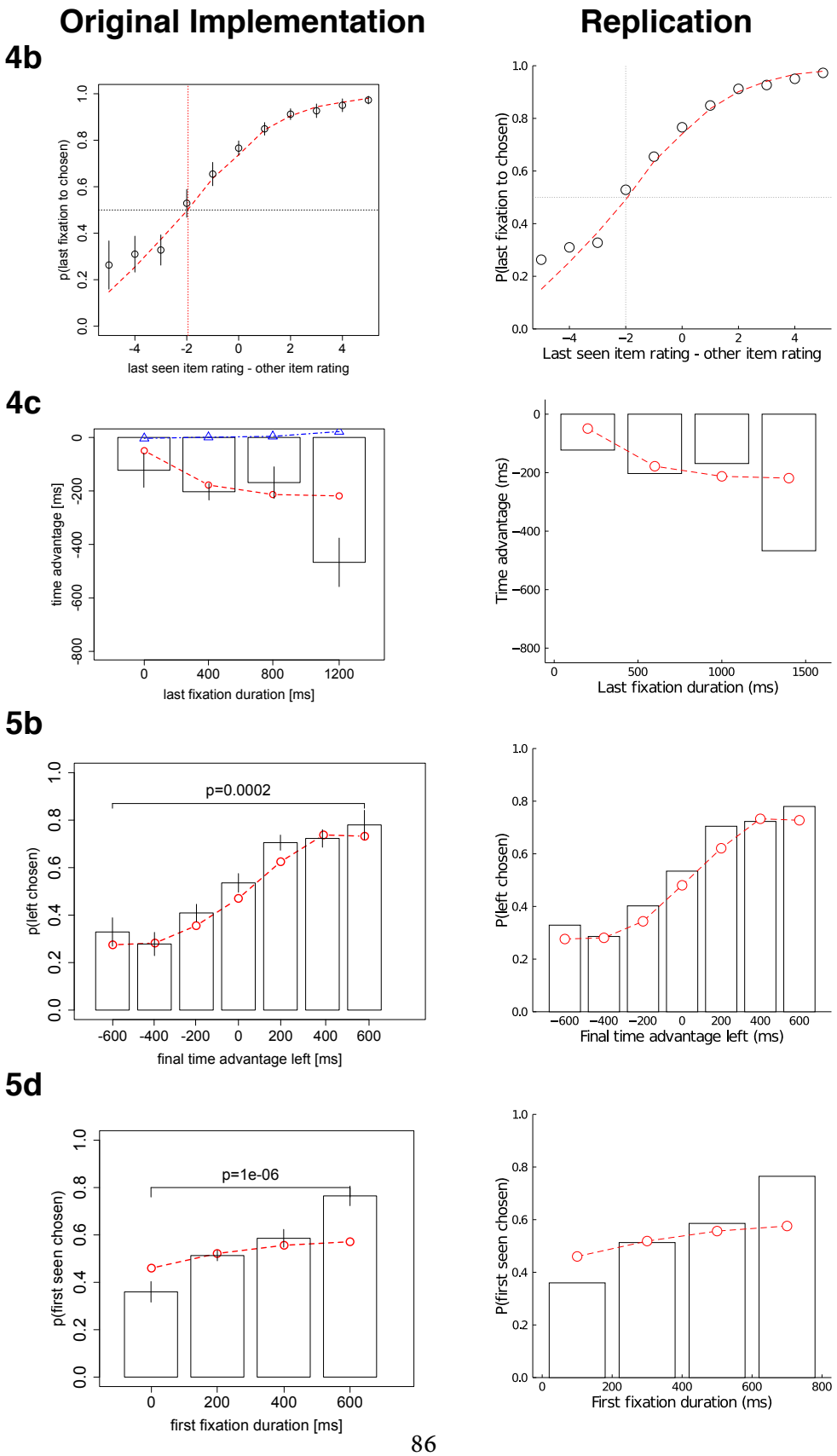

Figure A.3: ***Replication of*** *Krajbich et al. (2010)*. Note that x axis labels in the orginal plots sometimes reflected the left tail of the bin; in these cases, we adjusted the tick locations accordingly.

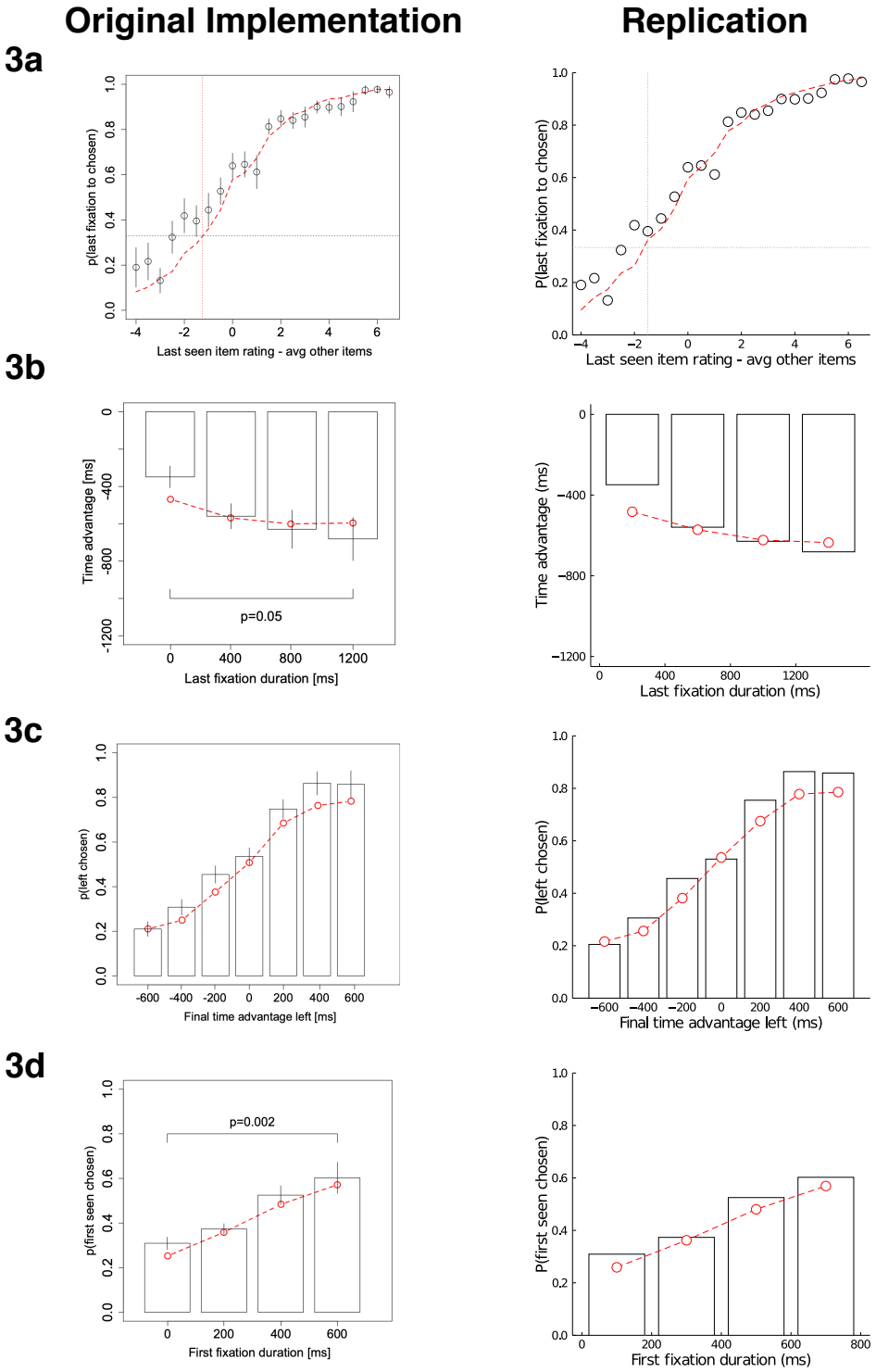

Figure A.4: ***Replication of*** *Krajbich and Rangel (2011)*. Note that there are slight deviations in model predictions due to noise in the simulations; the orginal plots are based on 2000 simulated trials.

A.5 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.5.1 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,

$$\mathrm{VOI}_{\mathrm{myopic}}(b_t, c) = \mathop{\mathrm{E}}_{b_{t+1}|b_t,c}[R(b_{t+1}, \perp)] - R(b, \perp).$$

In our model, this is equal to the expected value of the item that will be chosen after taking an additional sample minus the expected value of an item chosen based on the current beliefs. That is,

$$\mathrm{VOI}_{\mathrm{myopic}}(b_t, c) = \mathop{\mathrm{E}}_{\mu_{t+1}|\mu_t,\lambda_t}\left[\max_i \mu_{t+1}^{(i)}\right] - \max_i \mu_t^{(i)}.$$

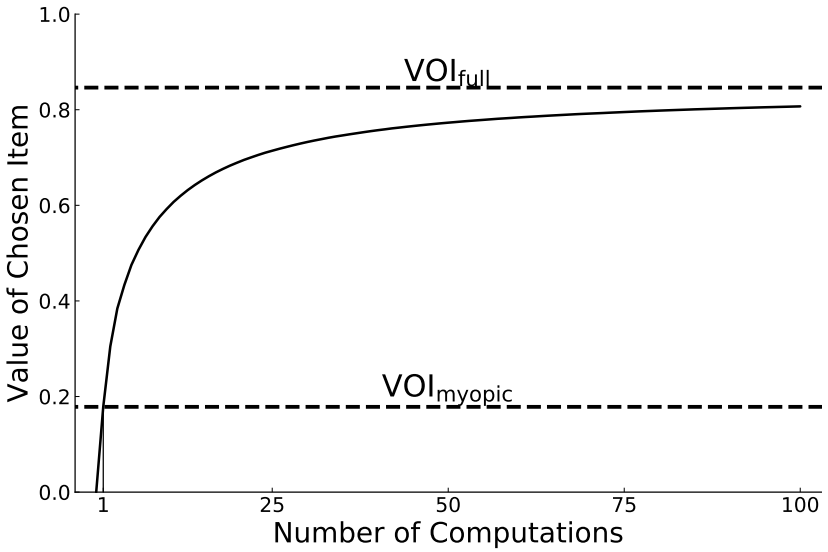

Figure A.5: ***Illustration of the value of information features.*** The solid line shows the average value of the item chosen after different numbers of computations selected by a near-optimal policy assuming no computational costs. The dashed lines show values for two of the VOI features in the initial belief 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 term as

$$\mathop{\mathrm{E}}_{\mu_{t+1}^{(c)} | \mu_t^{(c)}, \lambda_t^{(c)}} \left[\max \left\{ \mu_{t+1}^{(c)}, \max_{i \neq c} \mu_t^{(i)} \right\} \right]. \tag{A.1}$$

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.1, 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 3.4, we have

$$\mu_{t+1}^{(c)} = \frac{\sigma_x^{-2} x_t + \lambda_t^{(c)} \mu_t^{(c)}}{\lambda_t^{(c)} + \sigma_x^{-2}}. \tag{A.2}$$

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 belief 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.2, we have

$$\mu_{t+1}^{(c)} \sim \text{Normal}(\mu_\mu, \sigma_\mu^2)$$

where

$$\mu_\mu = \frac{\sigma_x^{-2}}{\lambda_{t+1}^{(c)}} \mu_t^{(c)} + \frac{\lambda_t^{(c)} \mu_t^{(c)}}{\lambda_{t+1}^{(c)}} = \mu_t^{(c)}$$

and

$$\sigma_\mu^2 = \left(\frac{\sigma_x^{-2}}{\lambda_{t+1}^{(c)}} \right)^2 \left(\frac{1}{\lambda_t^{(c)}} + \sigma_x^2 \right).$$

Having derived the distribution of Equation $\mu_{t+1}^{(c)}$, we now turn to the expected maximum in [A.1]. From basic probability theory we know that for any constant z and random variable X,

$$\mathrm{E}[\max\{X, z\}] = \Pr[X \leq z] \cdot z + (1 - \Pr[X \leq z]) \cdot \mathrm{E}[X \mid X > z]. \tag{A.3}$$

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

$$\Pr\left[\mu_{t+1}^{(c)} \leq \max_{i \neq c} \mu_t^{(i)}\right] = \Phi(\beta),$$

where Φ is the cumulative density function (CDF) of a standard Gaussian , and

$$\beta = \frac{\max_{i \neq c} \mu_t^{(i)} - \mu_\mu.}{\sigma_\mu}$$

Next, we apply the standard formula for the expectation of a truncated Gaussian, giving us

$$E\left[\mu_{t+1}^{(c)} \mid \mu_{t+1}^{(c)} > \max_{i \neq c} \mu_t^{(i)}\right] = \mu_\mu + \frac{\varphi(\beta)}{1 - \Phi(\beta)}\sigma_\mu,$$

where φ is the standard normal probability density function. Finally, putting this together we find that $\mathrm{VOI}_{\mathrm{myopic}}(b, c)$ is equal to

$$\Phi(\beta) \cdot \max_{i \neq c} \mu_t^{(i)} + (1 - \Phi(\beta)) \cdot \left(\mu_\mu + \frac{\varphi(\beta)}{1 - \Phi(\beta)}\sigma_\mu\right) - \max_i \mu_t^{(i)}.$$

A.5.2 Value of Perfect Information About One Item

Whereas $\mathrm{VOI}_{\mathrm{myopic}}$ captures the information value of a single sample, $\mathrm{VOI}_{\mathrm{item}}$ 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,

$$\mathrm{VOI}_{\mathrm{item}}(b_t, c) = \mathop{\mathrm{E}}_{u^{(c)} \mid \mu_t^{(c)}, \lambda_t^{(c)}}\left[\max\left\{u^{(c)}, \max_{i \neq c} \mu_t^{(i)}\right\}\right] - \max_i \mu_t^{(i)}.$$

The derivation is similar to that of $\mathrm{VOI}_{\mathrm{myopic}}$, 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, $u^{(c)} \mid \mu_t^{(c)}, \lambda_t^{(c)} \sim \mathrm{Normal}(\mu_t^{(c)}, 1/\lambda_t^{(c)})$. Thus, we apply the same steps beginning with [A.3], but replacing $\mu_{t+1}^{(c)}$ with $u^{(c)} \mid \mu_t^{(c)}, \lambda_t^{(c)}$. This results in $\mathrm{VOI}_{\mathrm{item}}(b, c)$ equal to

$$\Phi(\beta') \cdot \max_{i \neq c} \mu_t^{(i)} + \left(1 - \Phi(\beta')\right) \cdot \left(\mu_t^{(c)} + \frac{\varphi(\beta')}{1 - \Phi(\beta')}\sqrt{1/\lambda_t^{(c)}}\right) - \max_i \mu_t^{(i)}$$

where

$$\beta' = \frac{\max_{i \neq c} \mu_t^{(i)} - \mu_t^{(c)}}{\sqrt{1/\lambda_t^{(c)}}}.$$

A.5.3 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,

$$\text{VOI}_{\text{full}}(b) = \mathop{\text{E}}_{u|\mu_t^{(c)},\lambda_t^{(c)}} \left[\max_i \left\{ u^{(i)} \right\} \right] - \max_i \mu_t^{(i)}. \tag{A.4}$$

For the case of N items, the conditional expectation term is given by the integral

$$\int \dots \int \left[\max_i u^{(i)} \prod_{i=1}^{k} \text{Normal}(u^{(i)}; \mu_t^{(i)}, 1/\lambda_t^{(i)}) \right] du^{(1)} \dots du^{(N)}.$$

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.

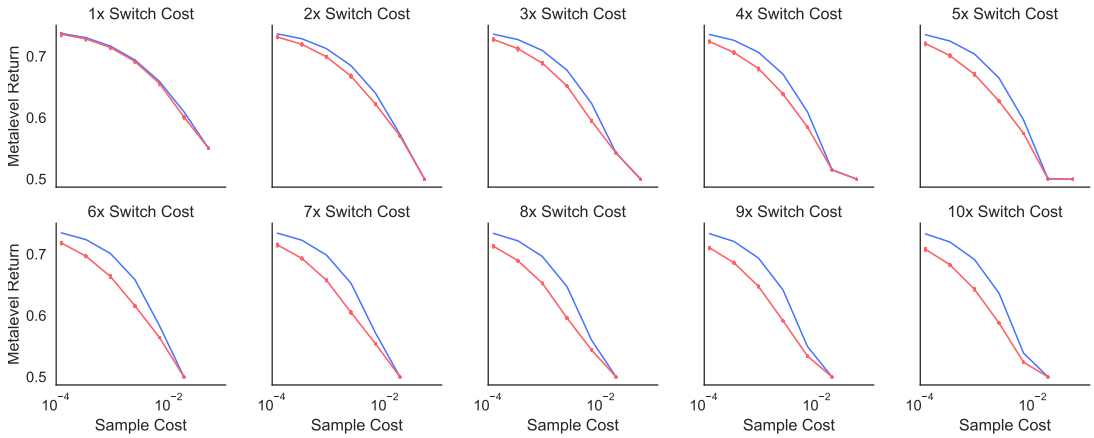

Figure A.6: ***Performance of BMPS on the Bernoulli model with switching costs.*** Each line shows the average metalevel return achieved by the BMPS policy (red) or the true optimal policy (blue). The red line shows mean performance from the top 80 policies identified by the UCB algorithm. Additionally, 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,

$$\mathrm{E}[X] = -\int_{-\infty}^{0} F_X(x)\,dx + \int_{0}^{\infty} (1 - F_X(x))\,dx, \tag{A.5}$$

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,

$$F_{\max \mathcal{X}}(x) = \prod_{X \in \mathcal{X}} F_X(x), \tag{A.6}$$

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 belief distributions for each item. Letting M denote $\max \mathcal{X}$, we can define its CDF as

$$F_M(m) = \prod_{i=1}^{N} \Phi\left(\sqrt{\lambda_t^{(i)}}\left(m - \mu_t^{(i)}\right)\right).$$

Combining Equations [A.4], [A.5], and [A.6], we arrive at the following expression for $\mathrm{VOI}_{\mathrm{full}}(b)$:

$$-\int_{-\infty}^{0} F_M(x)\,dx + \int_{0}^{\infty} (1 - F_M(x))\,dx - \max_i \mu_t^{(i)}.$$

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 $\widehat{\mathrm{VOC}}$.

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

$$c^* = \underset{c \neq \perp}{\operatorname{argmax}} \left\{ w_1 \cdot \mathrm{VOI}_{\mathrm{myopic}}(b, c) + w_2 \cdot \mathrm{VOI}_{\mathrm{item}}(b, c) - \mathrm{cost}(c) + w_4 \right\}. \tag{A.7}$$

Second, since $\mathrm{VOC}(b, \perp) = 0$, it follows that if $\widehat{\mathrm{VOC}}(b, c^*) > 0$, the policy should sample from item c^*, and otherwise it should stop sampling. In general, determining the sign of $\widehat{\mathrm{VOC}}(b, c^*)$requires evaluating $\mathrm{VOI}_{\mathrm{full}}(b)$. However, in some cases the sign 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 $\widehat{\mathrm{VOC}}(b, c)$ by replacing $\mathrm{VOI}_{\mathrm{full}}(b)$ with $\mathrm{VOI}_{\mathrm{item}}(b, c)$ in Equation [A.7]. 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 $\exp \beta \widehat{\mathrm{VOC}}(b, c)$. In particular, the standard method for sampling from this distribution requires fully evaluating $\widehat{\mathrm{VOC}}(b, c)$. However, we can circumvent this issue using the Gumbel-max trick (**?**), which provides a way to sample from a Boltzman (softmax) distribution by taking the argmax of the unexponentiated values corrupted by Gumbel noise. Formally,

$$\Pr\left[\underset{i}{\operatorname{argmax}} \{x_i + \varepsilon_i\} = j \right] = \frac{\exp x_j}{\sum_i \exp x_i}.$$

As a result, we can rewrite the soft-max policy as

$$\pi(b; \mathbf{w}, \beta) = \underset{c}{\operatorname{argmax}} \left\{ \beta \widehat{VOC}(b, c; w) + \varepsilon_c \right\},$$

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.7, and comparing the lower-bound VOC to an independent Gumbel sample, $\varepsilon_\perp$, rather than 0 to capture the noise applied to $\mathrm{VOC}(b, \perp)$.

A.6 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$(1 + a, 1 + b)$, where a and b are respectively the number of times 1 and 0 have been sampled for the given item. Critically, the resulting belief space is discrete because a and b 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 5

B.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], $z = 9.9$, $p < .001$.

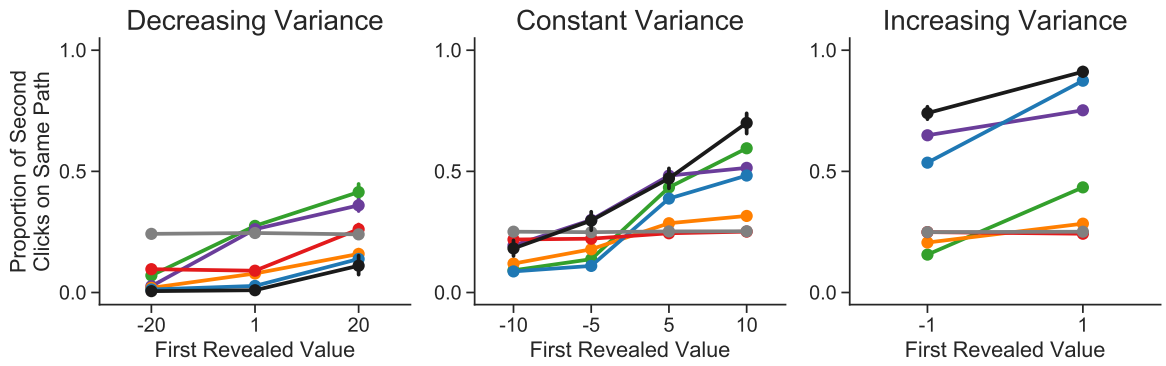

Figure B.1: ***Alternative version of Figure 5.4D.*** Here, the predictions of all alternative models are genreated with the best-fitting set of heuristic mechanisms. As in Figure 5.4D, points show means and error bars show bootstrapped 95% confidence intervals, both computed across participants.

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 main-text Figure 4D, we pre-registered that we would plot the predictions of the same models shown in main-text Figure 4D. However, as illustrated in Figure B.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

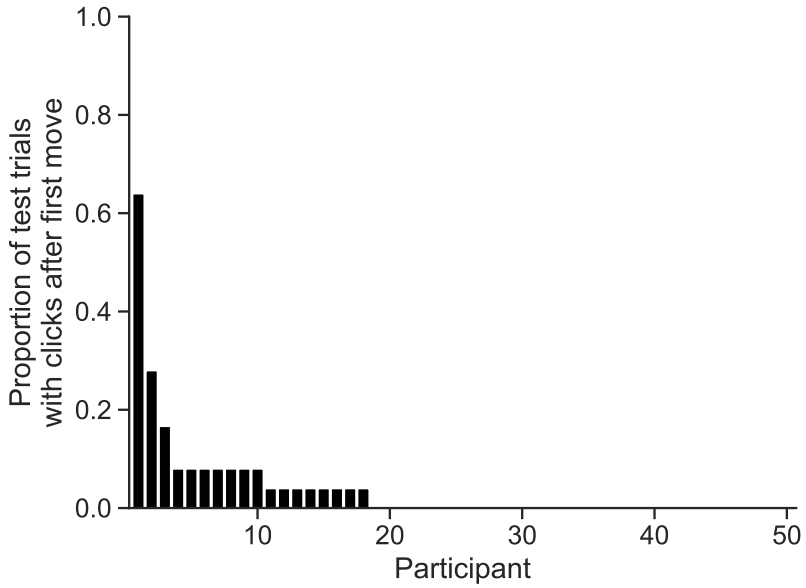

Figure B.2: ***Experiment 5 results.*** The proportion of test trials in which each participant made at least one click after moving the spider.

(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.

B.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 B.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.

B.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.

B.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

$$\begin{aligned} f_{\text{stop}}(s) = & \beta_{\text{satisfice}} \cdot \mathrm{E}[V_b] + \\ & \beta_{\text{bestnext}} \cdot \left(\mathrm{E}[V_b] - \max_{i \neq b} \mathrm{E}[V_i] \right) + \\ & \beta_{\text{pbetter}} \cdot \Pr(V_b \geq \alpha) + \\ & \beta_{\text{pbest}} \cdot \Pr\left(V_b \geq \max_{i \neq b} V_i \right) + \theta_{\text{stop}}. \end{aligned} \tag{B.1}$$

B.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 B.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 B.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 B.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

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. (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. and Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4):703–719.

Armel, K. C., Beaumel, A., and Rangel, A. (2008). Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making*, 3(5):396–403.

Armel, K. C. and 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. and Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, 39(2):216–233.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H. R., Reid, A., Verfaellie, M., Shadlen, M. N., and Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *eLife*, 8:undefined–undefined.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305.

Berkowitsch, N. A., Scheibehenne, B., and 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., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98.

Biderman, N., Bakkour, A., and Shohamy, D. (2020). What Are Memories For? The Hippocampus Bridges Past Experience with Future Decisions. *Trends in Cognitive Sciences*, 24(7):542–556.

Bitzer, S., Park, H., Blankenburg, F., and 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., and 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.

Botvinick, M. and Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, 16(10):485–488.

Botvinick, M. M., Niv, Y., and Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280.

Busemeyer, J. R. and Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3):432–459.

Callaway, F., Gul, S., Krueger, P., Griffiths, T. L., and Lieder, F. (2018). Learning to select computations. In *Uncertainty in Artificial Intelligence: Proceedings of the Thirty-Fourth Conference*.

Callaway, F., Rangel, A., and 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., Lieder, F., and Griffiths, T. L. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, pages 1–14.

Camerer, C. and Ho, T. (2004). A Cognitive Hierarchy Model of Games*. *Quarterly Journal of Economics*, 119(3):861–898.

Caplin, A. and 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., and Ludwig, C. J. H. (2013). Adaptive sampling of information in perceptual decision-making. *PLoS ONE*, 8(11).

Cavanagh, J. F., Wiecki, T. V., Kochar, A., and Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, 143(4):1476–1488.

Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1):55–81.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410–418.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective and Behavioral Neuroscience*, 8(4):429–453.

Dayan, P. and Huys, Q. J. M. (2008). Serotonin, Inhibition, and Negative Mood. *PLOS Computational Biology*, 4(2):e4.

De Groot, A. D. (1965). *Thought and Choice in Chess*. De Gruyter Mouton, The Hague.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The Cost of Accumulating Evidence in Perceptual Decision Making. *Journal of Neuroscience*, 32(11):3612–3628.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, 11(5):1–36.

Fisher, G. (2017). An attentional drift diffusion model over binary-attribute choice. *Cognition*, 168:34–45.

Ford, J. K., Schmitt, N., Schechtman, S. L., Hults, B. M., and Doherty, M. L. (1989). Process tracing methods: Contributions, problems, and neglected research questions. *Organizational behavior and human decision processes*, 43(1):75–117.

Frömer, R., Dean Wolf, C. K., and 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.

Fudenberg, D., Strack, P., and Strzalecki, T. (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12):3651–3684.

Gabaix, X., Laibson, D., Moloche, G., and Weinberg, S. (2006). Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model. *American Economic Review*, 96 (4)(4):1043–1068.

Gershman, S. J., Horvitz, E. J., and 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. and Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, 62(1):451–482.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological review*, 103(4):650.

Gigerenzer, G. and Todd, P. M. (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press, USA.

Glaholt, M. G. and Reingold, E. M. (2009). Stimulus exposure and gaze bias: A further test of the gaze cascade model. *Attention, Perception & Psychophysics*, 71(3):445–450.

Gluth, S., Kern, N., Kortmann, M., and 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., and Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *eLife*, 7:e39659.

Gold, J. I. and Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36(2):299–308.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Gopnik, A. (1998). Explanation as Orgasm*. *Minds and Machines*, 8(1):101–118.

Gottlieb, J. and 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., and Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11).

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., and 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., and 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.

Gureckis, T. M. and Markant, D. B. (2012). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science*, 7(5):464–481.

Hay, N. (2016). Principles of Metalevel Control.

Hay, N., Russell, S., Tolpin, D., and Shimony, S. E. (2012). Selecting computations: Theory and applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pages 346–355, Arlington, Virginia, USA. AUAI Press.

Hébert, B. and Woodford, M. (2017). Rational inattention with sequential information sampling. *Working Paper*, pages 1–141.

Hebert, B. and Woodford, M. (2019). Rational inattention when decisions take time. *Journal of Chemical Information and Modeling*, 53(9):1689–1699.

Ho, M. K., Abel, D., Cohen, J., Littman, M., and Griffiths, T. (2020). The efficiency of human cognition reflects planned information processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1300–1307.

Holmes, W. R., Trueblood, J. S., and Heathcote, A. (2016). A new framework for modeling decisions about changing information: The Piecewise Linear Ballistic Accumulator model. *Cognitive Psychology*, 85:1–29.

Horvitz, E. J. (1987). Reasoning about beliefs and actions under computational resource constraints. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, pages 429–447.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26.

Howes, A., Lewis, R. L., and Vera, A. (2009). Rational Adaptation Under Task and Processing Constraints: Implications for Testing Theories of Cognition and Action. *Psychological Review*, 116(4):717–751.

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., and Kolling, N. (2021). Formalizing planning and information search in naturalistic decision-making. *Nature Neuroscience*, 24(8):1051–1064.

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F., and 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., and Dolan, R. J. (2016). Approach-Induced Biases in Human Information Sampling. *PLOS Biology*, 14(11):e2000638.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and 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., and 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. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306.

Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489–1506.

James, W. (1890). *The Principles of Psychology, Vol I.* The Principles of Psychology, Vol I. Henry Holt and Co, New York, NY, US.

Jang, A. I., Sharma, R., and Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *eLife*, 10:e63436.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604.

Just, M. A. and Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kahneman, D. and Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American psychologist*, 64(6):515.

Keramati, M., Dezfouli, A., and 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., and 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.

Knill, D. C. and Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press.

Kool, W. and Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, 2(12):899–908.

Kool, W., Gershman, S. J., and Cushman, F. A. (2017). Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, 28(9):1321–1333.

Krajbich, I. (2018). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, 29:6–11.

Krajbich, I., Armel, C., and 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., and Rangel, A. (2012). The Attentional Drift-Diffusion Model Extends to Simple Purchasing Decisions. *Frontiers in Psychology*, 3.

Krajbich, I. and 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., and 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., and Speekenbrink, M. (2018). Adaptive planning in human search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64.

Lewis, R. L., Howes, A., and Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 6(2):279–311.

Lieder, F. and Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, 124(6):762–794.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43.

Lohse, G. L. and Johnson, E. J. (1996). A Comparison of Two Process Tracing Methods for Choice Tasks. *Organizational Behavior and Human Decision Processes*, 68(1):28–43.

Ludwig, C. J. H. and Evens, D. R. (2017). Information Foraging for Perceptual Decisions. *Journal of Experimental Psychology. Human Perception and Performance*, 43(2):245–264.

MacGregor, J. N., Ormerod, T. C., and 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.

Manohar, S. G. and Husain, M. (2013). Attention as foraging for information and value. *Frontiers in Human Neuroscience*, 7(November):1–16.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: WH Freeman.

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. and Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21(11):1609–1617.

McFadden, D. (2001). Economic Choices. *American Economic Review*, 91(3):351–378.

McMillen, T. and Holmes, P. (2006). The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology*, 50(1):30–57.

Milosavljevic, M., Malmaud, J., and 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.

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., and Hayden, B. Y. (2020). Heuristics and optimal solutions to the breadth–depth dilemma. *Proceedings of the National Academy of Sciences*, 117(33):19799–19808.

Najemnik, J. and Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391.

Newell, A., Shaw, J. C., and Simon, H. A. (1959). Report on a general problem solving program. In *IFIP Congress*, volume 256, page 64. Pittsburgh, PA.

Newell, A. and 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.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154.

Noguchi, T. and Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological Review*, 125(4):512–544.

Norris, D. and Cutler, A. (2021). More why, less how: What we need from models of cognition. *Cognition*, 213:104688.

Oaksford, M. and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4):608–631.

O'Donoghue, T. and Rabin, M. (1999). Doing It Now or Later. *American Economic Review*, 89(1):103–124.

Ongchoco, J. D., Jara-Ettinger, J., and 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*, pages 904–910.

Orquin, J. L. and Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, 144(1):190–206.

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., and 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., and Johnson, E. J. (1993). *The Adaptive Decision Maker*. The Adaptive Decision Maker. Cambridge University Press, New York, NY, US.

Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, 8(9):095118.

Pirrone, A., Azab, H., Hayden, B. Y., Stafford, T., and 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.

Polanía, R., Krajbich, I., Grueschow, M., and Ruff, C. C. (2014). Neural Oscillations and Synchronization Differentially Support Evidence Accumulation in Perceptual and Value-Based Decision Making. *Neuron*, 82(3):709–720.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rahnev, D. and Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, 41:e223.

Ramírez-Ruiz, J. and 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. and McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural computation*, 20(4):873–922.

Ratcliff, R. and 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., and Mckoon, G. (2016). Diffusion Decision Model : Current Issues and History. *Trends in Cognitive Sciences*, 20(4):260–281.

Roe, R. M., Busemeyer, J. R., and Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108(2):370–392.

Russell, S. and Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence*, 49(1-3):361–395.

Russell, S. J. and Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ.

Russo, J. E. and Dosher, B. A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 9(4):676–696.

Savage, L. J. (1954). *The Foundations of Statistics*. The Foundations of Statistics. John Wiley & Sons, Oxford, England.

Schulte-Mecklenbeck, M., Kuehberger, A., and Johnson, J. G. (2011). Visiting the decision factory: Observing cognition with MouselabWEB and other information acquisition methods. In Schulte-Mecklenbeck, M., Kühberger, A., and Johnson, J. G., editors, *A Handbook of Process Tracing Methods for Decision Research*, pages 37–58. Psychology Press.

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., and De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *eLife*, 9:e60705.

Sezener, C. E., Dezfouli, A., and 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., and Cohen, J. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240.

Shi, S. W., Wedel, M., and 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., and Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12):1317–1322.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Sims, C. A. (1998). Stickiness. *Carnegie-Rochester Conference Series on Public Policy*, 49:317–356.

Smith, S. M. and Krajbich, I. (2018). Attention and choice across domains. *Journal of Experimental Psychology. General*, 147(12):1810–1826.

Smith, S. M. and Krajbich, I. (2019). Gaze Amplifies Value in Decision Making. *Psychological Science*, 30(1):116–128.

Snider, J., Lee, D., Poizner, H., and 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. and 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., and Botvinick, M. M. (2014). Optimal Behavioral Hierarchy. *PLOS Computational Biology*, 10(8):e1003779.

Song, M., Wang, X., Zhang, H., and 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. and Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization*, 25(3):309–327.

Stewart, N., Chater, N., and Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, 53(1):1–26.

Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J., and 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.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and 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*, pages 216–224.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

Tajima, S., Drugowitsch, J., Patel, N., and Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature Neuroscience*, 22(9):1503–1511.

Tajima, S., Drugowitsch, J., and Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, 7(1):12400.

Tavares, G., Perona, P., and Rangel, A. (2017). The attentional Drift Diffusion Model of simple perceptual decision-making. *Frontiers in Neuroscience*, 11(AUG):1–16.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640.

Teodorescu, A. R. and Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review*, 120(1):1–38.

Thomas, A. W., Molter, F., Krajbich, I., Heekeren, H. R., and Mohr, P. N. C. (2019). Gaze bias differences capture individual choice behaviour. *Nature Human Behaviour*, 3(6):625–635.

Todd, P. M. and Gigerenzer, G. (2003). Bounding rationality to the world. *Journal of economic psychology*, 24(2):143–165.

Tomov, M. S., Schulz, E., and Gershman, S. J. (2021). Multi-task reinforcement learning in humans. *Nature Human Behaviour*.

Towal, R. B., Mormann, M., and 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., and Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, 121(2):179–205.

Turner, B. M. and Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin and Review*, 21(2):227–250.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3):550–592.

Usher, M. and McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, 111(3):757–769.

van Opheusden, B., Acerbi, L., and 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., and Ma, W. J. (2017). A computational model for decision tree search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, US.

Wang, S., Feng, S. F., and 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., and Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, 367(6484):1362–1366.

Yang, L. C., Toubia, O., and 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.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. 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 TODO: url.