Rational representations of uncertainty: a pluralistic approach to bounded rationality

Isaac Davis

Department of Psychology, Yale University

Abstract

An increasingly prevalent approach to studying human cognition is to construe the mind as optimally allocating limited cognitive resources among cognitive processes. Under this *bounded rationality* approach (Icard 2018, Simon 1980), it is common to assume that resource-bounded cognitive agents approximate normative solutions to statistical inference problems, and that much of the bias and variability in human performance can be explained in terms of the approximation strategies we employ. In this paper, we argue that this approach restricts itself to an unnecessarily narrow scope of cognitive models, which limits its ability to explain how humans flexibly adapt their representations to novel environments. We argue that more attention should be paid to how we form our cognitive representations in the first place, and advocate for pluralistic framework which jointly optimizes over both representations *and* algorithms for manipulating them. We identify several fundamental trade-offs that manifest in this joint optimization, and draw on recent work to motivate a unified formal framework for this analysis. We illustrate a simplified version of this analysis with a case study in social cognition, and outline several new directions for research that this approach suggests.

1 Introduction

1.1 Background and motivation

Much of our everyday cognitive and perceptual activity involves inference under uncertainty: across many different contexts, we make judgments that are undetermined by the (often sparse and noisy) data available to us. *Are these pants black or dark blue? Will my friend enjoy this horror movie, or will it make them uncomfortable? Is this car merging into my lane, or did they leave their blinker on by mistake?* In the study of human cognition, it has become increasingly common to interpret our cognitive capacities through this lens, an approach known as *rational analysis* (Anderson 1990). Rational analysis is motivated by an assumption that the human mind has adapted to solve certain kinds of environmentally-grounded decision problems with limited information, and we can gain key insights into human cognition by precisely characterizing these problems and their optimal solutions.

Rational analysis is traditionally formulated at the computational level of analysis (Marr 1982), aiming to capture the formal structure of the inference problems we solve (i.e.: the information content of inputs and outputs), and comparing human performance against the optimal solutions to those problems. As such, rationalist models are typically posited as useful *descriptions* of human behavior, rather than genuine *explanations* of the neural and cognitive mechanisms underlying that behavior. More recent work, however, has sought to bridge this explanatory gap, extending the methodology of rational analysis to the algorithmic level of description. This approach, referred to as *boundedly rational analysis* (Icard 2018) or *resource-rational analysis* (Lieder & Griffiths 2020), explicitly accounts for the limited computational resources (e.g.: time and memory) with which the mind operates. Rather than modeling cognitive activity as (approximately) optimal inference under uncertainty, boundedly rational analysis models

cognitive activity as the (approximately) optimal allocation of cognitive resources. Recent work has leveraged this assumption to show how many of the apparent biases and errors that characterize human reasoning (Tversky & Kahneman 1974) actually reflect optimal performance under certain assumptions about the cost of computation (e.g.: Lieder et al 2012, Vul et al 2014, Lieder et al 2018).

1.2 Our contribution

In this paper, we argue that certain traditional approaches to boundedly rational cognitive modeling focus on an unnecessarily narrow scope of plausible cognitive models, and while they may provide a normative justification for *why* humans would produce certain patterns of behavior, they fall short of explaining *how* we develop these patterns of behavior. This approach is characterized by first defining a computation-level representation of a problem, deriving the optimal solution to that problem for an unbounded agent, then considering the optimal algorithm through which an agent with finite computational resources should approximate that solution. Underlying this approach are certain assumptions about how the agent can represent uncertainty and manipulate those representations: at a computational level, the optimal solution involves exact computation over explicit representations of uncertainty (e.g.: Bayesian posterior inference over a prior probability distribution- Griffiths et al 2008). At the algorithmic level, the most common assumption in the literature is that agents leverage sampling-based algorithms for approximating probabilistic computations (e.g.: Bonawitz et al 2014, Denison et al 2014).

We argue that the focus on approximating optimal solutions, and the particular focus on sampling-based approximations, is neither immediately demanded nor immediately justified by the assumptions of boundedly rational analysis. This is due in part to the fact that approximating a normatively ideal solution is not always the most rational

strategy for an agent with limited computational resources; in fact, there are cases in which approximation may be *less efficient* than exact computation. More generally, we argue that this approach glosses over another relevant dimension of optimization: the representations themselves. Given that there exist plausible neurophysiological accounts to support a range possible representations (Buesing et al 2011, Knill & Pouget 2004, Ma et al 2006, Moreno-Bote et al 2011), as well as preliminary behavioral evidence suggesting some flexibility in how we represent uncertainty (Vilares et al 2012, Houlsby et al 2013), we advocate for a more pluralistic approach to bounded rationality which optimizes over representational forms *and* algorithms for manipulating those representations.

We present two main arguments to advocate for this approach. Our first argument is posed at the scale of a single well-defined task: we will demonstrate in Section 3 that, even for a fairly restricted space of tasks, there may exist non-trivial interactions between how we represent uncertainty in key variables, and how we can efficiently manipulate that representation to solve the task. Thus, by fixing a single representation and adjusting our algorithm for manipulating it, we are only considering one half of the full optimization problem. Our second argument is posed at the more general (and realistic) context in which a bounded agent may encounter many different possible tasks, with varying degrees of uncertainty about which tasks will appear when. At this "zoomed out" perspective, it is clear that a bounded agent could not optimize for every possible task individually: the relevant question, then, is how the agent develops sufficiently flexible strategies and adapts them to novel contexts. As we will investigate in Section 4, this broader perspective makes the need for flexible representations especially clear, and highlights several fundamental trade-offs that have been under-studied in much of the literature on bounded rationality.

1.3 Outline

In the next section we provide more detail on the motivation and use of rational analysis, concerns about the explanatory capacity of rationalist cognitive models, and how boundedly rational analysis seeks to resolve these concerns. We then review recent work on boundedly rational cognitive modeling, how the scope of this work may be overly constrained, and what this implies about the rationalist justification of such models. Finally, we consider what the scope of our focus *ought* to be, and the requirements for an analysis framework that covers the appropriate scope.

In section three, we sketch out the requirements for such a framework, and point to some existing formal tools that are well-suited for the task. In particular, we show how Probabilistic Programming Languages (PPLs) provide a unified framework for formalizing both a space of representations, and a space of algorithms for manipulating those representations, in a way that exposes certain trade-offs relevant for our analysis. Further, we argue that parameterized complexity theory (Blokpoel et al 2010, Downey & Fellows 2012) provides a useful lens through which to characterize these trade-offs in a way that supports joint optimization. We then provide a simple case study to demonstrate that, even with a very restricted set of tasks, this optimization can be quite non-trivial. As we shall argue, however, a fully formalized framework for this joint optimization requires attention to certain trade-offs and constraints that go beyond the scope of traditional approaches. In section 4, we consider the higher-level problem faced by a bounded agent who must navigate a (potentially unknown) distribution of (potentially very different) tasks. We highlight three trade-offs that become especially salient in this context, and draw on recent work to motivate an optimization framework that unifies these trade-offs. We then argue that such a framework is better suited for understanding *how* people *actually* develop certain representations and strategies, which provides more explanatory power than a normative justification for *why* people *ought* to

use certain strategies. Finally, we conclude in section 5 with a brief summary of our findings, before considering future directions for research suggested by this approach.

2 Background

While not an entirely novel concept (Simon 1955, 1980), bounded rationality has seen a recent surge of interest in cognitive science and psychology, largely motivated by an apparent tension between two different bodies of psychological research. Here we provide more background on these two approaches to studying the mind, how bounded rationality seeks to resolve the tension between them, and the degree to which earlier approaches can fulfill this purpose.

2.1 Rational analysis

The study of human cognition faces a persistent identifiability problem. As we cannot directly observe or intervene on a subject's cognitive states, we generally have to rely on (often sparse and noisy) behavioral data to distinguish hypotheses. Furthermore, the space of hypotheses (i.e.: high-level cognitive models) is largely unbounded in the absence of any strong theoretical assumptions. Given the sparsity of available data streams, relative to the vast space of possible hypotheses, there will usually be many (sometimes infinitely many) competing explanations compatible with the same data (Pylyshyn 1980). Rational analysis seeks to address this problem by narrowing our focus: by modeling normative solutions to the problems being solved by the mind, we can both reduce the space of possible alternatives to consider, and provide more quantifiable metrics for comparing competing models (Anderson 1990). Thus, a rational analysis of cognitive behavior proceeds by identifying the problems solved by the mind, developing normative models of the ideal solutions to those problems, and comparing

human performance against those ideal solutions. Importantly, this approach is typically framed at the computational level of analysis (Marr 1982), aiming to characterize our cognitive behavior in terms of rational inferences while remaining agnostic about the cognitive or neural mechanisms underlying these inferences.

While there are multiple formalizations of rational analysis, the most prevalent by far is the Bayesian implementation (Chater & Oaksford 2007, Griffiths, Kemp, & Tenenbaum 2008). Although our arguments are aimed at rational analysis more generally, grounding these arguments in a particular implementation will help illustrate them more saliently, and we choose the Bayesian implementation due to its tremendous presence in modern cognitive science. Formally, we represent a cognitive agent's uncertainty about the world as a prior probability distribution $P(w)$ over possible world-states. Given some evidence E, a Bayesian agent (henceforth referred to as the *observer*) will update the degree to which they believe in world-state w according to Bayes' rule:

$$P(w|E) = \frac{P(E|w)P(w)}{P(E)} \tag{2.1}$$

where $P(w|E)$ is the agent's updated degree of belief in w (i.e.: the *posterior probability* of w), $P(E|w)$ denotes the probability of observing E given that w is the true state of the world (i.e.: the *likelihood* of E given w), $P(w)$ is the observer's prior degree of belief in w (before observing any evidence), and $P(E)$ is the total probability of observing evidence E. Faced with the problem of inferring the true hypothesis after observing evidence E, the most rational strategy[1] is to compute equation (2.1) for each possible world state, and return the state w^* which maximizes the posterior probability $P(w^*|E)$ (de Finetti 1937, Huttegger 2013). This provides a normatively ideal solution for inference under uncertainty and has therefore been widely used as a basis for rational analysis of human cognition.

[1]In the sense that no other strategy can outperform this strategy

The main concern when constructing a Bayesian cognitive model is how the observer represents the probability distributions in equation 2.1. The most common approach is to assume some internal mental model which specifies a set of variables (both observable and latent), and a set of probabilistic causal relations between these variables (i.e.: a *generative model*- Gerstenberg et al 2021, Icard 2016). To make this more concrete, we introduce a simple demonstration of such a model, which we will refer to throughout the rest of the paper as an illustrative example. To this end, imagine we are watching an agent navigate some environment (e.g.: a shopping mall). We observe the agent's first few steps $x_1, \ldots, x_{t-1}$, and wish to predict the agent's next step x_t. Figure 1a depicts such a task.

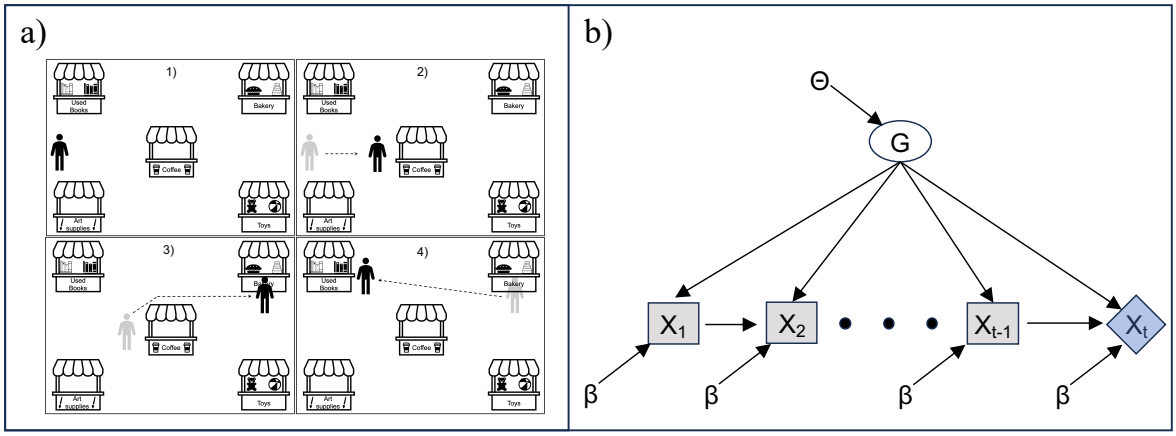

Figure 1: Illustration of an action-prediction task (panel a) and a simplified "rational planning" model for solving such a task (panel b). Panel a) depicts an agent navigating a shopping mall, where each sub-panel depicts the agent's next few steps. The observer's task is to predict the agent's next steps. Panel b) depicts a simplified "rational planning model" (Baker et al 2009) for solving such a task. Variables in grey boxes are observed (i.e.: the agent's previous steps). Variables in circles are posited latents (i.e.: the agent's goal state). Variables in blue diamonds are the targets of inference (i.e.: the agent's next step). Variables without borders are parameters that, together with the structure of the model, define a joint probability distribution over all variables in the model.

Behavioral inference tasks like these have been widely studied through the Bayesian framework. Figure 1b depicts a simplified version of a "rational planning model," a common generative model used to study this capacity (Baker et al 2009, Baker et al 2011). Under a rational planning model, we assume that the agent has some latent goal state G (e.g.: to acquire a certain type of item), where the parameter Θ captures the prior probability that the agent will have a particular goal (i.e.: the agent's preferences over different states of the world). At each step, the agent either moves along a shortest path to the goal state with probability $1 - \beta$, or moves in a random direction with probability β. This assumption provides the likelihood term $P(x_t|G, x_1, \ldots, x_{t-1}; \beta)$ (i.e.:

the probability of taking a particular action, given the agent's goal and prior actions), and the parameter Θ provides the prior distribution over goals $P(G;\Theta)$. Thus, this model encodes all of the information necessary to compute the posterior distribution in equation 2.1. This general approach- encoding a probability distribution in a generative model, and manipulating that model via Bayesian inference- has been used to study nearly every aspect of human cognition, including object perception and categorization (Kersten et al 2004, Salakhutdinov et al 2012, Stengård & Van den Berg 2019), language production and interpretation (Degen 2023, Goodman & Frank 2016), a range of intuitive theories such as physics (Smith & Vul 2013, Xu et al 2021) and psychology (Baker et al 2011, Jara-Ettinger et al 2016), social reasoning (Davis et al 2023; Gershman et al 2017), and the very process of cognitive development itself (Gopnik & Wellman 2012, Perfors et al 2011, Ullman & Tenenbaum 2020).

2.2 How rational are we, really?

Despite the success of rational analysis at accurately *describing* human inferences across a wide range of domains, there remain concerns (both theoretical and empirical) about the *explanatory* capacity of these models. The primary theoretical concern is tractability: outside of simple cases, the computations underlying optimal statistical analysis are generally intractable, in the sense that the amount of computation required increases exponentially (or worse) in the size of the input. In causal inference, for example, the number of possible causal structures over a set of variables increases exponentially in the number of variables, and exact Bayesian inference would require computing the posterior probability (equation 2.1) of each possible structure. Furthermore, many Bayesian cognitive models involve either continuous or infinitely recursive hypothesis spaces (e.g.: Griffiths & Ghahramani 2005), rendering exact inference completely infeasible. Thus, a rationalist explanation of human cognition must

address how we, as cognitive agents, perform these seemingly intractable computations quickly enough to make real-time decision (Jones & Love 2011).

On the empirical side, there are many cases in which the claims of rationality underlying this framework don't seem to manifest in human responses. Indeed, it is quite well established that human statistical judgments contain systematic errors and biases (Tversky & Kahneman 1974) that deviate from the predictions of Bayesian inference. For example, our estimations are often improperly biased or "anchored" towards numerical values we have previously considered, even when those values have no relation to the values we are estimating (Epley & Gilovich 2006), and we consistently over-weigh the probability of unlikely events with extreme consequences (Lichtenstein et al 1978). Furthermore, there is often a great deal of variability in human responses, both between and within individuals (Mozer et al 2008), which conflicts with the predictions of a rational Bayesian decision-maker. In particular, a rational Bayesian observer should always "posterior maximize," i.e.: deterministically choose the hypothesis with the highest posterior probability. In many cognitive studies, however, there is significant variability among participants' responses, and the overall empirical distribution of these responses tends to match the Bayesian posterior distribution, a phenomenon known as "posterior matching." While this may intuitively seem like an approximately rational strategy, it has been shown that posterior matching has (under a computation-level rational analysis) no rational justification, and should therefore *not* be interpreted as evidence that people are (approximately) rational (Eberhardt & Danks 2011).

2.3 A different kind of rationality

The bounded rationality program seeks to address both the theoretical and empirical challenges to rational analysis with a single conceptual reframing. Whereas computation-level rational analysis models agents with unbounded cognitive resources,

but limited information access, boundedly rational analysis explicitly considers the limited resources (e.g.: time, memory, etc.) to which the human mind has access. This new set of constraints introduces a fundamental trade-off: in general, more accurate solutions require more computation, which in turn makes them more costly. Thus, a boundedly-rational agent should weigh the benefit of having a more accurate solution against the increased cost of computing a more accurate solution, and allocate cognitive resources up to the point where the benefit of increased accuracy is outweighed by the cost.

This reframing has two benefits with respect to the aforementioned concerns. On the theoretical side, it helps alleviate concerns about intractability by suggesting that people are not actually performing intractable statistical inferences, but are instead approximating these computations in a more efficient way. Second, many of these approximation methods involve random sampling, often in a fashion that produces biased or auto-correlated outputs (e.g.: MCMC sampling). Thus, boundedly rational analysis aims to provide a normative justification for our apparently sub-normative biases and heuristics, by showing that they actually reflect a rational allocation of limited cognitive resources. Indeed, many apparent biases in human judgments have been shown to reflect the behavior of certain kinds of algorithms for approximating Bayesian inference: our tendency to anchor estimations to previously considered values, or to base our decisions on a small number of guesses, reflects the optimal behavior of certain kinds of sampling algorithms when generating additional samples is costly (Bonawitz et al 2014, Lieder et al 2012, Vul et al 2014). The over-weighting of unlikely events with extreme consequences reflects optimal sampling behavior for certain resource-constrained algorithms that approximate Bayesian inference (Lieder et al 2018). Posterior-*matching* can be more rational than posterior *maximizing* under certain constraints on memory (Icard 2021). Thus, the bounded rationality paradigm seems to provide a promising resolution to both

the theoretical and empirical concerns levied against the rational analysis framework.

2.4 The scope of boundedly rational cognitive models

The concerns and insights that motivated the bounded rationality paradigm suggest a certain intuitive approach to deriving boundedly-rational cognitive models. First, we define a problem (e.g.: predicting an agent's behavior) at the computation-level, and compute the optimal solution to that problem for an unbounded observer. In the Bayesian framework, this means defining a generative model of some relevant part of the world (e.g.: an agent's mental states and how those states causally relate to behavior), and using this model to compute a posterior distribution over possible answers. However, as these computations are typically intractable, an optimal *bounded* agent should approximate this posterior inference as well as is rational[2], given their cognitive resources. While intuitively appealing, this approach raises several concerns.

The first concern is that approximation does not always solve intractability: for many commonly occurring statistical inference problems, even approximate solutions (for any fixed degree of accuracy) cannot be tractably computed in general (Kwisthout et al 2011). Even when approximation does enable a tractable solution, it may not necessarily yield the boundedly optimal solution: in some cases, non-Bayesian heuristics can outperform approximate Bayesian inference with the same limited resources (Icard 2018). Thus, even if approximate Bayesian inference is tractable, there is no guarantee that such an algorithm is actually resource-rational without assuming substantial restrictions on the observer's space of plausible algorithms, and the space of representations over which those algorithms operate.

This leads to a second, more general concern about these models: how do we

[2]i.e.: up to the point where the cost of additional computation exceeds the benefit of additional accuracy

determine the appropriate set of cognitive constraints, including how the agent represents uncertainty for a given problem, and the set of algorithms through which the agent can manipulate those representations? If our assumptions are too general or minimal, we risk glossing over important factors that can influence the true "cost" of a solution. For example, one approach is to assert that the agent has a method for drawing unbiased, independent samples from the relevant posterior distribution at a fixed cost per sample, while remaining agnostic about the details of the sampling process itself (e.g.: Bonawitz et al 2014, Vul et al 2014). However, generating unbiased, independent samples from a posterior distribution often requires just as much computation as exact posterior inference (or else requires a prohibitive number of random decisions to approximate). Thus, glossing over the details of the sampling process in this fashion makes it difficult to assess how well this approach can inform a resource-rational understanding of human cognition.

On the other hand, if our assumptions are too strong, we risk omitting other possible representations or algorithms that may be more efficient. Another approach, for example, is to assume the observer uses a fixed approximation algorithm (e.g.: some form of MCMC sampling), and evaluate the optimal use of that algorithm (e.g.: the optimal number of samples to draw) for a particular task representation (e.g.: Dasgupta et al 2017, Lieder et al 2012, Milli et al 2021). While this exposes the relevant details of the sampling process, these details are only applicable if humans do, in fact, use algorithms with the same properties as those assumed in the model. This assumption is complicated, however, by the fact that there are many approximation methods than can be implemented with the same core machinery posited by these models. For example, given the cognitive machinery required to implement a Metropolis-Hastings algorithm (a form of MCMC algorithm- Robert & Casella 1999), one could implement a range of alternate algorithms for approximating the same distribution, including various forms of

exact inference, rejection sampling, particle filters, and other MCMC algorithms. This fact makes it difficult to assert that one particular approximation algorithm is *the* right one to use in an algorithmic-level cognitive model.

In response to these concerns, some have proposed a different approach to understanding how resource-bounded agents could solve these seemingly intractable problems. In particular, this approach suggests that, rather than finding efficient algorithms for approximating intractable computations, perhaps the mind simply forms representations for which tractable solutions already exist (e.g.: Correa et al 2023, Kwisthout et al, 2011, Tomov et al 2020). The notion of simplifying representations- rather than using approximate algorithms over exact representations- is not novel to machine learning or computer science. In variational methods, for example, an intractable probabilistic computation is made tractable by approximating the true distribution with a family of simpler distributions (e.g.: by assuming independence between variables- Sanborn 2017). Similarly, certain algorithms for probabilistic graph inference impose sparsity constraints on the inferred graphs (i.e.: restricting the number of connections between variables) as a means of trading off representational accuracy for computational efficiency (Bishop 2006, ch 8).

More recently, some authors have suggested that the mind may employ similar tricks to form tractable mental representations, motivated in part by insights from parameterized complexity theory (Downey & Fellows 2012). This approach to complexity theory aims to break down the computational cost of solving a problem into different "dimensions" that characterize the structure of the problem. That is, suppose we have a class M of intractable problems, and we identify a set of parameters of interest $K = \{k_1, \ldots, k_n\}$ which characterize individual instances of problems within this class (e.g.: the number of latent variables in the problem, the number of values that each variable can assume, etc.). Given these parameters of interest, the aim of parameterized

complexity analysis is to determine whether it is possible to solve problems in M efficiently when the values of these parameters are held fixed, even as the size of the input increases arbitrarily. If this is the case, then the parameters in K are said to be the *source of intractability* for M, and M is said to be *fp-tractable for* K. It has been shown, for example, that a common class of Bayesian inference problem which is generally intractable is fp-tractable for two parameters- the maximum number of latent variables in the network, and the degree of certainty of the most probable configuration of latent states (Blokpoel et al 2011).

2.5 Moving forward

The previous section suggests two distinct conceptual approaches to boundedly-rational cognitive modeling. The first starts with a computation-level representation of a problem (i.e.: a particular generative model), and considers the resource-optimal algorithm for manipulating that representation (either exactly or approximately). The second approach starts with a description of a *task*, and considers what kinds of representations enable tractable solutions. This distinction between positing a "task" and positing a "representation" is a subtle but important one. Consider, for example, the task illustrated in Figure 1a. If we characterize the task as one of "goal inference," as is common in the Bayesian Theory of Mind literature, this entails certain assumptions and restrictions on how an observer represents the task (i.e.: it assumes that the observer explicitly represents a latent goal state for the agent). On the other hand, if we simply characterize the task in terms of the relevant inputs (the agent's environment and previous behavior) and outputs (the agent's next action), this imposes fewer assumptions on the observer's representation, and leaves more flexibility for the observer to "optimize" their representation for that particular task (i.e.: form a representation for which tractable solutions exist in that context, rather than using the same kind of

representation across all contexts in which that problem occurs).

In the remainder of this paper, we advocate for a unified framework that jointly optimizes over the representations we employ for solving a task, *and* the algorithms we use to manipulate those representations. We present two arguments for such a framework, each posed at a different conceptual "scale." In the next section, we will sketch out a formal framework for performing this joint optimization at the scale of individual tasks. We will then argue that, even for a fixed task (or a very restricted space of similar tasks), there may be important interactions between the structure of our representations and the cost of manipulating these representations via particular algorithms. Furthermore, by drawing on insights from parameterized complexity analysis, we can see that the cost of using two different algorithms might not "scale up" along the same dimensions. That is, if we have a set K of parameters that characterize a space of representations, and two different algorithms A_1 and A_2 for manipulating those representations, there may be cases (as we shall demonstrate in section 3.3) where parameter $k_1 \in K$ causes intractability in A_1, while a different parameter $k_2 \in K$ causes intractability in A_2. Thus, choosing an optimal algorithm depends on the structure of our representations, but choosing the optimal representation also depends on the set of algorithms we can use to manipulate it. This creates a bi-directional interaction between the choice of how to represent uncertainty in a particular context, and the optimal choice of algorithm for manipulating that representation. Thus, both dimensions of optimization can be important even at the scale of individual tasks.

To make the importance of this joint optimization especially salient, however, we must "zoom out" from the scale of individual, pre-known tasks, and consider the much more plausible context in which a bounded agent faces a (possibly unknown) distribution of (possibly very different) tasks. In reality, people face uncertainty across many different kinds of tasks in their daily lives, as well as uncertainty about which tasks they will

encounter when, and the exact nature of future tasks. Even within the scope of a single, routine workday, we might have to reason about several different people whom we know to varying degrees, across different contexts, each with its own set of social expectations demanding that we draw different kinds of inferences (e.g.: "how do I write this report in a way my boss will approve of?" "how do I decline my coworker's invitation without hurting their feelings?" "how do I avoid colliding with this stranger who isn't paying attention while they walk?"). It would almost certainly be infeasible for a bounded agent to develop separate, resource-rational solutions for each individual task they might encounter. In section 4, we will consider resource-rational analysis within the context of this broader problem and identify three additional trade-offs that become especially relevant at this scale of analysis. Although these trade-offs are relatively understudied in the literature, we will point to some recent work that explores these constraints, and sketch out how they might be incorporated into a unified theory of bounded rationality. Finally, we argue that such a framework could not only provide normative justification for *why* a bounded agent *ought* to employ certain strategies, but explain *how* a bounded could *actually* develop such strategies from complex, dynamic, and highly uncertain environments.

3 Framework sketch

The analysis we outline in the previous section requires two core components. The first is a unified framework for formalizing both a space of possible representations of a task, and a space of possible algorithms for manipulating these representations. As we show in the following section, Probabilistic Programming Languages (PPLs) are particularly well suited for this purpose (Goodman 2013), and are compatible with the assumptions underlying much of the current literature on bounded rationality. The second component

is a methodology for computing a cost profile for each algorithm as a function of the representation to which it is applied. Drawing on parameterized complexity theory, the aim is to identify a set of relevant dimensions that characterize the different representations within this space, and compute the cost of each algorithm in terms of these dimensions. This will enable a joint optimization over representations *and* algorithms. As we shall argue in section 4, a full analysis would require dimensions of comparison that are under-studied in the current literature (e.g.: expectations over future data streams). However, we provide a case study in section 3.3 which demonstrates that, even when we restrict our analysis to single task, this optimality analysis can still be fairly non-trivial.

3.1 Probabilistic programming languages and generative models

A probabilistic programming language (PPL) extends a deterministic programming language with a set of stochastic primitive functions. For example, we can define a stochastic primitive $flip(w)$ that returns a 1 with probability w, or a 0 with probability $1 - w$, and a function $roll(n)$ which returns an integer between 1 and n uniformly at random. We can derive more complex functions from stochastic primitives via composition and recursion. For example, the program below simulates flipping a coin with bias w, then rolling a three-sided die if the flip comes up heads, or a six-sided die if the flip comes up tails:[3]

[3]For these examples, we use a condensed, intuitive psueodocode based on WebPPL, a probabilistic programming language for generative models (Goodman et al 2016)

```
flip_and_roll(w){
    f = flip(w)
    if (f == 1) {return roll(3)} else {return roll(6)}
}
```

Note that, as a probabilistic program, repeated calls to $flip\_and\_roll(w)$ with the same input value will result in a distribution of different output values. However, the PPL contains an operator that enables analytic computations of these probabilities as well: for a stochastic primitive function f and a value x in its range, the operator $Prob(f, x)$ returns the probability that f will output x. Thus, given a stochastic primitive function f, we can analytically compute the distribution it encodes by applying $Prob(f, x)$ to each x in its range, or we can approximate this distribution by running f repeatedly on the same input and tabulating the frequency of each output. These two basic operators enable a range of methods for computing or approximating more complex distributions. For a purely analytical computation, we can enumerate each possible execution history of the program and multiply the probabilities associated with each primitive decision (see Figure 2a for an example), while a purely stochastic approximation simply requires repeatedly running the program and tabulating its outputs. This also enables a range of intermediate algorithms, by applying the analytical operator $Prob(f, x)$ to certain primitive random decisions, while approximating other random decisions via sampling.

While this enables a range of algorithms for computing (or approximating) a distribution over *outputs*, most inference problems of interest involve further manipulation of this distribution. Suppose, for example, that a program involves some set X of random variables, including a subset $E \subset X$ that we get to observe (e.g.: an agent's behavior), and a subset $Q \subset X$ that we don't get to observe (e.g.: an agent's

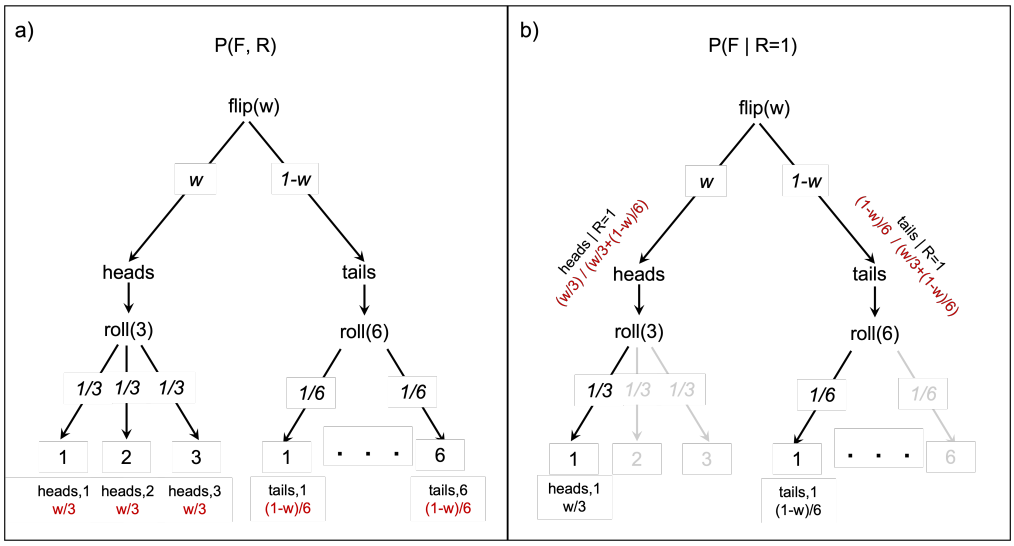

Figure 2: Diagram of procedure for computing the distributions implied by probabilistic programs. Panel a) depicts the procedure for computing the full joint distribution over all of model variables. Panel b) depicts the procedure for computing a conditional distribution, given the observed value of one model variable. The probabilities corresponding to the target distribution are highlighted in red.

mental states). In order to reason about the value of the latent variables Q given the observations E, we need access to the posterior distribution $P(Q|E)$. With the *flip_and_roll* function, for example, suppose we observe that the die roll R resulted in a value of 1, and we must infer the result of the initial coin-flip F. At a computational level, this corresponds to computing the posterior distribution $P(F|R = 1)$. At the algorithmic level, the basic operators of the PPL enable multiple ways to compute or approximate the target distribution.

To compute the distribution analytically, we can enumerate each possible execution history of the program, multiplying the weights of the primitive distribution at each random choice, and omitting any execution history which violates the observations (in this case, any history with a die roll not equal to 1). The conditional probability that the coin flip was heads, given that the die roll was 1, is equal to the total probability mass of all (non-excluded) executions where the coin flip was heads, divided by the total

probability mass of all executions where the die roll was 1 (see Figure 2b). This procedure allows the system to analytically compute any conditional distribution from any probabilistic program (with finite and discrete outputs).

At the other extreme, we can generate unbiased samples from the posterior distribution $P(Q|E)$ using a straightforward technique called rejection sampling: we simply run the program repeatedly until it outputs a value that matches the observation (in this case, until it results in a die roll of 1), then return the value of the query variable (in this case, the coin flip) associated with the final execution. Note that the analytic procedure uses only deterministic computation, but involves no random decisions, while rejection sampling may involve many random decisions, but involves no deterministic computation. We can therefore interpret these two algorithms as opposite endpoints of a spectrum, with fully deterministic algorithms at one extreme and fully stochastic algorithms on the the other. Between these two endpoints lie a range of intermediate methods that can be implemented using the same core machinery, including particle filters and various Markov Chain Monte Carlo (MCMC) methods. These methods use a mix of both random decisions and deterministic computation to generate (usually biased and autocorrelated) samples from the target distribution $P(Q|E)$. This trade-off between deterministic computation and random decisions is just one of several possible dimensions relevant to our analysis, but we shall focus on this dimension for our case study in section 3.3, to demonstrate that this trade-off alone entails a non-trivial optimization problem, even within a fairly restricted problem space.

3.2 Rational representations of uncertainty

The previous section demonstrates how PPLs can simultaneously encode a particular representation (i.e.: generative model) of a problem, and a set of algorithms for manipulating that representation. It is clear, however, that given a set of stochastic

primitives and arbitrary recursion, we can define a rich space of possible representations for the same problem. We therefore need some systematic way of comparing these possible representations. One way of evaluating potential representations is to contrast "task-dependent" or "opportunistic" representations, which only represent uncertainty in the decision variable itself, with "constitutive" representations, which explicitly represent not just the decision variable, but a slew of latent variables thought to be involved in the causal process that generates the decision variable (Koblinger et al 2021). Consider our earlier example, in which we observe an agent's initial movements $x_1, \ldots, x_{t-1}$ through some environment W, and must then predict the agent's next move x_t. A fully task-dependent representation for this task would maintain a probability distribution over x_t as a direct function of the input variables $W, x_1, \ldots, x_{t-1}$ (Figure 3a). Intuitively, this representation encodes an assumption that the agent follows some fixed "script," such that the probability of the next action x_t is directly implied by the previous actions.

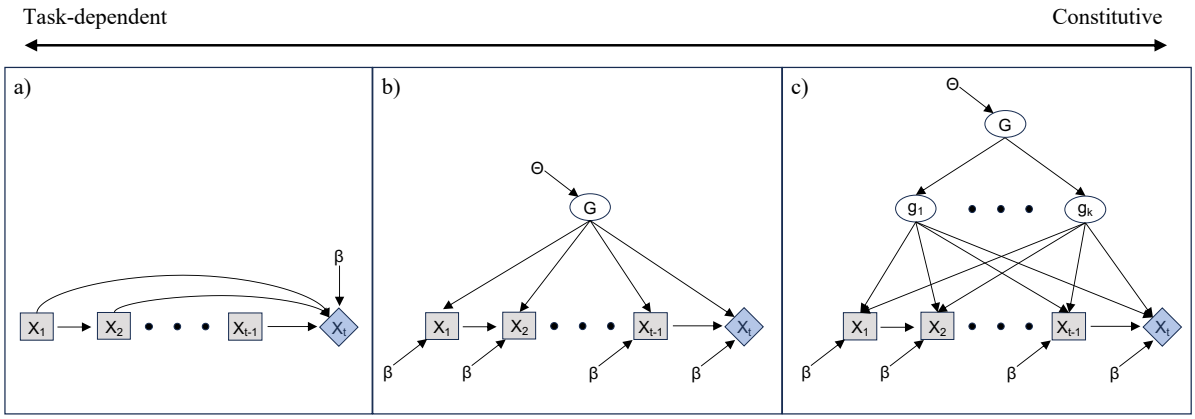

Figure 3: Examples of possible representations for an action-prediction task. Variables shaded in grey are observed (in this case, the agent's previous actions). Variable shaded in blue is the target of prediction (agent's next action). Variables in ovals are latent states (e.g.: goals). Variables without borders are parameters that encode the relevant probability distributions (e.g.: probability of taking an action given previous action and goal). Panel a) depicts a fully task-dependent model that only represents uncertainty in the target variable. Panel b) depicts a simple mentalistic model that posits a single latent goal state. Panel c) depicts a more complex mentalistic model that posits a hierarchical goal state with planning over sub-goals.

A constitutive representation, on the other hand, explicitly represents the latent mental states that are thought to cause the agent's behavior. For example, Figure 3c shows a complex goal model that posits a high-level goal G (e.g.: a certain recipe), which entails a set of sub-goals $g_1, \ldots, g_k$ that are required to fulfill the high-level goal G (e.g.: a list of ingredients), which in turn determine the agent's plan (e.g.: path through the grocery store that meets all the required ingredients). Of course, this distinction between task-dependent and constitutive representations is a graded, rather than binary notion, and we might consider a range of intermediate representations that include certain latent states but omit others. For example, Figure 3b depicts an intermediate representation

that posits a single static goal or preference state. Intuitively, we might interpret this as encoding the assumption that the agent is following one of several possible "scripts," and the single goal variable encodes which script the agent is executing (e.g.: Davis & Jara-Ettinger 2022).

Given these different representations of the same task, how do we evaluate and compare them? One obvious dimension is the cost of manipulation: in general, richer and more constitutive representations are costlier to manipulate and compute than simpler, more task-dependent representations, though the exact rate at which these costs scale depends on the nature of the algorithms being used (e.g.: the number of deterministic computations versus random decisions required). On the other hand, task-dependent representations tend to be highly inflexible, requiring a distinct representation for each possible task, even within a similar context, while constitutive representations enable much greater generalization and flexibility (Koblinger et al 2021). In our action prediction example, although we could predict an agent's behavior by memorizing a set of scripts that they tend to follow, it is likely that those scripts would vary widely across contexts (e.g.: the scripts one follows in the grocery store are unlikely to be the same scripts one follows at an airport). On the other hand, if we represent the mental states that cause the agent's actions, we can generalize those mental states across contexts (e.g.: knowing that the agent likes hamburgers improves our ability to predict the agent's behavior in both a grocery store and an airport). Other potentially relevant dimensions for comparison include the memory cost of storing the representation (in particular, how many independent parameters must be stored), and the amount of data required to effectively learn the representation. The current literature on bounded rationality has paid considerably less attention to these last three factors- generalizability, memory requirements, and learnability- focusing primarily on computation time. We will return to this in section 4, when we consider how to expand

the current scope of boundedly rational cognitive modeling to enable a proper optimality analysis over representations and algorithms.

3.3 Case study: action prediction

We now provide a simple case study to motivate what this analysis might look like. As described in the previous sections, a fully formalized framework for this joint optimization is beyond the scope of this paper, in part because it will necessarily involve dimensions of comparison that have been under-explored in the current literature (see section 4). Thus, rather than a full demonstration, our aim here is to show that, even with a fairly constrained and simplified problem, and even with only a single dimension of comparison (deterministic computation versus random decision-making), this optimization can still be quite non-trivial, and may entail different optimal solutions even for different instantiations of the same basic task. To this end, we return to our action prediction example, where we observe an agent's environment W and first few steps $\bar{x} = x_1, \ldots, x_{t-1}$, and wish to predict the agent's next step x_t. Of course, there is a wide range of generative models that one could use to represent this task, and a wide range of algorithms one could use to manipulate these representations. We will restrict our analysis to the candidate models shown in Figure 3, and the two algorithms described in section 3.1: enumeration, which only involves deterministic computation and no random decisions, and rejection sampling, which only involves random decisions and no deterministic computation.

We shall start with the fully task-dependent model in Figure 3a: under this model, the probability that the agent will take action x_t is a direct function of the agent's previous steps and parameter β: $P(x_t|\bar{x};\beta)$. Importantly, we are assuming that the observer has already learned the representation and relevant parameters, so the cost of acquiring the representation is not currently a factor (though we will return to this in

section 4), and we only consider the cost of manipulating this representation to solve the task. For the fully task-dependent representation, the cost of manipulation is quite low-in fact, there is barely any computation required. For an analytic solution, the posterior probability $P(x_t|\bar{x};\beta)$ for any x_t is already stored in the parameter vector β, so computing this probability exactly involves a single step (essentially, looking up the corresponding probability). Similarly, we can obtain an unbiased sample from $P(x_t|\bar{x};\beta)$ with a single random decision. Thus, the fully task-dependent representation enables an extremely efficient solution that only requires a single computation or random decision. Of course, this is not the full story: in order to enable such efficient computation, we must store a significant number of independent parameter values in β (essentially, one vector for each possible sequence of previous actions $x_1, \ldots, x_{t-1}$). Furthermore, this representation is specific to one particular environment, so our knowledge of these parameters is unlikely to be of any use in a slightly different context.

On the other end of the spectrum, we shall now consider the cost of manipulating the complex goal model (Figure 3c) for solving this task.[4] For an unbounded observer with this representation, predicting the agent's next action requires marginalizing out the latent goal variables, i.e.:

$$P(x_t|\bar{x}) = \sum_{g_1,\ldots,g_k} P(x_t|x_{t-1}, g_1, \ldots, g_k)P(g_1, \ldots, g_k|\bar{x})$$

Intuitively, this requires computing, for each combination of sub-goals $g_1, \ldots, g_k$, the probability that the agent would take action x_t, given those sub-goals, weighted by the posterior probability that the agent has those sub-goals, given their prior behavior. The term $P(x_t|x_{t-1}, g_1, \ldots, g_k)$ is already encoded into the model by the parameter β, so the

[4]We omit the analysis for the intermediate model in Figure 3b, as this is essentially a special case of Figure 3c where the number of sub-goals k is fixed to 1

bulk of the work is to compute $P(g_1, \ldots, g_k|\bar{x})$ for each possible combination of sub-goals. Thus, analytically computing $P(x_t|\bar{x})$ for a single value of x_t involves M^k total computations, where M is the number of values that each sub-goal variable g_i can assume, and k is the number of possible sub-goals. The cost of the fully deterministic solution using this representation therefore grows exponentially in the number of sub-goals k.

Now we contrast this against the cost of obtaining a single unbiased sample of $P(x_t|\bar{x})$ via rejection sampling, which does not involve any deterministic computation, and only involves random decisions. Recall that rejection sampling involves repeatedly running the generative model "forward" until we obtain a sample for which the observable variables $x_1, \ldots, x_{t-1}$ have the same values that we have observed in $\bar{x}$, then observing the value of the target variable x_t. A single forward run of this model involves first sampling the sub-goal vector $g_1, \ldots, g_k$ from the goal prior Θ, then simulating the agent's behavior for t steps. Of course, we cannot precisely compute the number of samples required before we obtain one that matches our evidence. However, it is straightforward to compute that, on average, we should expect to generate $1/P(\bar{x})$ samples before we obtain one that matches, where $P(\bar{x})$ is the overall probability of the evidence $\bar{x}$. Thus, obtaining a single unbiased sample of $P(x_t|\bar{x})$ via rejection sampling requires an average of $1/P(\bar{x})$ samples from the forward model, each of which involves $k + t$ random decisions. Of course, this is not the full story, as the analytic solution is always guaranteed to provide the correct answer, while rejection sampling only gives us a single unbiased sample from the predictive distribution $P(x_t|\bar{x})$. The full analysis would require some measure for the value of accuracy- that is, how much does it cost to get the wrong answer?

Although we cannot properly establish which strategy is optimal without this extra piece of information, the present analysis is still sufficient to draw some useful conclusions: first, while the cost of analytically computing $P(x_t|\bar{x})$ grows exponentially

in the number of sub-goal states k, the cost of generating an unbiased sample from $P(x_t|\bar{x})$ is only polynomial in the number of sub-goals. On the other hand, the cost of an unbiased sample grows exponentially in the surprisal of the evidence (i.e.: $-log(P(\bar{x}))$, while the cost of an analytic solution is completely independent of the probabilities. Thus, even within this fairly restricted example, we see an important trade-off emerge: as our representations become richer and involve a larger number of interconnected latent variables, the cost of analytic computations grows exponentially, while the cost of generating unbiased samples only grows in polynomial time. On the other hand, as the surprisal of the evidence increases, the cost of generating an unbiased sample grows exponentially, while the cost of an analytic solution remains constant. This suggests that deeper and richer representations may be most efficiently manipulated via algorithms that rely more heavily on random decisions, while flatter and simpler representations enable more efficient analytic solutions.

4 What's missing?

The previous section outlines the general requirements for the sort of analysis framework we advocate, and demonstrates that, even with a simplified toy problem, there are non-trivial interactions between how we represent the uncertainty in the problem (e.g.: the latent depth of the generative model) and the optimal way to manipulate that representation (e.g.: via deterministic computation or unbiased sampling). However, it should be clear that the analysis in this simple demonstration is not the full story. After all, our analysis showed that the cost of computing the solution via the fully task-dependent representation is constant, so we might expect a resource-rational observer to always form fully task-dependent representations. It is well established, however, that humans regularly represent uncertainty beyond the decision variable alone

(Dension et al 2018, Houlsby et al 2013). Indeed, the analysis in the previous section ignores several other factors that are clearly relevant, such as the amount of memory required to store a task-dependent representation (which may require a very large number of independent parameters), or the time it takes to learn the relevant parameters (which may require a very large amount of data). The importance of these factors becomes especially salient when we consider the broader problem of a bounded agent navigating a complex, dynamic environment involving many different types of tasks, especially if there is uncertainty about which tasks will appear when. In this context, it is almost certainly infeasible for the agent to develop optimized, task-specific representations for each task they might encounter.

In this section, we highlight three additional constraints that become relevant in the context of this more general problem. We motivate the trade-offs that these constraints introduce to a bounded rationality analysis, and point to some recent work that explores how bounded cognitive agents might plausibly manage these trade-offs. We conclude the section by motivating how these constraints might be integrated into a unified framework for resource-rational analysis, and how such a framework could provide greater explanatory power than earlier approaches.

4.1 Memory

While much of the existing literature on resource-rationality has focused on computation time, another important cognitive constraint is memory. Within the context of a single task, we can interpret this constraint in terms of the number of independent parameter values that must be maintained in order to store a representation. In general, we can trade off computation time for memory by storing the output of a particular computation as a fixed parameter value. On an intuitive level, the fully task-dependent representation is only able to achieve such computational efficiency by making an

extreme trade-off: rather than relying on internal computations to predict the probability of an action given previous actions, this representation stores all of these probabilities as fixed parameter values. Thus, in order to maintain this representation for a particular task, one must store a separate, independent parameter vector for each possible state-sequence $x_1, \ldots, x_{t-1}$. In some contexts, this might be feasible: suppose, for example, that we are watching an agent navigate a very small grocery store with only three stands, and we know that we will only ever observe the agent in this particular environment. In this case, it may be perfectly tractable to simply memorize the agent's most frequent trajectories through the store, and avoid having to compute anything about the agent's internal states.

Outside of a highly constrained environment, however, this strategy would likely impose prohibitive memory requirements, as the observer would need to memorize a large list of parameters for each individual context (e.g.: for every different environment we observe the agent traverse). This highlights the first major benefit of a more constitutive representation: by leveraging a richer representation of the latent variables and processes that generate observable behavior, we can drastically reduce the number of independent parameter values that must be stored. For example, the complex goal model in Figure 3c requires a single parameter vector for the agent's goals (e.g.: the agent's preferences over possible goal states), and a single parameter that captures the agent's degree of "noisiness" when executing a plan (e.g.: how deterministically they follow the most efficient path). Thus, richer and more constitutive representations can drastically decrease the memory requirements for storing a representation at the expense of increased computation costs for manipulating the representation. On the other hand, if a particular computation is especially time-consuming, it may be beneficial to store the output of that computation as its own parameter, thus bypassing the need to recompute it in the future. In computer science, this technique is known as "memoization," and can

significantly improve the efficiency of probabilistic computations (Pfeffer 2007).

When we zoom out from the context of a fixed task, memory constraints introduce a second type of trade-off relevant to the bounded cognitive agent. In particular, rather than learning a completely new representation whenever a novel type of task is encountered, an agent could store previously generated representations in memory and draw on those as a "starting point" for new tasks. The effectiveness of this strategy depends on the generalizability of the stored representations (see section 4.3): a representation that is only useful in a very limited context may not be worth the extra storage space to retain. On the other hand, overly general representations might be of limited use in any one specific context, or require significant additional adaptation to utilize in a specific context. Some recent work explores how bounded cognitive agents could effectively manage this trade-off through a kind of "representational caching" (e.g.: Dechter et al 2013, Zhao et al 2023). In these frameworks, a bounded agent generates new candidate representations by sampling them from a probabilistic grammar, then evaluates how well that representation supports a particular task (or set of tasks). As new representations are generated, they may, with some probability dependent on their performance, be stored in memory and recalled in subsequent iterations. These "adaptor grammars" (Johnson et al 2006) thereby enable a bounded agent to trade-off memory (i.e.: storing additional rules and representations in a grammar) for computation time (i.e.: bypassing the need to regenerate and recompute those representations), and recent experimental work has shown that these models can replicate certain order-effects and other sub-optimalities that manifest in human concept learning (e.g.: Fränken et al 2022, Zhao et al 2023).

4.2 Learnability

A second constraint that was glossed over in section 3 is learnability. In the example from 3.3, we assumed that the observer already had access to fully parameterized versions of each representation. Thus, this analysis ignored the cost of *learning* these representations in the first place (i.e.: inferring the values of relevant parameters). In reality, however, the cost of learning a representation may be a significant factor in deciding how to represent uncertainty in a task. In general, richer and more constitutive representations can be learned more quickly, and often from less data, than fully-task dependent representations, a phenomenon sometimes referred to as the "blessing of abstraction" (Goodman et al 2011). There are two high-level reasons for this difference: first, in a fully task-dependent representation, the only data relevant for learning the representation are data observed in same context being represented. For example, if we represent an agent's trajectories through an environment as an explicit distribution over trajectories, the only data useful for inferring that distribution are observations of the trajectories themselves. On the other hand, if we also represent the agent's mental states, we can integrate information from across multiple contexts to learn the relevant parameter values. For example, if we explicitly represent an agent's preferences (rather than a direct distribution over trajectories), we can leverage data from multiple contexts (e.g.: any context in which the agent makes a choice of what to eat) to infer the parameter values that capture the agent's preferences.

A second difference is that task-dependent representations require fully labelled data (e.g.: observation of the agent's full trajectory), while constitutive representations can leverage unlabelled data as well (Koblinger et al 2021). For example, suppose we observe the agent go to the dairy counter, then take two more steps in another direction, but we don't get to observe the rest of the trip. Using a mentalistic model of the agent's behavior, an observer could infer a posterior distribution over the agent's possible goals,

based on the partial trajectory, then use that inferred goal distribution to predict the probability of each possible next step. The observer could then update the parameters in the model by averaging over all possible completions of the trajectory, weighted by the posterior probability of that trajectory. Intuitively, this means that the observer can use the latent causal processes encoded in the representation to simulate the missing portion of the data, and use that simulated data to perform additional learning. Thus, even though task-dependent representations are more efficient to manipulate, they generally require more data, more specific data, and more labelled data in order to learn, compared to constitutive representations. This shows that choosing the optimal representation may depend in part on our expectations about the availability and cost of future data. Furthermore, a learner may have some influence over which evidence they observe when, which they can potentially exploit to improve the efficiency of learning. In causal learning, for example, a learner may be able to choose which interventions to apply. Given the set of causal hypotheses the learner is currently considering, certain interventions will provide more decisive evidence than others. Recent work on "active learning" shows that people are able to identify informative interventions in causal learning tasks (e.g.: Coenen et al 2015, Bramley et al 2017), suggesting another method that bounded cognitive agents might employ to manage this learnability trade-off.

In addition to the availability of new evidence, representation learning is also constrained by the computational cost of incorporating new evidence into the representation. In general, suppose our representation is parameterized by some parameter vector Θ. In order to learn the relevant parameter values from some evidence E, a Bayesian observer must compute the posterior probability $P(\Theta|E) \propto P(E|\Theta)P(\Theta)$, which typically requires computing the likelihood function $P(E|\Theta)$ (i.e.: the probability of observing evidence E, given the true parameter values Θ). For complex representations with many latent states, computing this likelihood is often intractable,

thus imposing a restrictive cost on the process of incorporating new evidence into the observer's representation, independent of the cost of *obtaining* that evidence. Thus, although constitutive models can typically be learned from less data, they may be be costlier to update from that data. Some recent work draws on tools from Bayesian statistics to explore how a bounded agent could tractably incorporate new evidence into a complex representation. These models leverage "likelihood-free inference," a class of methods for approximating Bayesian inference without directly computing the likelihood function, instead relying on summary statistics that are easier to compute (e.g.: Gutmann & Corander 2016). Similar approximation methods have been used to develop algorithmic-level models of causal and physical learning in from temporal data, which replicate certain patterns of suboptimality in human inference (Gong & Bramley 2023, Ullman et al 2018).

4.3 Generalizability

Perhaps the most crucial factor that the task-specific perspective from section 3 failed to highlight is generalizability. While a task-dependent representation can enable efficient solutions for one particular context, it is ill-suited for generalizing beyond that context. For example, memorizing an agent's most frequent behavior trajectories in a particular environment will enable fast, efficient prediction of the agent's behavior in that environment. However, this strategy would require learning and memorizing a whole new set of trajectories for every new environment (and agent) we observe. On the other hand, by representing the latent states that cause the agent's behavior, we can generalize to new contexts much more effectively: if, instead of memorizing the agent's behavior, we represent the agent's *preferences* and use our model to compute the agent's behavior, we can utilize our representation across a much wider range of contexts, even if it requires somewhat more computation within each context. Thus, we must manage another

trade-off: do we develop a larger set of more task-specific representations, or a smaller set of richer and more flexible representations? Managing this trade-off requires careful attention to the distribution of tasks and domains that we expect to face, even moreso than the previously discussed constraints.

One potential approach for managing this trade-off is to model the broader learning problem directly: rather than considering a set of problems in which we learn individual representations for solving individual tasks, we consider the unified problem of learning a collection of representations for one or more domains of tasks simultaneously. Hierarchical Bayesian models have proven tremendously useful for this type of learning, enabling an observer to simultaneously learn a set of representations at multiple levels of abstraction from data collected across multiple domains. For example, Goodman et al (2011) derive a model that simultaneously learns a set of causal models for specific domains, and a higher-order "theory" of causality that constrains the lower-level models. In a similar vein, Kemp & Tenenbaum (2008) propose a hierarchical model of "structure-learning," which simultaneously learns a set of latent data structures from data collected across multiple domains, and a higher order distribution over latent structure types that constrains the lower-level representations. More recent work has explored how bounded cognitive agents could efficiently learn a dynamic library of representations for solving a range of different tasks (e.g.: Bramley et al 2023, Zhao et al 2023), suggesting another potential strategy for managing this representational trade-off.

4.4 Putting it all together: the value of representation

While a full specification of a framework that unifies all of these trade-offs is beyond the scope of this paper, we can motivate how this might be done by looking toward some recent work. In a standard approach to resource rational analysis, we consider some task t, and a set $\mathcal{A}$ of possible algorithms for solving the task. These algorithms may differ in

both the reward they yield if applied to this task (e.g.: how accurately or consistently they produce the right answer), which we can denote by $R(a,t)$ for $a \in \mathcal{A}$, as well as the cost of implementing the algorithm in the task, which we can denote by $C(a,t)$. The overall utility derived from applying algorithm a to task t is thus $U(a,t) = R(a,t) - C(a,t)$, and the resource-optimal algorithm for solving the task is defined as $a^* = argmax_a U(a,t)$ (Lieder & Griffiths 2020). We can then generalize this to account for potential uncertainty about the distribution of tasks we will face. If we have some belief about the set T of possible tasks, and the probability $P(t)$ that we will face a certain task $t \in T$, we can compute the *expected* utility of applying algorithm a as

$$EU_P(a) = \sum_t U(a,t)P(t) \tag{4.1}$$

and define the optimal algorithm as $a^* = argmax_a EU_P(a)$.

In this context, our proposal requires generalizing this optimization to a space $\mathcal{R}$ of possible representations for tasks in T, *and* a set $\mathcal{A}$ of algorithms for manipulating those representations. For a fixed representation $r \in \mathcal{R}$, we can extend the above definitions to $R(a,r,t)$, $C(a,r,t)$, and $U(a,r,t)$, to respectively define the reward, cost, and overall utility of applying algorithm a to representation r for task t. We can then define the overall utility of representation r for solving task t as $U(r,t) = max_a U(a,r,t)$ (i.e.: the utility obtained by applying the best algorithm for that representation). Similarly, for a given distribution over tasks $P(t)$, we can define the expected utility of a representation for solving those tasks as

$$EU_P(r) = \sum_t U(r,t)P(t) \tag{4.2}$$

If $P(t)$ is highly concentrated on a small set of similar tasks, it may be more efficient to use a cheaper, task-dependent representation. On the other hand, if $P(t)$ has non-trivial support over a large set of different tasks, we may require a richer but more generalizable

representation to adequately solve those tasks.

Importantly, the standard cost function $C(a, r, t)$ only captures the cost of an implementation- that is, the cost of applying a to r to solve a single instance of t. However, two of the "costs" described in the previous section apply outside the scope of a particular implementation: the cost of maintaining the representation r in memory (i.e.: the number of independent parameter values required), and the cost of learning the representation (i.e.: inferring the values of those parameters). These costs are specific to the representation itself, and are independent of the algorithm used to manipulate that representation, or the task(s) for which the representation applies. Thus, when computing the expected utility of the representation, these costs would appear outside the scope of the expectation operator, i.e.:

$$EU_P(r) = \left(\sum_t U(r,t)P(t) \right) - C_{mem}(r) - C_{learn}(r) \tag{4.3}$$

The normative solution to the bounded agent's optimization problem is then given by the representation (or set of representations) that optimizes this expected utility.

While this motivates how the standard bounded rationality framework could be generalized to optimize over representations, there remain two significant conceptual challenges to this account. The first is that computing this expected utility requires full knowledge of the space of possible tasks we may encounter, and the probability that we will encounter a given task. In reality, however, this assumption rarely holds. Indeed, one of the main motivations for this line of research is to explain how humans can so effectively adapt to completely novel or unexpected contexts, and how we can quickly identify efficient (if not globally optimal) representations for solving those tasks. Thus, the first challenge is how this approach can account for the "unknown unknowns:" we can't always know what we don't know about future possibilities. A second remaining

challenge is how we efficiently navigate the massive space of possible representations (or worse, the space of possible *libraries* of representations). After all, the motivation behind this approach to cognitive modeling is the observation that humans have limited cognitive resources. If the normative solution to this problem requires yet another massively intractable optimization, then what does this account really explain? To put this problem another way: if the mind really is an "adaptive toolbox" of heuristics that we can flexibly combine and adapt to novel contexts (Gigerenzer & Todd 1999), then the overall value of any given toolbox should be captured by something like equation 4.3. But if this optimization is itself intractable, how does a bounded cognitive agent develop such a toolbox, and how does that agent flexibly combine and adapt their tools to novel contexts?

Some recent work seeks to resolve this tension by showing how we can leverage general purpose algorithms to dynamically generate a toolbox of useful representations in an efficient way (e.g.: Bramley et al 2017; Bramley et al 2023; Dasgupta et al 2017; Franken et al 2022; Zhao et al 2023). These accounts incorporate many of the strategies described in the previous three sections to manage the key trade-offs we have identified thus far. In particular, they leverage representational caching to store previously used representations in memory for efficient re-use in future tasks; they leverage structured, dynamic priors (e.g.: adaptor grammars) to learn an ensemble of representations for multiple contexts simultaneously; they make use of local sampling methods (e.g.: MCMC) to incrementally adapt existing representations for new contexts; and they employ multiple strategies to allow for efficient parameter learning, such as adaptively selected interventions to generate informative evidence, and efficient approximations for incorporating that evidence into our representations.

These accounts therefore take a crucial step beyond the scope of early research into bounded rationality. In particular, earlier work aimed to show *why* human reasoning is

biased in certain systematic ways, by arguing that these biases and heuristics are optimal in another sense- they reflect optimally efficient approximations of normatively ideal solutions (e.g.: Lieder & Goodman 2012; Parpart et al 2018; Vul et al 2014). As we have argued, however, the lack of clarity about the proper scope of this analysis left it unclear as to *how* humans develop these biases, especially given that the computations required to derive these optimal approximations are themselves largely intractable. However, by drawing on a much more general set of representational and algorithmic tools, this new line of research, makes valuable progress towards answering the crucial "how" question.

5 Discussion and future work

Bounded rationality is a promising research program that seeks to resolve a longstanding tension in cognitive science and psychology. On the one hand, the rational analysis paradigm has proven a tremendously useful tool for studying how humans deal with uncertainty. Across a wide range of contexts, human judgments seem to reflect (approximately) rational statistical inference, and rationalist cognitive models have been used to provide computation-level accounts for nearly every aspect of human cognition (Griffiths et al 2008, Oaksford & Chater 2007). On the other hand, a computation-level rational analysis does not explain how people are able to perform these seemingly intractable computations, nor does it explain the seemingly sub-rational biases and errors we systematically display across a wide range of inference tasks (Tversky & Kahneman 1974, Epley & Gilovich 2006, Lichenstein et al 1978, Mozer et al 2008). The bounded rationality paradigm seeks to resolve this tension by considering the limited cognitive resources with which real-world human minds operate, and justifying our apparently sub-rational biases as the rational allocation of limited resources.

However, it is difficult to determine the appropriate scope of focus for boundedly

rational cognitive models. Early work in the field aimed to characterize an inference problem and its optimal solution at the computational-level, then consider algorithms for tractably approximating that solution. As we and others have argued, however, this approach is neither immediately demanded nor immediately justified by the assumptions of bounded rationality. First, approximation does not, in general, make intractable problems tractable: in many cases, approximate solutions can be just as prohibitively expensive as exact solutions (Kwisthout et al 2011). Even in cases where approximation *is* tractable, there is no general guarantee that approximating the optimal solution is more rational or efficient than some other context-specific heuristic (Icard 2018). Furthermore, there are many different ways that an agent could represent uncertainty in a given task, and many different algorithms for manipulating those representations, all of which could be implemented using the same cognitive machinery typically assumed by these models. This issue is especially salient when we zoom out from the perspective of a single task, and consider the broader problem of navigating a complex, dynamic environment with uncertainty about the nature of future tasks. Thus, we argue that this approach may unnecessarily limit our search for plausible cognitive models, and in doing so, limits its usefulness as a genuine explanation of human cognition. In short, we argue that traditional approaches may help us justify *why* humans *should* use the heuristics they do, but falls short of explaining *how* humans *actually* develop these heuristics given limited cognitive resources.

For this reason, we advocate for a more pluralistic approach to boundedly rational cognitive modeling, where we consider the representational and computational primitives to which an agent has access, and optimize over the full space of representations and manipulations that could be implemented with those primitives. In section 3, we demonstrated how, even with a fairly simple and restricted task space, there are non-trivial interactions between the way we represent uncertainty (e.g.: the richness of

the latent structure encoded in our representations) and cost of manipulating those representations via different algorithms (e.g.: via exact enumeration or unbiased sampling). In section 4, we considered the higher level problem of optimizing for multiple tasks across multiple (potentially unknown) domains, and identified three additional constraints that, while highly relevant in this context, have been relatively understudied in the early bounded rationality literature. We then pointed to recent work that has begun to take these additional constraints more seriously, described the progress they have made towards understanding how bounded cognitive agents can develop and adapt their representations for novel contexts, and motivated what a universal framework for balancing these trade-offs might look like.

The arguments we presented point to some important future work, both theoretical and empirical. On the theoretical side, our notion of "value of representation" will require some additional work to fully formalize and implement. In particular, integrating memory constraints into a unified notion of cost may be challenging. This is partially due to the fact that memory constraints are spatial in nature, while computational constraints are typically calculated in terms of time. Furthermore, memory can impose different kinds of constraints depending on the type of representations being used. When maximizing a posterior distribution analytically, for example, even though the agent must compute the posterior probability for each possible answer, they need only *remember* one possible answer at a time- if a new answer is determined to have a higher posterior probability than the previously remembered answer, the agent is free to "forget" the previous answer and only retain the new one. On the other hand, if the agent is, say, approximating the distribution with a set of samples, then the agent's memory limitation will directly constrain the number of samples the agent can retain, and thus the accuracy of the approximation. In future work, it will be important to consider how memory limitations fit into this broader analysis framework, and some recent work has already

begun to investigate this issue (e.g.: Patel et al 2020). Additionally, the cost of *learning* a representation may be highly dynamic and depend on our knowledge or expectations about the distribution of environments we will face, and the availability (and cost) of relevant data in those environments. Finally, while recent work with adaptor grammars has shown how a bounded agent can learn a library of representations through local, incremental changes, these incremental changes will likely affect the cost of manipulation via different algorithms. For example, as our representations become incrementally deeper (i.e.: more latent variables), they become incrementally costlier to manipulate analytically. Thus, future work may explore how to augment these frameworks so that they can incrementally develop both a library of representations, *and* a library of algorithms for manipulating those representations in the context of specific tasks. This would further our theoretical understanding of how a bounded agent could tractably learn to navigate a highly dynamic, complex, potentially unknown environment.

On the empirical side, this pluralistic modeling approach suggests several directions for future research. First, although there is some behavioral and neural evidence that people can represent uncertainty in multiple ways (e.g.: Denison et al 2018, Houlsby et al 2013), there is little empirical work that explores how flexibly people can *adjust* their representations in response to specific task demands or environments. Several recent papers have highlighted the importance of measuring how experimental subjects spontaneously represent experimental stimuli (e.g.: Davis 2021, Szollosi et al 2023), and Koblinger et al (2021) outline a general approach to behavioral studies into the task-specificity of people's cognitive representations. A similar approach could be leveraged to investigate the flexibility of those representations across different task environments. This also suggests several experimental manipulations worth investigating through these frameworks, such as participants' expectations about the availability of future data, or the distribution of environments in which they will need to make

judgments.

Furthermore, these insights may lead to novel predictions about when we expect people to rely on sampling-based approximations versus exact computation. The case study we presented in section 3.3 suggests such a study: if, as current work suggests, the variability in human responses reflects an underlying sampling process, then an agent who solves a problem exactly should display significantly less variance in their responses than an agent who approximates a solution via sampling. We can therefore leverage this principle to derive testable hypotheses about how people manipulate different representations of uncertainty. Some work has already demonstrated that people can be motivated to make more accurate inferences with less variability by increasing the potential payout of a correct answer (Vul et al 2014) or increasing the noisiness of a stimulus (Hamrick et al 2015). Given the principle that richer representations are more costly to manipulate analytically, it should also be the case that increasing certain parameters of an inference problem (e.g.: the number of possible hidden states or the likelihood of an observation) beyond a certain threshold should induce a switch from exact computation to approximation, or vice versa. Such a switch would be characterized by a sharp increase or decrease in response variability as a problem moves above or below one of these thresholds. Thus, theoretical development of this framework will both necessitate and generate a plethora of novel behavioral studies.

References

[1] Anderson, J. R. (1990). The Adaptive Character of Thought. *Psychology Press.*

[2] Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, No. 33).

[3] Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329-349.

[4] Bishop, C. (2006). Pattern recognition and machine learning. *Springer publishing*.

[5] Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). *Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference*. Cognitive psychology, 74, 35-65.

[6] Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, 124(3), 301.

[7] Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (2023). Local search and the evolution of world models. *Topics in Cognitive Science*.

[8] Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS computational biology*, 7(11), e1002211.

[9] Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology*, 79, 102-133.

[10] Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., & Griffiths, T. L. (2023). Humans decompose tasks by trading off utility and computational cost. *PLOS Computational Biology*, 19(6), e1011087.

[11] Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from?. *Cognitive psychology*, 96, 1-25.

[12] Davis, I. (2021). How do we know what babies know? The limits of inferring cognitive representations from visual fixation data. *Philosophical Psychology*, 34(2), 182-209.

[13] Davis, I., Carlson, R., Dunham, Y., & Jara-Ettinger, J. (2023). Identifying social partners through indirect prosociality: a computational account. *Cognition*, 240, 105580.

[14] Davis, I., & Jara-Ettinger, J. (2022). Hierarchical task knowledge constrains and simplifies action understanding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44 (44)

[15] Dechter, E., Malmaud, J., Adams, R. P., & Tenenbaum, J. B. (2013). Bootstrap learning via modular concept discovery. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press/International Joint Conferences on Artificial Intelligence.

[16] De Finetti, B. (1937). Le prevision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincare* (Vol. 7, No. 1, pp. 1-68).

[17] Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9, 519-540.

[18] Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, 115(43), 11090-11095.

[19] Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, 126(2), 285-300.

[20] Downey, R. G., & Fellows, M. R. (2012). Parameterized complexity. Springer Science & Business Media.

[21] Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines*, 21(3), 389-410.

[22] Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological science*, 17(4), 311-318.

[23] Fränken, J. P., Theodoropoulos, N. C., & Bramley, N. R. (2022). Algorithms of adaptation in inductive inference. *Cognitive Psychology*, 137, 101506.

[24] Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science*, 41, 545-575.

[25] Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*.

[26] Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In Simple heuristics that make us smart (pp. 3-34). *Oxford University Press*.

[27] Gong, T., & Bramley, N. R. (2023). Continuous time causal structure induction with prevention and generation. *Cognition*, 240, 105530.

[28] Goodman, N. D. (2013). The principles and practice of probabilistic programming. *ACM SIGPLAN Notices, 48(1)*, 399-402.

[29] Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11), 818-8

[30] Goodman, N.D., Tenenbaum, J.B., & The ProbMods Contributors (2016). Probabilistic Models of Cognition (2nd ed.) https://probmods.org/

[31] Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological review*, 118(1), 110.

[32] Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6), 1085.

[33] Griffiths, T. L., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. In NIPS (Vol. 18, pp. 475-482).

[34] Griffiths, T. L., Kemp, C., & Tenenbaum, J.B. (2008). Bayesian models of cognition. In *The Cambridge handbook of computational psychology* (R. Sun, ed.), ch. 3, 59-100, Cambridge University Press.

[35] Gutmann, M. U., & Cor, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125), 1-47.

[36] Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the 37th annual conference of the Cognitive Science society*.

[37] Houlsby, N. M., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M., & Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. *Current Biology*, 23(21), 2169-2175.

[38] Huttegger, S. M. (2013). In defense of reflection. *Philosophy of Science*, 80(3), 413-433.

[39] Icard, T. (2016). *Subjective probability as sampling propensity.* Review of Philosophy and Psychology, 7(4), 863-903.

[40] Icard, T. (2018). Bayes, bounds, and rational analysis. *Philosophy of Science*, 85(1), 79-101.

[41] Icard, T. (2021). Why be random?. *Mind*, 130(517), 111-139.

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

[43] Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and brain sciences, 34(4), 169.

[44] Johnson, M., Griffiths, T., & Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in neural information processing systems*, 19.

[45] Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31), 10687-10692.

[46] Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.*, 55, 271-304.

[47] Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12), 712-719.

[48] Koblinger, Á., Fiser, J., & Lengyel, M. (2021). Representations of uncertainty: where art thou?. *Current Opinion in Behavioral Sciences*, 38, 150-162.

[49] Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cogn. Sci.*, 35(5), 779-784.

[50] Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of experimental psychology: Human learning and memory*, 4(6), 551.

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

[52] Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, 125(1), 1.

[53] Lieder, F., Griffiths, T., & Goodman, N. (2012). *Burn-in, bias, and the rationality of anchoring*. Advances in neural information processing systems, 25, 2690-2798.

[54] Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11), 1432-1438.

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

[56] Milli, S., Lieder, F., & Griffiths, T. L. (2021). A rational reinterpretation of dual-process theories.

[57] Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30), 12491-12496.

[58] Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? Cognitive science, 32(7), 1133-1147.

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

[60] Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(1), 111-132.

[61] Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive psychology*, 102, 127-144.*Cognition*, 217, 104881.

[62] Patel, N., Acerbi, L., & Pouget, A. (2020). Dynamic allocation of limited memory resources in reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 16948-16960.

[63] Pfeffer, A. (2007, July). Sampling with memoization. In *AAAI* (Vol. 7, pp. 1263-1270).

[64] Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, 120(3), 302-321.

[65] Robert, C. P., & Casella, G. (1999). The Metropolis-Hastings Algorithm. In Monte Carlo Statistical Methods (pp. 231-283). Springer, New York, NY.

[66] Salakhutdinov, R., Tenenbaum, J., & Torralba, A. (2012, June). One-shot learning with a hierarchical nonparametric bayesian model. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* (pp. 195-206). JMLR Workshop and Conference Proceedings.

[67] Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and cognition*, 112, 98-101.

[68] Simon, H. A. (1980). Bounded rationality. In *Utility and probability* (pp. 15-18). Palgrave Macmillan, London.

[69] Simon, H. (1955). A behavioral model of bounded rationality. Quarterly Journal of Economics, 69(1), 99-118.

[70] Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, 5(1), 185-199.

[71] Stengård, E., & Van den Berg, R. (2019). Imperfect Bayesian inference in visual perception. *PLoS computational biology*, 15(4), e1006465.

[72] Szollosi, A., Donkin, C., & Newell, B. R. (2023). Toward nonprobabilistic explanations of learning and decision-making. *Psychological Review*, 130(2), 546.

[73] Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLoS computational biology*, 16(4), e1007594.

[74] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185(4157), 1124-1131.

[75] Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive psychology*, 104, 57-82.

[76] Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2, 533-558.

[77] Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology*, 22(18), 1641-1648.

[78] Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive science*, 38(4), 599-637.

[79] Xu, K., Srivastava, A., Gutfreund, D., Sosa, F., Ullman, T., Tenenbaum, J., & Sutton, C. (2021). A bayesian-symbolic approach to reasoning and learning in intuitive physics. *Advances in Neural Information Processing Systems*, 34, 2478-2490.

[80] Zhao, B., Lucas, C. G., & Bramley, N. R. (2023). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, 1-12.

Acknowledgments

We are very grateful to David Danks, Yarrow Dunham, Julian Jara-Ettinger, as well as the editor and anonymous reviewers for their invaluable feedback on this manuscript.