Bayes, Bounds, and Rational Analysis

(To appear in *Philosophy of Science*)

Thomas F. Icard

July 4, 2017

Abstract While Bayesian models have been applied to an impressive range of high-level cognitive phenomena in recent years, methodological challenges have been leveled concerning their particular use in cognitive science, and specifically their role in the program of rational analysis. The focus of the present article is on one strand of these criticisms, namely, computational impediments to probabilistic inference, and related puzzles about empirical confirmation of these models. The proposal is to rethink the role of Bayesian methods in rational analysis, and specifically to adopt an independently motivated notion of rationality appropriate for computationally bounded agents, and then to explore broad conditions under which (approximately) Bayesian agents would in fact be rational. The proposal is illustrated with a characterization of computational costs in an abstract manner inspired by ideas in thermodynamics and information theory.

1 Introduction

Many normative questions—questions about what we ought to do or think—obviously depend on various factual questions about what the world is like. In particular, answers to these questions may depend on details about what kinds of cognitive agents we are and how our minds in fact work. In the other direction, there is an equally powerful idea, that one productive method for discovering how our minds in fact work is by asking questions and developing hypotheses about what would be good ways for the mind to work, that is, how the mind ought to work. Versions of this thesis can be traced back at least to the American pragmatists, who particularly stressed the adaptation of mind to an environment, and

many variations of the strategy have been pursued in psychology (e.g., Green and Swets 1966; Peterson and Beach 1967; Marr 1982; etc.). Much contemporary research in cognitive science takes inspiration from a formulation due to John Anderson, based on a methodological program called *rational analysis* (Anderson, 1990). A rational analysis of some cognitive function involves conjecturing a precise problem that the mind is assumed to be solving, finding a good or even optimal solution to this problem, and using that solution to guide the scientist's search for cognitive models of how that function works.

The rational analysis program has been applied across a number of domains, from low-level perception to high-level cognitive tasks such as categorization and language understanding. Because many such tasks can be construed as problems of inference under uncertainty, it has been common to characterize them as Bayesian inference problems, with the problem specified in terms of a prior probability model to be updated with some data or observations, and the optimal solution given by application of Bayes' Rule. Bayesian cognitive science has seen a remarkable surge of interest over the past several decades, and the approach has been applied to high-level phenomena as diverse as naïve physics and linguistic pragmatics (see Tenenbaum et al. 2011 for a review).

At the same time, a number of critical articles have questioned some of the methodological assumptions underlying Bayesian rational analysis (Murphy, 1993; Kwisthout et al., 2008; Jones and Love, 2011; Eberhardt and Danks, 2011; Bowers and Davis, 2012; Marcus and Davis, 2013). Some of the criticisms are directed at rational analysis generally, echoing similar criticisms to optimality modeling in biology. Others are aimed specifically at the claim that we should expect Bayesian theory to be the default framework for understanding rationality or optimality when it comes to studying human cognition. One central theme in this critical work is that the calculations required by most Bayesian models are intractable and thus could not describe computations performed by a resource-limited brain. If one instead considers approximations to Bayesian models, which in many cases provide better fit to individual-level data anyway (cf. the discussion in §5 below), then the claim of rationality or optimality comes into question, since tractable approximations can often be arbitrarily less accurate than what is prescribed by the ideal model. Thus, the original motivation for rational analysis—to help guide the search for cognitive models—is undermined, since there is nothing obviously distinguishing an approximation to a Bayesian model from any other non-Bayesian model, at least from a rational point of view.

The aim of the present article is to propose a different way of thinking about Bayesian models in cognitive science and their relation to the rational analysis

program. In short, rather than construing the problems to be solved by the mind as problems of inference under uncertainty, we should instead think of cognitive functions as solving constraint optimization problems, where the relevant constraints include not (only) making accurate inferences, but making rational use of limited computational resources.[1] On this version of *boundedly rational analysis*, it cannot be assumed from the start that a given cognitive function should be modeled in a Bayesian manner. However, it does leave open the possibility that in many cases Bayesian models, or approximations thereto, will nonetheless be promising hypotheses even from the boundedly-rational constraint-optimization perspective. This shift in perspective is arguably consonant with more recent trends in Bayesian cognitive modeling, but it also makes definite prescriptions, and raises pressing and precise questions about when we should expect a Bayesian (approximation) model to be apt.

After presenting the view in more detail and sketching a very general account of bounded (instrumental) rationality, we shall consider this question of when approximate Bayesian computations, specifically Monte Carlo algorithms for Bayesian inference, could be seen as a boundedly rational solution to an underlying constraint optimization problem. We discuss and assess several possibilities, including a speculative proposal inspired by thermodynamics, which characterizes the problem to be solved in terms of an abstract notion of computational cost derived from the entropy of an agent's action distribution (the probabilities of taking various possible actions). Whether or not the overall proposal ultimately succeeds in characterizing the tasks to which human cognition might be adapted, it will be argued that some such account should be given in order to motivate Bayesian models and their approximations as a priori plausible candidates in a rational analysis.

2 Bayesian Rational Analysis

The study of high-level cognition—phenomena such as categorization, causal reasoning, moral judgment, language understanding and production, etc.—relies largely on behavioral data. There are methods that allow more fine-grained data to be collected than mere input-output pairs, e.g., eye-tracking, response-time studies, brain lesion analyses, etc. However, even with these more probing techniques, there remains a problem of *identifiability* (Pylyshyn, 1984; Anderson, 1990). The

[1] The present article develops a proposal made earlier in Icard (2013, 2014).

problem is two-fold. First, there is an obvious question of where to begin the search for high-level cognitive models. Second, there are often cases in which competing models cannot yet be distinguished by available measurement. Rational analysis promises to ameliorate both of these problems by focusing the researcher's attention on models that involve a good, or even optimal, way of carrying out the function. Anderson (1990, 29) summarizes the methodology in six steps: (1) Precisely specify the goals of the cognitive system. (2) Develop a formal model of the environment to which the system is adapted. (3) Make the minimal assumptions about computational limitations. (4) Derive the optimal behavioral function given items 1 through 3. (5) Examine the empirical literature to see if the predictions are confirmed. (6) If the predictions are off, iterate. He then goes on to demonstrate the efficacy of the approach in four cases studies on memory, categorization, causal inference, and problem solving. In the intervening decades rational analyses have been applied to scores of phenomena (Chater and Oaksford 1999; Oaksford and Chater 2007; Tenenbaum et al. 2011; Lewis et al. 2014).

Rational analysis is similar in spirit to the use of optimal models in evolutionary biology, in that both involve an optimality assumption. In the biological case, the assumption is often used to support a kind of adaptationist explanation: the organism has this observed feature *because* it optimal (Parker and Maynard Smith, 1990). The status of rational analysis as a method for deriving *explanations* of cognitive phenomena is a matter of considerable controversy (Danks, 2008; Jones and Love, 2011; Eberhardt and Danks, 2011; Bowers and Davis, 2012). However, the primary use of the methodology in cognitive science—and the one that is our focus in this paper—is to address the identifiability problem, that is, to uncover reasonable models of cognition that make sense and have a priori plausibility, especially in cases where empirical evidence provides inadequate hints about where to start. Even critics of optimality modeling in biology are sympathetic to this kind of motivation (Kitcher 1987). The question is how we should theorize about human rationality in order to arrive at plausible cognitive models in the first place.

Across a number of domains of cognition, it often seems sensible to characterize cognitive functions in terms of inference problems under uncertainty. The idea that even unconscious mental operations could be understood as inferences goes back at least to Helmholtz, who construed vision as the problem of inferring latent causes of sensory impressions. The framing nicely generalizes: categorization involves inducing categories for novel objects in order to predict unobserved features (e.g., Anderson 1990), language understanding involves inferring a speaker's intended message from the utterance produced (e.g., Frank and Goodman 2016), and so on. Furthermore, inference and prediction make sense as abilities that

would be good for an organism, and would thus seem eligible for improvement and perhaps even optimization through adaptation.[2]

Bayesian rational analysis is premised on two assumptions. The first is that many cognitive functions can indeed be understood as solving inductive inference problems. The second is that Bayesian methods provide a rational approach to such problems. The Bayesian approach to statistical inference can be understood as incorporating the following broad claims:

- Uncertainty can and should be characterized in terms of a probability distribution $P(\vartheta)$ over some relevant class of possible states of affairs, together with appropriate likelihood functions $P(s|\vartheta)$ for any conceivable stimuli s, specifying how the alternative hypotheses would be expected to produce observations.

- Upon receiving some stimuli s, one ought to update one's view about ϑ by conditionalizing and using Bayes' Rule, so that $P(\vartheta|s) \propto P(s|\vartheta)P(\vartheta)$. That is, the posterior probability of ϑ is proportional to the likelihood times the prior on ϑ.

This broad characterization by no means picks out a unique statistical methodology,[3] but it is enough to distinguish the Bayesian approach to cognitive science. For a given cognitive task, the Bayesian cognitive scientist will posit a reasonable prior probability distribution $P(\vartheta)$ assumed to capture participants' prior expectations (perhaps guided in part by empirical elicitation), as well as a likelihood function $P(s|\vartheta)$ specifying what one would expect to observe were each of the hypotheses true (which may also be determined partly empirically). Once these two elements have been specified, the scientist hypothesizes that participants' responses in the experimental task can be captured in some way in terms of the posterior distribution $P(\vartheta|s)$.

What exactly is supposed to be rational about these models? There are at least three aspects that might be characterized as rational, or specifically as Bayesian:

1. that people are behaving as though they are performing Bayesian inference,

2. the priors and likelihoods themselves may be rational,

3. people are making rational decisions on the basis of their inferences.

[2] See Geisler and Diehl (2002) for a detailed account of how optimal inference in perceptual systems could evolve.

[3] Famously, Good (1971) calculates that there at least 46,656 different versions of Bayesianism.

While it is typically presumed that people make reasonable decisions on the basis of their updated view of the situation (e.g., predicting the hypothesis with highest posterior probability), and that the assumed prior and likelihood at least roughly reflect the statistics of the environment, I take it that what unites the numerous Bayesian analyses of cognition is adherence to 1, which concerns the logic of the inductive computations performed.[4] In other words, even if people maintain inaccurate priors and likelihoods and sometimes make bad choices given their information, the more people's inductive leaps on different tasks are shown to accord with Bayes' Rule, the more Bayesian rational analysis will be vindicated. As long as we can assume that the mind is (at least approximately) performing Bayesian computations, inferring the shape and parameters of people's priors and likelihoods is not an insurmountable task (see, e.g., Hemmer et al. 2015). The very idea that it would be reasonable to approach a novel cognitive modeling problem by assuming that the mind is effectively performing (approximate) Bayesian inference would already mark considerable progress on the identifiability problem.

Having now made clear what we take to be distinctive of Bayesian rational analysis, we can still ask, what is so rational about Bayesian conditionalization, particularly given this highly subjectivist variety of Bayesianism? Authors from Laplace to Pólya have considered Bayesian reasoning to be nothing more than tutored common sense, perhaps with no need of justification: Bayes' Rule naturally trades off prior plausibility of a hypothesis with its ability to explain the data. Others have proposed formal justifications, e.g., based on Dutch book arguments or epistemic utility arguments (see Huttegger 2013 and references therein). Some in the cognitive science literature (e.g., Perfors et al. 2011) follow similar arguments from de Finetti (1937) claiming that a Bayesian predictor is bound to out-predict any non-Bayesian in the long run. Without settling which, if any, of these arguments succeeds, let us momentarily accept that, for certain kinds of (perhaps highly idealized) agents making statistical inferences, adherence to Bayesian norms is indeed rational.

3 Challenges to Bayesian Rational Analysis

What does it take to show that a Bayesian model has successfully elucidated some cognitive phenomenon? Experimentalists almost never require that each individual participant, or even the majority of participants, give the Bayes optimal re-

[4] In fact, 2 and 3 point to two of the dimensions distinguishing varieties of Bayesianism according to Good (1971).

sponse on a task. Rather, often the posterior distribution associated with the model is compared to (some statistic of) the distribution of responses in an experiment (see Eberhardt and Danks 2011; Vul et al. 2014 for many examples). The usual justification for this degree of latitude is that Bayesian models are intended to capture cognition at the "computational level" in the sense of Marr (1982). This may engender more mechanistic "algorithmic level" models aimed at predicting (variation in) individual responses, but the computational level model is thought to be of interest in itself, giving a clean, intuitive portrayal of the abstract problem that the mind is assumed to be solving, and perhaps even shedding explanatory light on why the function works the way it does (e.g., Perfors et al. 2011).

There is at least one sufficient reason why we should not expect participants in an experiment to give optimal Bayesian responses: the calculations required by most Bayesian cognitive models are intractable. This in-principle tractability problem is accompanied by a related empirical issue. A common finding, going back at least to Estes (1959), is that the distributions of responses in a study will often *match* the normative posterior distributions, a phenomenon called "posterior matching" (Peterson and Beach, 1967). For example, an analysis by Vul et al. (2014) of the data from an influential study by Griffiths and Tenenbaum (2006) showed that the two distributions, response and model-posterior, match almost perfectly. This posterior matching behavior of course means that most participants are giving non-Bayes-optimal responses.

These facts—that Bayesian models are usually intractable, and that people show systematic deviations from Bayesian predictions at an individual level—bolster some of the more pressing criticisms of the approach. The central problem is that they seem to show either that Bayesian norms do not set the right standard of rationality, or just that people fall far short of meeting it.

A common response has been to suggest that people may be approximating Bayesian calculations in some way. In particular, a number of authors have proposed sample-based, e.g., Monte Carlo, algorithms that approximate Bayesian inference in the sense of asymptotic convergence (see the discussions by Vul et al. 2014 and many references therein). It is often supposed that these algorithms, and thus people's behavior, ought to be considered (approximately) rational because they approximate the rational ideal, in the sense that they converge to the ideal in the limit. In other words, people are doing as well as they could given their limited resources.

By their very nature, approximation algorithms can give rise to very different predictions from the ideal models they approximate, and this can even be important in accounting for experimental data. However, from the normative per-

spective, if Bayesian reasoning determines the standard, this means approximations can also deviate dramatically from what would be rational. The challenge to Bayesian models then becomes acute: why should we expect an approximation to Bayesian inference to be more rational than any number of alternative models that do not in any straightforward sense approximate Bayesian calculations? And if they are not, what distinguishes them as especially worthy of attention from the perspective of rational analysis?

This worry about rationality of approximate Bayesian calculations is compounded with a more general worry, that even for idealized agents without resource limitations Bayesian norms may not be uniquely rational, particularly when the environment statistics are unknown or constantly changing (Gigerenzer, 1991; Sutton and Barto, 1998). For instance, Gigerenzer and Brighton (2009) show that even a relatively tractable and a priori reasonable Bayesian model may be empirically worse at prediction than very simple heuristics—heuristics that systematically ignore potentially relevant information—for wide classes of problems. Douven (2013) argues that the rationality of Bayesian inference depends on the agent having specific epistemic goals, where other equally legitimate goals might favor other inference methods. None of this is to say Bayesian approximation algorithms could not provide a good model for a given phenomenon. The objection is to the claim that we have good reason to expect this to be the case, based solely on rationality considerations and a methodological optimism that the mind will have happened upon rational solutions to problems. We need to understand what additional assumptions could justify such optimism.

4 Boundedly Rational Analysis

The response we propose to these challenges is to accept them, but to argue that there may in fact be good reason to think that in many circumstances approximate Bayesian algorithms could be rational after all. We cannot reach this conclusion by merely noting that they approximate Bayesian calculations, e.g., by asymptotic convergence. Instead, we must reconsider our understanding of rationality—and in particular, of rational analysis—and show that these models do in fact meet the standards of this general and independently motivated concept of rationality.

Here we return to a point made at the outset. Rational analysis is premised on the idea that we can reach descriptive hypotheses from normative claims. But we might also think that what we consider normative depends to some extent on what kinds of agents we are. Classical accounts of rationality do this to a minimal

degree. Traditional Bayesian theory assumes that we are observationally limited, but that our information about the world can be represented by a probability distribution. Critics have often countered that this assumes too much and too little: we are limited not just observationally, but also computationally, and in part because of these additional limitations we cannot necessarily assume that our view on the world be captured by a probability representation. Herbert Simon (1956, 1976) famously argued that useful theorizing about human agents should focus not on *substantive rationality*, but what he called *bounded* or *procedural rationality*, which concerns the computational processes and algorithms we use to make inferences and decisions. Significantly, the extent to which a given algorithm or procedure is rational will also depend on details of the architecture on which it is running. Thus, on this view of rationality, details not only about the environment and task, but also about human psychology—the way the mind actually works—will be relevant to answering normative questions about inference.

This view of rationality may seem anathema to rational analysis. In fact, Simon (1991) criticized Anderson's rational analysis of categorization for being too focused on optimality, and paying too little attention to human agents as they are:

> Our interest is in the learning process itself—not a hypothetical one or an optimal one, but the one that people use. We want a learning theory precisely because people do not arrive at optimal classifications immediately or costlessly. (Simon, 1991, 35)

In the other direction, although the method includes a step for characterizing computational limitations, it is often said that a rational analysis is effective in as far as it makes the minimal assumptions at this step. Anderson (1990) himself characterized the methodology as a "nearly mechanism-free casting of a psychological theory" (30), and suggested that the potential need to specify too many architectural details is "the potential Achilles' heel of a rational approach" (32).

The core claim of this article is that there is a promising middle ground, and that finding it may help us understand when and why Bayesian analyses of cognition will lead to successful modeling. In this direction, we describe a framework for theorizing about procedural or resource-bounded rationality, which neither assumes nor rules out the possibility that Bayesian agents are rational. The framework combines aspects of the Bayesian evolution framework of Geisler and Diehl (2002) with an analysis of bounded optimal agents in artificial intelligence by Russell and Subramanian (1995), though it is more general than both of these.[5]

[5]For instance, unlike in these articles, we do not assume that *fitness* or *goodness* is necessarily

The advantages of this generality are (1) that it can be applied flexibly to a diverse array of agents facing a diverse array of tasks across different environments, and more importantly for present purposes, (2) that it highlights the considerable assumptions involved in distinguishing (approximate) Bayesian agents as optimal.

Imagine that we are modeling from outside an agent who will receive some data s, before choosing some action a. The utility U of taking action a depends on the state of the world ϑ, with $U(a, \vartheta)$ a real-valued measure of how good action a is in state ϑ. While it may draw upon the agent's own internal representation of value or desirability (if such there be), the utility function is to be understood from the external perspective of the theorist. Finally, suppose that we have some prior model on states of the environment, $P(\vartheta)$, as well as a likelihood function $P(s|\vartheta)$ over stimuli. Again, this is to be understood as capturing a theorist's perspective on the agent's situation.

Let us first think abstractly of agents as defining *agent functions* $\mathcal{A}$, which map observations s to distributions over actions, so $\mathcal{A}(s)(a)$ gives the probability that $\mathcal{A}$ will respond with a after receiving stimulus s. If we know the state of the world ϑ and the data s that the agent will receive, then we can define a score function σ, relative to some way Σ of combining action probabilities with utilities:

$$\sigma[\vartheta, s, \mathcal{A}] \;=\; \Sigma\big(\langle \mathcal{A}(s)(a_i), U(a_i, \vartheta)\rangle_{a_i}\big)$$

Here $\langle \mathcal{A}(s)(a_i), U(a_i, \vartheta)\rangle_{a_i}$ denotes the set of probability-utility pairs with a_i ranging over all possible actions. This just says that Σ is some function of this set. As an example, we might take Σ to sum over utilities of possible actions, weighted by their probabilities according to $\mathcal{A}(s)$:

$$\sigma[\vartheta, s, \mathcal{A}] \;=\; \sum_{a_i} \mathcal{A}(s)(a_i) \times U(a_i, \vartheta) \tag{1}$$

Prior to knowing anything about the world, we can define a measure of fitness ϕ, relative to functions Φ and Ψ, by combining the scores with the world probabilities in some manner:

$$\phi[\mathcal{A}] \;=\; \Phi\Big(\big\langle \Psi\big(P(\vartheta, s)\big) \times \sigma[\vartheta, s, \mathcal{A}]\big\rangle_{\vartheta, s}\Big) \tag{2}$$

We assume Ψ is weakly monotone increasing, so that state/observation pairs with higher probability have a higher impact on fitness, all else being equal. Again, this

determined by taking an arithmetic average over states of the environment. Meanwhile, Geisler and Diehl (2002) do not define any notion of *agent program*, and Russell and Subramanian (1995) do not analyze stochastic agent behavior, whereas these are both central to our framework.

expression means that $\phi[\mathcal{A}]$ is a function of probability/score pairs, now ranging over all possible state/observation pairs.

For example, taking Ψ to be the identity function, and taking Φ to define an expectation, $\phi[\mathcal{A}]$ gives a straightforward expected score:

$$\phi[\mathcal{A}] \;\; = \;\; \sum_{\vartheta,s} P(\vartheta, s) \times \sigma[\vartheta, s, \mathcal{A}] \tag{3}$$

We imagine nature chooses a state of the world ϑ and generates some stimulus s; the agent must then take an action a_i, and the payoff is the weighted sum of utility for each of the actions the agent might take in state ϑ. This is a natural and common way of assessing fitness of an agent, appropriate for many purposes. However, we leave the official definition of fitness in (2) more general, to incorporate minimax, maximin, geometric averaging, and other proposals. This allows evaluating agents in a way that, e.g., punishes variance or going below some minimum acceptable utility. Though we will not be using this generality further in the present article, it is important to highlight the fact that moving from the more general (2) to the much more specific (3) is substantive.

In the specific case of a prediction problem, we might stipulate that the action space and the state space coincide, so that the agent's task is simply to predict the state. This naturally suggests a utility function where $U(\vartheta, \vartheta') = 1$ if $\vartheta = \vartheta'$, and $U(\vartheta, \vartheta') = 0$ otherwise. We might, for example, think of categorization in these terms: guessing the correct category gives utility 1, while guessing the wrong category gives 0. Note then that, assuming (1), we have $\sigma[\vartheta, s, \mathcal{A}] = \mathcal{A}(s)(\vartheta)$. By a *standard prediction problem* we mean any setup with this utility function, with Σ and Φ defining expectations, and with Ψ weakly monotone increasing. It is easy to show that no agent function has better fitness in such a setting than an agent $\mathcal{A}^*$ who acts according to Bayes' Rule, choosing ϑ that maximizes $P(\vartheta|s)$ with probability one. We state this as a fact. (The derivation is routine but somewhat lengthy. See Okasha 2013, who essentially shows this in a very similar context.)

Fact 1. No agent outperforms a Bayesian agent in a standard prediction problem.

Fact 1 identifies one clear class of situations in which a Bayesian agent is uncontroversially—indeed, almost by stipulation—optimal. The assumptions built in to this fact are substantial. Aside from the stipulation of a standard prediction problem, it is also assumed that the agent antecedently knows the prior and likelihood functions. Rather than dwell on these assumptions, which can certainly be questioned, we turn to computational considerations.

A significant problem with $\mathcal{A}^*$, and thus with this characterization of optimality, is that in many cases no possible agent, not to mention any actual agent, could instantiate it. In order to incorporate computational limitations, we may refine agent functions by considering somewhat more concrete representations of agents and the actual computations they perform. Suppose we have a class Π of possible *agent programs*, from some "programming language" (cf. Pylyshyn 1984). This could be based on actual programs written in a concrete language such as Java, classes of neural networks with associated algorithms, or any other (possibly more abstract) type of representation that we would like to consider as candidates for some mental process.

We make two key assumptions. First, each program π in Π instantiates an agent function $\mathcal{A}_\pi$. Second, we can associate a *cost* to π for processing stimulus s, written $C_\pi(s)$. Then we define the cost-adjusted fitness of a program in an analogous manner:

$$\phi[\pi] \;=\; \Phi\Big(\langle \Psi\big(P(\vartheta, s), \sigma[\vartheta, s, \mathcal{A}_\pi], C_\pi(s)\big)\rangle_{\vartheta,s}\Big) \tag{4}$$

where ϕ now depends on the costs $C_\pi(s)$ of computation. As before, one reasonable (but not unique or uncontroversial) instantiation of this definition would simply give expected score minus costs:

$$\phi[\pi] \;=\; \sum_{\vartheta,s} P(\vartheta, s) \times \big(\sigma[\vartheta, s, \mathcal{A}_\pi] - C_\pi(s)\big) \tag{5}$$

Let us say an agent program π is boundedly rational to the extent that $\phi[\pi]$ is high.

This framework is remarkably general, and does not by itself issue very strong claims about bounded rationality. However, with further assumptions it may do so. Focusing on P and U, we can explore the landscape of "ecologically rational" agents (Gigerenzer and Brighton, 2009; Oaksford and Chater, 2007). Focusing on Π and the associated costs C, we can ask questions about what representations and algorithms are best suited to a given task (Anderson, 1990; Russell and Subramanian, 1995; Lewis et al., 2014). By manipulating both of these, we may find natural scenarios in which, e.g., "satisficing" agents (Simon, 1956) turn out to be optimal, or in which accuracy of representation is sacrificed in favor of non-epistemic ends (Stich, 1990; Mark et al., 2010). Finally, we can go even further and explore alternatives to how we want to assess agents—in terms of Σ, Φ, and Ψ—which might give rather different pronouncements from the standard expected utility assessment in (5). This may be particularly important if we want to relate this type of fitness to varieties relevant to evolution and adaptation (Sober, 2001).

Notably, this framework in no way assumes that a Bayesian agent will always be optimal. Even under the most standard fitness measure in (5), it may be that no cost-effective program in Π is actually consistent with a Bayesian agent function $\mathcal{A}^*$, pace Fact 1. However, it does allow us to ask a precise question: are there plausible assumptions about the components of this framework that will make a Bayesian agent—or more interestingly, an agent performing approximate Bayesian computations—appear boundedly rational? This will be our main question from here on.

5 Case Study: Posterior Matching

The prevalent phenomenon of posterior probability matching has been at the center of many of the criticisms of Bayesian approaches to cognitive science, particularly those that target rationality claims. Given the central importance of this phenomenon, we would like to refine our question from the previous section: are there plausible assumptions about the components of our framework that would render boundedly rational an agent who displays posterior matching behavior?

The observation that people's choices match not just empirical frequencies, but also more "notional" probabilities derived from normative models, has a long history (Estes, 1959; Peterson and Beach, 1967). Much of the human data—including in many recent Bayesian analyses—can be modeled using a class of functions first explored in psychology by Luce (1963), and which have subsequently become a standard tool in the field of reinforcement learning under the name of softmax functions (Sutton and Barto, 1998). Using the notation established in the previous section, suppose we have an agent that needs to choose from among various actions a_i, each assigned some estimated measure of value $V_s[a_i]$. A Luce choice agent $\mathcal{L}_\tau$ will take action a_i with probability:

$$\mathcal{L}_\tau(s)(a_i) \quad = \quad \frac{e^{V_s[a_i]/\tau}}{\sum_{a_j} e^{V_s[a_j]/\tau}}$$

where the parameter τ determines how close the agent is to maximizing value. As $\tau \to 0$, the agent maximizes value with probability 1; while as $\tau \to \infty$, all actions become equally likely. A special case of this setup is the standard prediction problem. A convenient value function for prediction problems, often used in machine learning and in neuroscience, is the log posterior probability, so that the value of guessing a state ϑ given data s is equal to $\log P(\vartheta|s)$. If we

assume this, the Luce choice agent $\mathcal{L}_\tau$ is then proportional to the following:

$$\mathcal{L}_\tau(s)(\vartheta) \quad \propto \quad e^{\log P(\vartheta|s)/\tau} \tag{6}$$

When $\tau = 1$, this agent posterior matches exactly: $\mathcal{L}_1(s)(\vartheta) = P(\vartheta|s)$. It is very common in Bayesian cognitive modeling to compare $\mathcal{L}_1$, rather than the "ideally rational" agent $\mathcal{A}^*$, to human data. There are also cases where the parameter τ is fit to some value potentially less than one, in order to provide a more flexible model (see, e.g., the discussion in Sanborn et al. 2010).

We saw that in a standard prediction problem, ignoring costs, $\mathcal{A}^*$ will be deemed rational, indeed optimal (Fact 1). In what sense, if any, is $\mathcal{L}_\tau$ rational? In a thorough discussion, Eberhardt and Danks (2011) consider a number of possible answers to this question, but find them all wanting. The problem is that in a single-shot inference problem there is no evident advantage to guessing a state with probability lower than the maximum.

One step in the direction of answering this challenge has been suggested by Vul et al. (2014). In a standard prediction problem, $\mathcal{L}_1$ properly describes the behavior of an agent who draws a sample state ϑ^* from $P(\vartheta|s)$ and uses ϑ^* as a response, while an agent who draws multiple samples and chooses the mode will correspond to $\mathcal{L}_\tau$ for some $\tau < 1$. Under certain assumptions about the cost of drawing samples from a posterior distribution, compared to the potential improvement in utility afforded by a more accurate approximation to the posterior, Vul et al. (2014) show that it is often optimal to draw very few samples, sometimes as few as one. Griffiths et al. (2015) extend this analysis by suggesting that, for concrete sampling algorithms such as those based on the Metropolos-Hastings method, optimal decisions involve not just relatively few samples, but also unavoidable bias in the sampling process, which is used to account for a number of psychological effects. The motivation behind this work is in the spirit of bounded rationality: because the computations required by application of Bayes' Rule are intractable, use of Monte Carlo sampling might make sense as "behavior that is actually optimal in the context of the limitations of the agent" (Vul et al., 2014, 26). The fact that Monte Carlo algorithms are also a favored engineering solution to hard probabilistic inference problems lends further credence to this suggestion.

At the same time, this response to the challenge arguably assumes too much at the outset. In terms of the framework from the previous section, Vul et al. (2014) essentially assume that Π consists solely of agents that draw some number of samples from the posterior before making a guess, differing only in how many samples are drawn. That is, the analysis only compares approximately Bayesian agents to

one another. This does not show that any of these agents will be boundedly rational in any more robust sense, in particular as compared to agents employing heuristics whose choice behavior may differ markedly. This is especially pressing given the possibility of scenarios in which simple heuristics perform demonstrably *better* than even very costly Bayesian calculations, not to mention algorithms that merely approximate those calculations. Agents whose behavior is described by $\mathcal{L}_\tau$ may well be boundedly rational in many cases of interest. But we cannot simply assume this without a compelling reason.

The difficulty in substantiating a claim of bounded rationality is related to the "potential Achilles heel" of rational analysis. We could perhaps learn that, as a matter of fact, the brain represents (conditional) probability distributions by a particular kind of sampling process (cf. Icard 2016). This would then allow us to sharpen hypotheses about what class Π of programs should be our focus, before then focusing on more specific issues such as the determination of an optimal value for τ. This kind of detailed analysis is possible, and can be profitable.[6] Yet, for high-level cognition we simply do not have such understanding, and the point of rational analysis is to help guide that very search, so that we know what to look for. Moreover, it seems likely that the details may be quite different from task to task. The diversity of tasks for which $\mathcal{L}_\tau$ has been used as part of a Bayesian rational analysis speaks further to the need for some more general type of justification.

As an analogy, consider the question of when we should expect a flexible architecture, as opposed to a fixed behavioral repertoire, to be adaptive for an organism. On the one hand it is certainly not obvious that flexibility will always be advantageous. But on the other hand, it would not be very illuminating to say merely that it depends on details of the case. One would like a more general characterization of conditions under which flexibility would be adaptive. Using a similar (but much simpler) decision-theoretic framework of assessment to what we proposed here, Godfrey-Smith analyzes this question, showing that, at least in certain simplified circumstances, flexibility will be useful when there is variability in environmental conditions that matter to the organism, but also stable correlations between these conditions and states of the agent (Godfrey-Smith, 1996). Our question is about a more specific, and more complicated, type of flexibility—namely, when it makes sense to behave in an approximately Bayesian manner according to $\mathcal{L}_\tau$—and we will thus need to make more substantial assumptions at the outset. However, the goal is roughly the same: to find a helpful level of

[6] See Lewis et al. (2014) for examples involving very detailed assumptions about architectural constraints involved in response ordering and eye movement tasks.

analysis that is sufficiently anchored to details that matter, while also achieving the kind of generality to which rational analysis aspires.

6 Entropic Cost Functions

One way of approaching questions about mental computation costs would be to find an interpretation of $C_\pi(s)$ in terms of relatively concrete *space* or *time* requirements incurred when program π processes stimulus s. This is undoubtedly appropriate when possible, and indeed, a fair amount is known about the biophysical and metabolic costs of neural computation in certain domains, e.g., in the sensory system (Niven and Laughlin 2008). The problem with this approach in the case of high-level cognition is that time and space costs are potentially sensitive to computational architecture, and the latter is much of what we are attempting to understand. Of course, as in the work on Monte Carlo sampling described above, one can profitably stipulate a certain kind of architecture together with associated costs and explore the consequences of this stipulation. The bounded rationality framework sketched earlier accommodates such approaches. A more ambitious aim, however, is to try to derive sampling-like behavior from first principles, based on (boundedly) rational analysis and reasonable, but perhaps more general, assumptions about cost. The proposal explored in this section is somewhat speculative, but maintains this more ambitious aim.

A number of authors have suggested that we might be able to understand computational resource bounds in terms of general thermodynamic costs associated with computing.[7] To help motivate the proposal, consider a simple thought experiment about ideal gases from Feynman (1998), further extended by Ortega and Braun (2013). Imagine a box X of gas with volume V_1 containing N atoms, and suppose we wanted to compress the gas to a smaller box Y with volume V_2. Assuming we can do this in a way that keeps the temperature T constant, the physical work W involved in this process is given by the well known equation:

$$W \quad = \quad NkT \log \frac{V_2}{V_1} \tag{7}$$

where k is the Boltzmann constant. In order to forge a bridge to information and computation, let us suppose there is only a single atom in X, and that moreover X can be partitioned into some cells $\vartheta_1, \vartheta_2, \ldots$, such that the smaller box Y is the

[7] The most detailed such account as applied to human decision making is offered by Ortega and Braun (2013).

union of some of these cells. Let us also define $P(\vartheta) = V_1(\vartheta)/V_1$, where $V_1(\vartheta)$ is the volume of ϑ in X, and likewise let $Q(\vartheta) = V_2(\vartheta)/V_2$. Assuming that the atom could equally likely be in any position within box X before compression, and anywhere in Y afterward, the probabilities P and $Q = P(\,\cdot\,|\,Y)$ quantify our uncertainty about which cell the atom occupies. The situation is as in Fig. 1.

Figure 1: Compressing X (with volume V_1) to Y (with volume V_2)

We can thus think of W as capturing the work involved in reducing our uncertainty about where the atom is, by shifting our distribution from P to the (sharper) distribution Q, as though we "learned" the proposition Y. Notice that with this shift in perspective on equation (7), we can see that the work is proportional to the following quantity (cf. Ortega and Braun 2013, 4):

$$\begin{aligned}
W \;&\propto\; \log\frac{V_2}{V_1} \\
&= \sum_\vartheta Q(\vartheta)\log\Big(\frac{V_2}{V_1}\,\frac{V_1(\vartheta)}{V_1(\vartheta)}\,\frac{V_2(\vartheta)}{V_2(\vartheta)}\Big) \\
&= \sum_\vartheta Q(\vartheta)\log\frac{V_2(\vartheta)}{V_1(\vartheta)} + \sum_\vartheta Q(\vartheta)\log\Big(\frac{V_2}{V_2(\vartheta)}\,\frac{V_1(\vartheta)}{V_1}\Big) \\
&= \sum_\vartheta Q(\vartheta)\log\frac{V_1(Y\cap\vartheta)}{V_1(\vartheta)} - \sum_\vartheta Q(\vartheta)\log\frac{Q(\vartheta)}{P(\vartheta)} \\
&= \sum_\vartheta Q(\vartheta)\log P(Y|\vartheta) + \sum_\vartheta Q(\vartheta)\log P(\vartheta) - \sum_\vartheta Q(\vartheta)\log Q(\vartheta) \\
&= \sum_\vartheta Q(\vartheta)\log P(Y,\vartheta) + H(Q)
\end{aligned}$$

That is, the work is proportional to (negative) total energy plus entropy $H(Q)$ of the smaller box Y. The negative of this quantity is sometimes referred to as the

free energy (Feynman, 1998).[8]

Our question is whether this way of thinking about resources could justify the sort of behavior characterized by Luce's choice rule. In that direction, let us relate this back to the problem of determining bounded fitness. The expression above for work immediately suggests a cost for an arbitrary agent function $\mathcal{A}$. Define the cost of agent function $\mathcal{A}$, given stimulus s, to be:

$$\begin{aligned} C_{\mathcal{A}}(s) &= \tau \sum_{\vartheta} \mathcal{A}(s)(\vartheta) \log \mathcal{A}(s)(\vartheta) \\ &= -\tau\, H\big(\mathcal{A}(s)\big) \end{aligned}$$

In other words, we assume very abstractly that the cost of implementing an agent function $\mathcal{A}$ is given by (some multiple of) the negative entropy of the distribution $\mathcal{A}(s)$ over actions.

The intuition here is simple. Without exercising any control over one's actions, any action would be a priori as likely as any other. Configuring oneself to ensure that a specific action is taken requires some physical work, including in situations where such work takes the form of mere thinking (about what to do, what the answer is, and so on). Even when one implicitly "knows" the right solution to a problem, distilling this solution sufficiently that one may act upon it is not a costless process. The assumption is that the cost of this work is inversely proportional to the uncertainty remaining in the agent's behavioral dispositions.[9]

Let us again assume $\Psi = \log$, so that value is given in terms of log probabilities, and let us otherwise assume we are in a standard prediction problem. For a fixed s, writing $Q = \mathcal{A}(s)$, we can treat agent functions $\mathcal{A}$ themselves as programs

[8]There has been much recent discussion in neuroscience and philosophy about a "free energy principle" as one way of fleshing out hypotheses about predictive coding and "active inference" in perception and cognition (see, e.g., Friston 2010). It is important to clarify that, apart from appeal to some of the same mathematical ideas from information theory and thermodynamics, the present proposal is in no way tied to these hypotheses. In particular, while the present proposal is certainly compatible with specific hypotheses about predictive coding, modular hierarchical processing, variational inference, and so on (see Sengupta et al. 2013), it in no way assumes them.

[9]It is worth noting that entropic cost functions have been independently proposed in economics and game theory to formalize a closely related concept, *cost of decision control*. For instance, Mattsson and Weibull (2002) draw upon representation theorems by Shannon and Hobson and interpret their axioms as conditions on a measure of cost of control, deriving (some multiple of) action entropy as the unique cost function.

with costs in a uniform manner:

$$\begin{aligned}\phi[\mathcal{A}] &= \sum_{\vartheta} \log P(s,\vartheta)\, \sigma[\vartheta, s, \mathcal{A}] - \tau \sum_{\vartheta} Q(\vartheta) \log Q(\vartheta) \\ &= \sum_{\vartheta} Q(\vartheta) \log P(s,\vartheta) + \tau\, H(Q) \end{aligned} \tag{8}$$

Thus we arrive right back to our quantity that we just saw was proportional to work. The crucial fact, which is a standard result in thermodynamics and information theory (Jaynes, 1957, 623), is that maximizing work, i.e., maximizing fitness—trading off by τ expected utility maximization with entropy maximization—gives us exactly the Luce choice rule with this value of τ:

Fact 2. Eq. (8) reaches a maximum with $\mathcal{A} = \mathcal{L}_\tau$. That is, the Luce choice rule is boundedly optimal in a standard prediction problem with entropic costs.

Fact 2 marks progress on our main question. Assuming a reasonable conversion factor τ, and assuming the entropic cost function is appropriate to begin with, this provides a boundedly rational analysis of posterior probability matching: what looks like a suboptimal version of a deterministic strategy may instead be optimal given the costs of effecting intelligent deterministic strategies. Specifically, this result singles out algorithms that (behave as though they) approximate Bayesian inference. An important question now is whether the entropic cost assumption is a sensible one. Acknowledging that the hypothesis is somewhat speculative, we offer here only a preliminary defense for why it is nonetheless worth taking seriously.

7 Role of Costs in (Boundedly) Rational Analysis

By defining costs in this very abstract way, applying directly to agent functions $\mathcal{A}$ rather than more concrete agent programs π, we essentially define a partition on any space Π of programs, whereby π_1 and π_2 are equally optimal if they lead to the same action distributions, i.e., if $\mathcal{A}_{\pi_1} = \mathcal{A}_{\pi_2}$. This is more fine-grained than merely looking at expected utilities, but it is certainly more coarse-grained than, for example, looking at time or space consumption. A legitimate concern is that this makes costs unhelpfully abstract. If we see that algorithm 1 produces the same behavior as algorithm 2, but consumes less time and space in the process, we would obviously say that algorithm 1 is more cost-efficient. The entropic

cost function is blind to this difference. It was claimed above that this level of generality should be seen as an advantage, and we now need to defend this claim.

The advantage emerges most clearly when we consider how assumptions about computational costs fit into the broader methodological program of rational analysis, the topic with which we began. When studying a given cognitive phenomenon we face the daunting task of narrowing down some vast set Π of potential programs. The point of a rational analysis is to focus our attention on those programs that *solve the underlying problem well*, and the point of *boundedly* rational analysis is to incorporate computational cost as a central component of what it takes to solve a problem well. Absent concrete information about costs, we would like to start somewhere.

The hypothesis is that this entropic cost function C is a good place to start: while it does not capture all aspects of computational costs, it does capture some important aspects. Quite generally, mental resources are required to make distinctions and to promote any particular action at the expense of potential alternatives. Thus, plausibly, C charges for behaviors that are truly resource-intensive. Yet, by design, such an abstract cost function as C cannot discriminate in a way that depends on concrete details of the computational architecture, and it is in virtue of this deliberate neutrality that C clearly cannot tell the whole story about cognitive costs. After all, if we knew the whole story we would probably not be faced with the identifiability problem in the first place. These considerations make it all the more striking just how much progress we can achieve. In particular, as revealed by Fact 2, with this relatively noncommittal assumption about cost we arrive at optimal behavior that is much closer to what we empirically observe in behavior.

To illustrate further what Fact 2 achieves, consider a prima facie deficiency of our cost model. Suppose there are just two equally likely states, ϑ_1 and ϑ_2, such that for any data point s they still appear indistinguishable: $P(\vartheta_1|s) = P(\vartheta_2|s) = 0.5$. Agent A takes the time and effort to draw a sample state from the posterior, conditioned on s, and thus has equal probability of choosing each. Agent B, by contrast, does not think at all and merely chooses randomly between ϑ_1 and ϑ_2. Somewhat counterintuitively, A and B suffer the same cost and are therefore equally optimal.

This result, however, is no more anomalous than the situation itself. From the perspective of rational analysis, we would not expect good performance in an environment like this to reveal any interesting cognitive structure at all. Imagine instead (as a simple idealization) that the possible data points come from the interval $(0, 1)$, and that we have likelihood density functions $f_{\vartheta_1}(s) = 2s$ and $f_{\vartheta_2}(s) = 2 - 2s$, so that $P(\vartheta_1|s) = s$ and $P(\vartheta_2|s) = 1 - s$. In this case, an agent

clearly stands to gain by attending to s. Assuming entropic costs with $\tau = 1$, the optimal agent is $\mathcal{L}_1$, who expects fitness of 0. Any agent who ignores s and steadfastly chooses ϑ_1 with probability p will expect fitness $-1 + H(p)$, which in the best case ($p = 0.5$) achieves only $-1 + \log 2 \approx -0.3$.

The broader lesson of this illustration is that acting optimally across a sufficiently wide range of problems, reflecting the true intricacy of typical cognitive tasks our minds face, is non-trivial. While boundedly rational agents in this setting may fall short of ideal Bayesian optimality—which make sense especially when conditional probabilities $P(\vartheta|s)$ are difficult to determine—they must nonetheless move in the direction of what Bayes' Rule dictates, acting as though they are appropriately, if only partially, incorporating their evidence. This already offers progress and potential insight into what kinds of agent programs should be the focus of our modeling attention (assuming, of course, that we are confident in our characterization of the underlying problem).

As we come to understand more details about inherent psychological costs and constraints, we should certainly expect C to be refined, which in turn may result in different optimality pronouncements. The working hypothesis is that such details will genuinely build on the present hypothesis, allowing us to make yet further distinctions. For example, perhaps among all possible algorithms instantiating the Luce choice rule in standard prediction problems, later insights about computational constraints will single out specific Monte Carlo sampling algorithms as uniquely optimal.

In closing, it is worth pointing out that our entropic cost assumption ought to be less contentious than the (very common) assumption we highlighted earlier in §4, namely that we are in a standard prediction problem. We have provided tools for justifying approximate Bayesian behavior, but this depends on a particular way of characterizing the problems cognition is assumed to be solving. For any specific cognitive function this characterization can be questioned (see, e.g., Murphy 1993 on Anderson 1990 on categorization). A full defense of (boundedly) Bayesian rational analysis would need to provide a convincing argument for why this assumption is justified in any given instance.

8 Conclusion

While Bayesian models of cognition have been used profitably to account for a wide range of psychological phenomena, the details of their application, coupled with their generally prohibitive computational costs, have generated a fair amount

of criticism. In particular, the question arises as to why it would make sense for an agent to employ inferential or decision making strategies that merely approximate "ideal" Bayesian solutions to these tasks, in the sense of asymptotic convergence given enough time and resources. We have addressed this question by proposing a very general way of theorizing about bounded rationality, which does not assume from the start that a Bayesian agent will be optimal. Instead, the framework allows a modular study of this question where we can explore different assumptions about utility structure, environments, representational capacity, and even methods of assessment. We demonstrated one example set of assumptions under which ideal Bayesian agents would be optimal (Fact 1), and used this to explore a very general hypothesis about computational cost under which sampling agents who approximate Bayesian calculations would in fact be boundedly optimal (Fact 2).

The resulting framework casts debates about Bayesian rationality in a different light, perhaps closer to an engineer's perspective. Indeed, Bayesian models and approximations thereto are ubiquitous in computational statistics, artificial intelligence, and other domains where inferences must be made under uncertainty as well as under tight resource constraints (MacKay, 2003). Such applications have directly inspired many of the concrete implementations of Bayesian models in cognitive science. This shift from ideal rationality to computationally bounded rationality makes sense from both perspectives: not only in engineering, but also in cognitive science, where we are searching for realistic, but nonetheless effective, mechanisms that a resource-bounded brain could plausibly implement, through some combination of incremental phylogenetic and ontogenetic improvement. The suggestion we have been exploring is that a kind of optimistic "computational Bayesianism" may be justified on these grounds, both as a generally good solution to a wide class of problems, and tentatively as part of a (boundedly) rational analysis of particular cognitive functions.

On a program like that sketched in this article, the coupling between normative and descriptive considerations—between 'ought' and 'is'—becomes intricate and delicate. On the one hand, knowing what kinds of agents we are, and equally critically, characterizing useful and realistic standards of rationality, is arguably essential to the task of assessing and improving our own reasoning and decision making behavior. On the other hand, according to the framework sketched here, this same kind of characterization—which we can use both for assessment and for generating hypotheses in cognitive modeling—should draw significantly upon facts about what we are like. Negotiating this precarious interdependence is clearly a significant challenge, but one that promises to be rewarding.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Lawrence Earlbaum Associates, Inc.

Bowers, J. S. and C. J. Davis (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin 138*(3), 389–414.

Chater, N. and M. Oaksford (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science 3*(2), 57–65.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater and M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science*, pp. 59–75. Oxford University Press.

Douven, I. (2013). Inference to the best explanation, Dutch books, and inaccuracy minimization. *The Philosophical Quarterly 63*(252), 428–444.

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

Estes, W. K. (1959). The statistical approach to learning. In S. Koch (Ed.), *Psychology: A Study of a Science*, Volume 2, pp. 380–491. McGraw-Hill.

Feynman, R. P. (1998). *Lectures on Computation*. Addison-Wesley.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré 7*, 1–68.

Frank, M. C. and N. D. Goodman (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science 20*, 818–829.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience 11*, 127–138.

Geisler, W. S. and R. L. Diehl (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London, B 357*, 419–448.

Gigerenzer, G. (1991). Does the environment have the same structure as Bayes' theorem? *Behavioral and Brain Sciences 14*, 495–496.

Gigerenzer, G. and H. Brighton (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science 1*, 107–143.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

Good, I. J. (1971). 46656 varieties of Bayesians. *American Statistician 25*, 62–63.

Green, D. M. and J. A. Swets (1966). *Signal Detection Theory and Psychophysics*. Krieger.

Griffiths, T. L., F. Lieder, and N. D. Goodman (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science 7*(2), 217–229.

Griffiths, T. L. and J. B. Tenenbaum (2006). Optimal predictions in everyday cognition. *Psychological Science 17*(9), 767–773.

Hemmer, P., S. Tauber, and M. Steyvers (2015). Moving beyond qualitative evaluations of Bayesian models of cognition. *Psychonomic Bulletin and Review 22*(3), 614–628.

Huttegger, S. M. (2013). In defense of reflection. *Philosophy of Science 80*(3), 413–433.

Icard, T. (2013). *The Algorithmic Mind: A Study of Inference in Action*. Ph. D. thesis, Stanford University.

Icard, T. F. (2014). Toward boundedly rational analysis. In P. Bello, M. Guarini, M. McShane, and B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting in Cognitive Science*, pp. 637–642.

Icard, T. F. (2016). Subjective probability as sampling propensity. *The Review of Philosophy and Psychology 7*(4), 863–903.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review 106*(4), 620–630.

Jones, M. and B. C. Love (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences 34*(4), 169–231.

Kitcher, P. (1987). Why not the best? In J. Dupré (Ed.), *The Latest on the Best*, pp. 77–102. MIT Press.

Kwisthout, J., T. Wareham, and I. van Rooij (2008). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science 35*, 779–784.

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology*, pp. 103–189. Wiley.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Marcus, G. F. and E. Davis (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science 24*(12), 2351–2360.

Mark, J. T., B. B. Marion, and D. D. Hoffman (2010). Natural selection and veridical perceptions. *Journal of Theoretical Biology 266*, 504–515.

Marr, D. (1982). *Vision*. W.H. Freeman and Company.

Mattsson, L.-G. and J. W. Weibull (2002). Probabilistic choice and procedurally bounded rationality. *Games and Economic Behavior 41*, 61–78.

Murphy, G. L. (1993). A rational theory of concepts. *The Psychology of Learning and Motivation 29*, 327–359.

Niven, J. E. and S. B. Laughlin (2008). Energy limitation as a selective pressure on the evolution of sensory systems. *The Journal of Experimental Biology 211*, 1792–1804.

Oaksford, M. and N. Chater (2007). *Bayesian Rationality*. Oxford University Press.

Okasha, S. (2013). The evolution of Bayesian updating. *Philosophy of Science 80*, 745–757.

Ortega, P. A. and D. A. Braun (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society of London, A 469*(2153).

Parker, G. A. and J. Maynard Smith (1990). Optimality theory in evolutionary biology. *Nature 348*, 27–33.

Perfors, A., J. B. Tenenbaum, T. L. Griffiths, and F. Xu (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition 120*, 302–321.

Peterson, C. R. and L. R. Beach (1967). Man as an intuitive statistician. *Psychological Bulletin 68*(1), 29–46.

Pylyshyn, Z. W. (1984). *Computation and Cognition*. MIT Press.

Russell, S. and D. Subramanian (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research 2*, 1–36.

Sanborn, A. N., T. L. Griffiths, and R. M. Shiffrin (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology 60*, 63–100.

Sengupta, B., M. B. Stemmler, and K. J. Friston (2013). Information and efficiency in the nervous system—a synthesis. *PLOS Computational Biology 9*(7), 1–12.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review 63*(2), 129–138.

Simon, H. A. (1976). From substantive to procedural rationality. In T. J. Kastelein, S. K. Kuipers, W. A. Nijenhuis, and G. R. Wagenaar (Eds.), *25 Years of Economic Theory*, pp. 65–86. Springer.

Simon, H. A. (1991). Cognitive architectures and rational analysis: Comment. In K. VanLehn (Ed.), *Architectures for Intelligence: The 22nd Carnegie Mellon Symposium on Cognition*, pp. 25–39. Lawrence Earlbaum Associates, Inc.

Sober, E. (2001). The two faces of fitness. In R. S. Singh, C. B. Krimbas, D. B. Paul, and J. Beatty (Eds.), *Thinking about Evolution*, Volume 2, pp. 309–321. Cambridge University Press.

Stich, S. P. (1990). *The Fragmentation of Reason*. MIT Press.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning*. MIT Press.

Tenenbaum, J. T., C. Kemp, T. Griffiths, and N. D. Goodman (2011). How to grow a mind: Statistics, structure, and abstraction. *Science 331*, 1279–1285.

Vul, E., N. D. Goodman, T. L. Griffiths, and J. B. Tenenbaum (2014). One and done? Optimal decisions from very few samples. *Cognitive Science 38*(4), 699–637.