

Commitments and Specificity in Bayesian Models

DRAFT: Do not quote without authors' permission.

Frederick Eberhardt

University of California, Berkeley; and Washington University in St. Louis

David Danks

Carnegie Mellon University; and Institute for Human & Machine Cognition

Abstract

Bayesian models of cognition are widely employed in the cognitive sciences. We argue that many of the commitments required for Bayesian models to be empirically and normatively adequate are insufficiently specified and defended. Bayesian models must be rational models, but the most common arguments for the rationality of Bayesian models suffer from fatal flaws. Bayesian models must specify a hypothesis space, but there is no account of how to respond when one realizes that one was fundamentally mistaken about the possibility space. Bayesian models must connect with behavior through some choice function, but the conceptually natural connection (maximize subjective expected utility) is different from the one that most cognitive scientists seem to believe is empirically correct (probability matching). In sum, we argue not that Bayesian models are useless or false, but rather that they are insufficiently specified, grounded, and defended.

Introduction

Bayesian models of learning and inference have become increasingly common in cognitive science. Several recent books focus on Bayesian models (Chater & Oaksford, 2008; Doya, Ishii, Pouget, & Rao, 2007; Oaksford & Chater, 2007), and a recent issue of *Trends in Cognitive Science* was devoted entirely to them (Chater & Manning, 2006; Chater, Tenenbaum, & Yuille, 2006; Courville, Daw, & Touretzky, 2006; Körding & Wolpert, 2006; Steyvers, Griffiths, & Dennis, 2006; Tenenbaum, Griffiths, & Kemp, 2006; Yuille & Kersten, 2006). A very incomplete sample of phenomena for which Bayesian models have been proposed includes: category learning and inference (Heit, 1998; Kemp, Perfors, & Tenenbaum, 2007; Kemp & Tenenbaum, 2003; Tenenbaum & Griffiths, 2001); causal learning and reasoning (Bonawitz, Griffiths, & Schulz, 2006; Griffiths & Tenenbaum, 2005; Sobel & Munro, 2006; Sobel, Tenenbaum, & Gopnik, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003); conditional inference (Oaksford & Chater, 2007; Oaksford, Chater, & Larkin, 2000); covariation assessment (McKenzie & Mikkelsen, 2007); imitation (Rao, Shon, & Meltzoff, 2004); information selection (Oaksford, Chater, Grainger, & Larkin, 1997); framing effects (McKenzie, 2004); memory effects (Schooler, Shiffrin, & Raaijmakers, 2001; Shiffrin & Steyvers, 1997); object perception (Kersten, Mamassian, & Yuille, 2004; Kersten & Yuille, 2003); repetition effects and priming (Mozer, Colagrosso, & Huber, 2002, 2003); and word learning (Xu & Tenenbaum, 2005, 2007).

We understand a Bayesian model to be committed to the following two claims:

1. Belief and uncertainty (i.e., “degree” of belief) about a domain, phenomena, or other suitably circumscribed collection are representable as a probability distribution over a (possibly infinite) set of exhaustive and mutually exclusive hypotheses H_1, \dots, H_n ; and

2. Given evidence E , the updated belief in a hypothesis H_i is given by $P(H_i | E)$, where this

$$\text{can be computed using Bayes's theorem: } P(H_i | E) = \frac{P(E | H_i)P(H_i)}{P(E)}.$$

In the jargon of Bayesianism, these models hold that an individual's strength of belief can be understood as an initial *prior probability distribution* over the space of possibilities that represents the reasoner's degree of (partial) belief in different propositions that characterize the world. In response to new data, the changes in an individual's beliefs can be modeled as the use of Bayes's theorem to update the prior probability distribution to a new *posterior probability distribution* over the various possibilities. This computation, also known as *Bayesian updating*, requires taking the product of two terms: (i) $P(E | H_i)$: the *likelihood* of observing evidence E if H_i really is true (often given by a so-called generative distribution that is relatively easy to specify); and (ii) the prior probability distribution. This product is divided by the *a priori* probability of the evidence, $P(E)$, which essentially acts as a normalization term. $P(E)$ need not be computed if we are interested only in the relative probabilities of the various hypotheses, though it is necessary to model the impact of rare (i.e., surprising) data.

As a very simple example of a Bayesian model, consider trying to determine the bias (if any) on a coin, given a series of flips. The hypothesis space for this problem is the infinite set of specific hypotheses about the bias of the coin: $P(H) = p$, for every p between 0 and 1. A Bayesian model must specify a prior probability distribution over these hypotheses; for example, one might use a uniform distribution over the $[0,1]$ interval if one is completely ignorant, or a truncated Gaussian with a peak at 0.5 if one thinks that a fair coin is more likely. The likelihood function for this learning problem is naturally given by the standard binomial distribution for the p of a particular hypothesis (e.g., $P(\text{Heads, Heads} | P(H) = 0.3) = (0.3)^2 = 0.09$). As one observes

more and more coin flips, the probability distribution over the infinite hypothesis space shifts by repeated applications of Bayes's theorem.

Bayesian models are appealing in cognitive science because they (i) allow for the representation of prior beliefs in the prior probability distribution; (ii) represent differences in background beliefs through different prior distributions in different individuals; (iii) model the integration of new evidence with prior beliefs through Bayesian updating; and (iv) explain gradual transitions in belief between various hypotheses by the use of probabilities to represent degrees of belief. Moreover, this representation of belief (using probabilities) and learning (as change in probability distribution) can be integrated cleanly with models of decision-making based on separable beliefs and desires (e.g., utilities).

To a first approximation, cognitive models can provide two different types of explanation—mechanistic and rational—corresponding to different sets of commitments for the model.¹ Models providing mechanistic explanations are by far the more common in cognitive science, and aim to characterize *how* some cognition or behavior is performed. Specifically, they describe the proximal causes that lead to particular behavior in particular situations. The representations and operations in these models are interpreted in a relatively realistic manner: they are thought to ultimately correspond (in a not-always-clear sense) to activity of collections of neurons or neural systems. One can, for example, try to find the neural bases of some computation in a mechanistic model, or compare the model's complexity with reaction time measurements. Bayesian models with mechanistic commitments have occasionally been offered (e.g., Doya, *et al.*, 2007; Lee & Mumford, 2003; Rao, 2005). Those models are interpreted and

¹ These types of explanations are technically not mutually exclusive; a model that says how some behavior arises can also explain why it arises. In practice, however, there are almost no models that attempt to provide both types of explanations.

evaluated in the “normal” manner in the cognitive sciences, and so we do not address them further.

More typically, Bayesian models are offered as models providing rational explanations, although the normative claim of the rationality of the models is often vague. We first argue that proponents of such Bayesian models must commit to substantial normative claims to ensure that the model has the relevant explanatory power and to justify the use of relatively “rich” Bayesian models. We thus turn to the potential rationality of Bayesian models, and argue that the two standard justifications of the rationality of Bayesianism both fall short: there is no positive argument that Bayesian updating is the rational method of belief change. We then argue that, even if Bayesianism can be shown to be rational, it faces a significant gap: there is no known, plausible account of rational belief change when the underlying hypothesis space changes. Finally, even if both of these challenges were somehow met, Bayesian models of learning and inference must somehow connect up with decisions and behavior in order to be testable. We argue that all of the obvious ways of doing so are either empirically or conceptually inadequate, or imply predictions that do not conform to the currently used method of empirical confirmation.

Our central thesis is not some form of naïve anti-Bayesianism; we believe that Bayesian models potentially have a place in the cognitive scientist’s toolbox. Our hesitation is based on a series of substantive challenges to the viability of Bayesian models *qua* rational models. *If these challenges are addressed*, then Bayesian models can provide powerful, useful, highly explanatory models. The antecedent is, however, our point of concern: there are no obvious candidates to satisfy the antecedent, and consequently Bayesian models often appear to provide only unmotivated, overly complex, purely instrumentalist summaries of data. Much of what we

aim to do here is to articulate, largely by a series of potential problems, the conditions that must be met for Bayesian models to have their full power and benefits.

Non-mechanistic Bayesian models must be rational

Ambiguity about the intended commitments of a Bayesian model arise because they are usually offered as so-called *computational level* models (Marr, 1982). Consider an abstract information processor trying to solve a problem in an environment. A computational-level description characterizes the <processor, problem, environment> triple in terms of input information, output behavior, and constraints (if any) on the possible computations in the processor. The name is thus something of a misnomer, as computational-level models specify what is (or should) be computed, but tell us nothing about what computations actually occur.² A computational-level description is entirely agnostic about the underlying algorithm or implementation. If a computational-level model shows how an information processor's behavior "solves" (in some sense) the given problem in a given environment, then it is a rational model. If it describes only the observed input-output function, then it only provides a summary of the data: it is instrumentalist.

Instrumentalist models have a poor reputation, but they are useful in certain situations. For relatively unexplored domains, we are often in the position of not even knowing what behavior is important, and a compact, parsimonious representation of the data can highlight behavioral features that have previously been ignored (e.g., sensitivity to base rate information). Even if we have a robust, well-confirmed cognitive (mechanistic) theory, an instrumentalist

² That information is specified in models at the algorithmic level (i.e., what is the actual underlying computational process?) and implementation level (i.e., how is the computational process performed in physical stuff such as neurons?)

theory can be pragmatically useful since it can often characterize the data using simpler computations. These two pragmatic considerations—compact representation of behavioral features and computational simplicity—both argue for using the simplest, weakest model that nonetheless captures the relevant data or predictions. Bayesian models, however, are relatively “rich” models requiring probability distributions over possibly-infinite hypothesis spaces, well-defined likelihood functions for all hypotheses, solutions to complicated integrals over parameter values, and so on. In many cases (e.g., sensitivity to base rates), much simpler models can do the same instrumentalist work as Bayesian models, and so pragmatic utility does not justify instrumentalist Bayesian models.

A proponent of an instrumentalist interpretation might instead argue that the use of Bayesian models is justified by their widespread applicability: the existence of Bayesian models for a wide range of domains provides some measure of unification to those disparate domains. This argument, however, only has non-pragmatic force when there is some similarity in either the underlying mechanisms that generate the phenomena, or the function/role of the phenomena in a containing system. “Unification” by an instrumentalist model is only valuable when there is a reason—shared mechanisms, or shared function/optimalty—for the data-level similarities. Thus, the instrumentalist Bayesian model has distinctive value only because of the existence of a shared mechanistic Bayesian model, or a shared rational Bayesian model. In either case, the instrumentalist interpretation of the Bayesian model is unjustifiably weak: stronger commitments can (and should) be made.³

³ Moreover, the idea that Bayesian models *should* be only instrumentalist would represent a surprising reversal for cognitive science. Much of the debate in the “cognitive revolution” was precisely about whether psychologists could justifiably talk about internal states. A restriction to instrumentalism would amount to giving up on that hard-won right.

Rationality of Bayesian models

Computational-level Bayesian models cannot be understood mechanistically, and there do not seem to be good grounds for using them with an instrumentalist interpretation. Thus, the remaining possibility is to understand Bayesian models as *rational* models that characterize the optimal response to a problem.⁴ A rational Bayesian model does not simply describe what people do; rather, it goes further and claims that people *should* act this way. Such models thus purport to answer “Why behavior *B*” by showing what the behavior provides to the organism, and in particular, how the behavior solves (optimally) some pressing challenge or problem faced by the reasoner. It should be noted that rationality of the model is not actually sufficient for this answer (Danks, 2008). Explanations in terms of optimality additionally require that the cognitive mechanism exists *because* it is optimal, and so one must show that the optimality of the mechanism played a causal role in its existence or maintenance. Accidents are not properly explained by their benefits, even if those benefits turn out to be optimal. Almost no computational-level Bayesian models provide serious, tested accounts of the development or acquisition of the cognitive mechanism. Instead, reference is usually made to possible accounts based in learning or natural selection. For the purposes of this paper, though, we will set aside the complex question of how to close this gap properly.

We understand ‘rational’ to be a descriptor that holds just when the reasoner’s input → output function (e.g., the perception → behavior function) provides an optimal solution relative to the reasoner’s abilities for a given problem or task in a particular environment. In

⁴ Proponents of computational-level Bayesian models sometimes suggest that the model need not actually be fully rational; “good enough” behavior might be sufficient. Claims of rationality are not, however, supererogatory add-ons to computational-level Bayesian models. They are absolutely necessary to justify many, and perhaps most, of the uses of such models.

characterizing rationality as a four-part relation—optimality on a task in an environment, relative to constraints—we explicitly follow the concept used in so-called *rational analyses* in cognitive science (Anderson, 1990, 1991; Chater & Oaksford, 2000; Oaksford & Chater, 1998, 2007). This conception of ‘rationality’ is stronger than those that require only the existence of *some* set of beliefs or preferences—perhaps a quite bizarre set—such that the responses are appropriate given that set; the notions of ‘representability’ and ‘rationality’ in revealed preference theory are of this latter type (Houthakker, 1950; Richter, 1966; Samuelson, 1938). At the same time, this notion of ‘rationality’ is weaker than those that require a belief or an action to be explicitly justifiable by the reasoner; many internalist accounts of epistemology have this requirement (e.g., Bonjour, 1985; Lehrer, 1988). The question thus arises: Are Bayesian models rational (according to this account)? It seems to be (almost) universally believed in the cognitive sciences that the answer is “yes”; we contend that the proper answer is “we do not know.” In particular, we focus here on the two standard arguments that Bayesian updating is the rational method of belief change: diachronic Dutch book and convergence to the truth.

Dutch book justification

This account of ‘rationality’ requires optimality of input-output functions for a task (typically, perception-behavior functions for solving a problem), but Bayesian models describe only inference and belief change, not behavior or action. In Dutch book arguments, Bayesian updating is connected with choice using standard decision theory: choose the action that maximizes subjective expected utility. More formally, given (i) a probability distribution $P(H)$ over possible states of the world (i.e., hypotheses), (ii) a probability distribution $P_A(O | H)$ over outcomes O given world-state H and action A , and (iii) utilities over outcome-state pairs $U(O,$

H), one should choose the action A that maximizes subjective expected utility: $SEU(A) = \sum_{h \in H} U(O, H) P_A(O | H) P(H)$. There is a long-running debate between causal and evidential decision-theorists (Jeffrey, 1965; Joyce, 1999; Lewis, 1981; Price, 1986) that centers on how $P_A(O | H)$ is determined; that question is irrelevant for Dutch book arguments, though, so we do not need to take a stand on the issue.

A *Dutch book* is a set of bets such that (i) the reasoner would agree to make all of the bets; but (ii) the reasoner is guaranteed to lose money on the set of bets, regardless of how the world turns out.⁵ An individual freely takes a bet if the subjective expected utility for one side of the bet is greater than the other side. For example, if a reasoner believes that a coin is fair, then she will accept a bet in which she wins \$4 if heads ($SEU = \2) and loses \$2 if tails ($SEU = -\1). Now suppose instead that I agree (for some bizarre reason) to simultaneously make the following two bets about that particular flip: (A) I win \$1 if heads, you win \$2 if tails; and (B) I win \$1 if tails, you win \$2 if heads. This pair of bets is a Dutch book since I am guaranteed to lose \$1, regardless of the outcome of the coin flip. A plausible necessary condition for rationality is that there are no Dutch books to be made against the reasoner: freely agreeing to a sure loss cannot possibly be optimal. Importantly, claims about Dutch books are not supposed to be contingent on the actual existence of malicious bookies, or other features of the setup. Rather, the existence of a Dutch book indicates an “irrational inconsistency of belief,” even if no actions are ever taken on its basis. This state is analogous to the irrationality on the basis of logical inconsistency of an agent who simultaneously believes P ; $P \Rightarrow Q$; and $\text{not-}Q$, even if she never acts on those beliefs.

Ramsey (1931) and de Finetti (1937) developed *synchronic* Dutch book arguments that showed that the unique way to avoid Dutch book at a particular point in time is for my degrees of

⁵ We use monetary gains and losses, rather than utilities, simply for ease of presentation.

belief to correspond to a probability distribution. That is, there exists a Dutch book against me at time t if and only if my degrees of belief (at time t) in a set of propositions do not correspond to a probability distribution over those same propositions. Thus, a necessary (and possibly sufficient⁶) condition on the degrees of belief of a *rational* reasoner at a particular point in time is that they must correspond to a probability distribution.

The Ramsey and de Finetti theorems say nothing about the rationality or irrationality of belief change. A *diachronic* Dutch book justification of the rationality of Bayesian updating involves proving that there is no possible Dutch book if one starts with coherent degrees of belief and responds to evidence by Bayesian updating. Teller (1973) credits David Lewis with providing the first concrete example of a diachronic Dutch book: a series of bets *distributed over time* that I freely accept, but that jointly guarantee that I lose money if I change my beliefs over time using some method other than Bayesian updating (and still accept the bets). Diachronic Dutch book arguments thus show that a particular type of “mind-change inconsistency” can only be avoided using Bayesian updating. More generally, Teller’s (1976) diachronic Dutch book theorems state: under certain (relatively strong) conditions, if there exists *any* updating method that is insulated from diachronic Dutch book, then Bayesian updating is insulated from diachronic Dutch book.⁷

⁶ Objective Bayesians argue against sufficiency, as they think that rational degrees of belief must satisfy other constraints. Subjective Bayesians typically argue that these additional conditions fall outside of the domain of rationality, and that correspondence to a probability distribution is sufficient.

⁷ Note that vulnerability to diachronic Dutch book does *not* necessarily imply vulnerability to synchronic Dutch book. A “learner” who randomly selects a new probability distribution after each piece of evidence will never be vulnerable to synchronic Dutch book (at any time), but will be vulnerable to Lewis-Teller diachronic Dutch books (though see the next paragraph).

Synchronic Dutch book arguments are widely accepted as showing that degrees of belief should be a probability distribution⁸; there are, however, several reasons to view diachronic Dutch book arguments with more suspicion.⁹ Most importantly, diachronic Dutch book arguments (for Bayesian updating) require that agents never change anything about their *conditional* beliefs (Levi, 1988, 2002). Diachronic Dutch book arguments for Bayesian updating imply that reasoners make initial, conditional commitments (formally, in the likelihood functions), and then never reconsider those conditional commitments, regardless of the evidence. The standard diachronic Dutch book arguments require a reasoner to commit to an enormous range of hypotheticals about how she will change her beliefs in almost any possible world, and then never revisit those commitments. Diachronic Dutch book arguments also require that the reasoner be incredibly myopic, since she is not permitted (in the arguments) to adapt her behavior based on predictions about her future states or retrospection on her past states. She cannot, for example, think about possible future or past bets and, if she recognizes that her conditional commitments have changed (or will change), then simply decline to enter into the sequences of bets that could lead to a diachronic Dutch book (Levi, 1988; Maher, 1992). Alternately, suppose a reasoner recognizes that she might have future belief changes that are truly irrational (e.g., because of drinking too much wine, or ingesting certain drugs). In these cases, the diachronic Dutch book arguments preclude the possibility of the agent acting preemptively to protect her future interests (e.g., by handing her car keys to someone else).

⁸ For example, Isaac Levi is one of the most prominent and persistent critics of diachronic coherence arguments, but even he agrees that synchronic rationality requires that degrees of belief correspond to probability distributions (e.g., Levi, 1988, 2002).

⁹ The following discussion simplifies some technical issues, but all of the objections can be stated formally. Also, most contemporary discussions focus on van Fraassen (1984)'s Reflection principle, but we focus on the special case of Bayesian updating (i.e., Reflection when we learn only that *E* definitely occurred).

Although the best world would be one in which reasoners never had irrational belief changes, such worlds are nonetheless possible (and actual). It seems bizarre to declare that it is irrational for reasoners to protect themselves when such irrational changes are foreseeable, but diachronic Dutch book arguments do exactly that (Maher, 1992).

We can weaken the diachronic Dutch book arguments so that they do not have some of these absurd consequences, but then they also no longer establish that Bayesian updating is the *unique* protection against irrationality.¹⁰ Instead, there are many different ways to avoid diachronic Dutch books, of which Bayesian updating is only one (e.g., Douven, 1999). This lack of uniqueness poses a problem for Bayesian models. The explanatory value of a rational model lies in its ability to answer questions of the form “Why this behavior (and not some other)?” Therefore, if there are multiple rational models that make distinct predictions, then any particular model’s answer to this question is immediately followed by the question: “Why that rational model/behavior/response, and not one of these other rational ones?” A rational model explains why people manifested some behavior when it privileges one set of behaviors over others on normative grounds. The explanatory value of a particular rational model is thus inversely correlated with the total number of rational models (given an account of ‘rationality’). Since multiple alternatives to Bayesian updating (in fact, infinitely many) also avoid diachronic Dutch book, Dutch book arguments give us no particular reason to privilege Bayesian models *qua* rational models.

¹⁰ Diachronic Dutch book arguments for Reflection similarly fail to show uniqueness (Hild, 1998; Maher, 1992).

Convergence justification

A different justification of the rationality of Bayesian updating focuses on learning and inference in the long run. A plausible necessary condition on any rational belief change method is that it should converge to the truth when possible. A rational learning method should eventually learn the truth when it is learnable (though the method might also value, e.g., short-run predictive accuracy). Bayesian updating satisfies this necessary condition for rationality: if the true hypothesis H is empirically distinguishable from other hypotheses and $P(H) \neq 0$, then Bayesian updating provably converges to maximal degree of belief in the truth given more and more evidence (a canonical expression of this result is provided in Savage, 1972).¹¹ That is, regardless of the initial beliefs of the Bayesian reasoner, she will eventually believe the truth (assuming certain technical conditions). Bayesian learning is sensitive to prior beliefs, but is still always “truth-directed.”

Of course, convergence is not a sufficient condition, as there are infinitely many convergent methods that are clearly *sub-optimal*: for example, the strategy of guessing randomly for 5000 years and then using Bayesian updating will converge to the truth, but is obviously not a rational learning method. One plausible further necessary condition focuses on speed: there should not be a reliable learning method that always converges faster. Given the choice between two methods that converge to the truth, one should rationally use the method that gets to the truth faster, or at least is no slower in getting there. Bayesian updating (for any non-dogmatic prior probability distribution) provably satisfies this latter condition: there is no method that gets to the truth faster than Bayesian updating in *every* “world” (i.e., regardless of which hypothesis is true, and the order of the randomly sampled evidence). There may be alternative non-Bayesian

¹¹ More precisely, for all ε , $P(\text{reasoner has degree of belief greater than } 1-\varepsilon \text{ in the truth}) \rightarrow 1$ as the number of datapoints goes to infinity.

methods that get to the truth faster in particular worlds, but none outperforms Bayesian updating in every world (Schulte, 1999). There are, however, infinitely many non-Bayesian methods that are similarly not speed-dominated: Bayesian updating might outperform them in some particular worlds, but it does not converge to the truth faster in every possible world. We thus have an argument based on “fast, reliable convergence to the truth” that Bayesian updating is a rational learning method, but it is also an argument that Bayesian updating is not uniquely rational.

The claim to rationality based on convergence is further undercut by a standard requirement for any Bayesian model: namely, the Bayesian reasoner must be instantaneously logically omniscient. A Bayesian reasoner must know all (relevant) logical and mathematical implications of the various hypotheses that she entertains.¹² In practice, this knowledge is encoded in the likelihood function: hypotheses that are logically inconsistent with the evidence assign zero likelihood to that evidence (i.e., $P(E | H) = 0$ if H and E are logically inconsistent). For many interesting problems, one must determine likelihoods by actually computing the probability of the evidence given the theory; a natural constraint on a likelihood function is thus that it be a computable function (i.e., there exists a machine that can compute it). This computability constraint is, however, sometimes incompatible with the logical omniscience requirement: there are learning problems that can be solved in the long run by computable falsificationist methods (e.g., Popper’s method of “assert the hypothesis until it is refuted”) that cannot be solved by a computable Bayesian reasoner (Juhl, 1993; Kelly & Schulte, 1995; Osherson, Stob, & Weinstein, 1988). The Bayesian reasoner only converges to the truth (whenever the truth can be learned) if she can sometimes “compute” uncomputable functions.

¹² We focus on the relevance of the logical omniscience requirement for the convergence argument, but there are other grounds on which to challenge this requirement. For example, it is *prima facie* inconsistent with mathematical practice. If reasoners are logically omniscient, then they should not, for example, be uncertain about most mathematical propositions.

Bayesian updating thus seemingly fails to satisfy even the original necessary convergence condition, unless the Bayesian reasoner can “compute” functions that no machine can possibly compute. One possible response by the Bayesian is to argue that likelihoods express the reasoner’s *current* beliefs about the evidence, not necessarily the actual likelihood. The challenging cases are all ones in which the true likelihood of the evidence is zero if the computing machine never halts, but where it could take arbitrarily long before actually halting. Thus, the Bayesian does not necessarily know the true likelihood in the short run. If she instead uses an appropriate “subjective likelihood” function¹³, then the convergence result pops back out. This revised reasoner, however, will sometimes find herself giving non-zero probability to propositions that are mathematically impossible, and so will not have degrees of belief that correspond to a probability distribution. She will thus be vulnerable to synchronic Dutch book. In particular, there are situations in which this “Bayesian” will be vulnerable to Dutch book at *every* point in time. This strongly subjectivist reasoner thus maintains the convergence argument at the expense of both the synchronic and diachronic Dutch book justifications.

A completely different response by the Bayesian would be to deny the relevance of these sorts of problems for Bayesian models in cognitive science. In fact, there are (to our knowledge) no Bayesian models proposed in cognitive science for problems that require uncomputable likelihood functions. Thus, one might suggest that the challenging problems are not practically relevant; they are (perhaps) just a “philosopher’s trick.” This response misunderstands the point of the critique, however. In practice, Bayesian models are claimed to be rational because of a positive answer to the *general* conceptual question: “Is Bayesian updating always rational?” These problematic cases demonstrate, however, that the general question cannot be answered

¹³ Any $P(E \mid \text{machine for } H \text{ has not halted in } n \text{ steps})$ that monotonically decreases in n suffices.

affirmatively on the basis of convergence arguments. These cases do not prove that no Bayesian models are rational; rather, they imply that proponents of Bayesian models in cognitive science must provide additional arguments for their specific models. There is no general convergence argument for the rationality of Bayesian models in all cases. And of course, even in cases for which there is such a convergence argument, there will still be infinitely many non-Bayesian models with the same convergence properties.

A weakened notion of ‘rationality’

In light of the significant difficulties with both Dutch book and convergence justifications, one might instead try to establish “Bayesian updating is rational” relative to some weaker account of ‘rational’ (e.g., Oaksford & Chater, 2007). In particular, one could focus on rationality as successful, goal-directed action, where the goals depend on the particular situation. The constraints used to judge this notion of ‘rationality’ are situation- *and behavior*-specific, rather than formal: “which...rational principles should be used to define a normative standard for particular...tasks...is constrained by the empirical human reasoning data to be explained” (Oaksford & Chater, 2007, p. 31). That is, the scientist assumes that people’s behavior is largely rational, and then finds a (small, coherent) set of formal principles that justify that behavior as normatively correct. Violations of a normative theory (e.g., the well-known Allais or Ellsberg “paradoxes”) are seen as challenges to the appropriateness of the putative normative theory, not indicators of irrationality. But this response misunderstands what normative principles of rationality are supposed to do, as they are exactly supposed to *not* be situation-dependent in this way. If they are situation-dependent, then they are only restatements of the observed behavior, and so simply instrumentalist. As Oaksford & Chater (2007) note, the appeal to a weaker notion

of ‘rationality’ only works if we avoid extreme situation-dependence by finding normative standards that are “consistent with other knowledge, independently plausible, and so on” (p. 31). But in that case, we are right back in the situation of searching for relatively abstract formal constraints that justify some particular method as ‘rational.’

We close by reiterating the overall theme of this paper: these arguments do not show that Bayesian models are never rational; rather, they show that much more thought and argument is required to justify the claim that Bayesian models as rational models. The two standard justifications of the rationality of Bayesian updating both suffer from serious flaws, and so we simply do not know at the current time whether Bayesian models are actually rational. The use of Bayesian models as rational models awaits some successful justification.

Lack of rational Bayesian responses to changes in the hypothesis space

Even if a suitable defense can be found for the rationality of Bayesian updating, standard Bayesian models have a notable lacuna. Any Bayesian model is defined only for a particular hypothesis space: the set of possibilities that the model entertains. In many psychologically plausible cases, however, the hypothesis space changes over the course of learning. No rational response to changes in the underlying hypothesis space has been offered, and we argue here that none of the obvious potential solutions will work. Thus, even if their rationality were established, Bayesian models would be suitable for only a limited type of learning and inference: namely, when the reasoner considers only one set of possibilities, and never adds or removes possibilities.

The hypothesis space of a Bayesian model includes all options that it can possibly learn, and anything not in this space is unlearnable, since long-run Bayesian learning is convergence to the most probable hypothesis *among those in the space*. Moreover, if the truth is excluded from

the hypothesis space, there are rarely guarantees that the Bayesian model will converge to the closest approximation within its hypothesis space (Grünwald & Langford, 2007). Consequently, a Bayesian model can only adequately capture learning in a domain if it represents all relevant, possible hypotheses, but this is often quite difficult (or impossible) to do *a priori*. A brief example illustrates the problem.

Recall the example from the Introduction of the reasoner trying to learn the bias of a coin. Her hypothesis space contained all $P(H) = c$, for c between 0 and 1. Initially, her prior belief in the coin coming up heads may be peaked at $P(H) = 0.5$ (a fair coin), but other $P(H)$ have non-zero probability. Suppose that the coin comes up heads on the first 10 tosses. Her posterior belief will shift towards $P(H) = 1$. Now suppose that the next 10 tosses are all tails, so the new posterior will be closer to $P(H) = 0.5$ again. Finally, suppose that, as she continues tossing, there are always 10 heads followed by 10 tails. At some point, as her posterior belief converges towards $P(H) = 0.5$ (i.e., the hypothesis of an unbiased coin), she will notice that the coin's behavior appears completely deterministic. Thus, her hypothesis space should not have been over different values of $P(H)$ but over different random and non-random sequences; it should consist of hypotheses about different possible features of the sequence of flips. This second hypothesis space is not an expansion or a contraction of the first; rather, it is a qualitatively different space.

The problem of changing hypothesis spaces is well known in the philosophy of science. Scientific revolutions are generally taken to be revolutions precisely because they include changes in fundamental assumptions, changes of ontological primitives, or both. Even if there are doubts about the nature of the scientific hypothesis space at a particular time, there is little disagreement about the fact that revolutionary theories are radically different from the hypotheses in use before. Kuhn (1962) refers to this as the incommensurability of paradigms;

Lakatos (1970) refers to different research programs; Feyerabend (1975) saw these changes as demonstrating a failure in principle of scientific method. Given such incommensurability, philosophy of science has principally been concerned with how the insights gained from one theory could (rationally) be transferred to the new one, when these theories do not share a common hypothesis space.

There seem to be two natural lines of response for a Bayesian model. First, one can argue that the “actual” hypothesis space throughout the process was the universal space of all different sequences, and the representation in terms of $P(H)$ was just a notational simplification made possible by a higher-order assumption that tosses were independent. Second, the Bayesian can argue that there actually was a change in the underlying hypothesis space as the pattern in the sequence was noticed, and somehow the degrees of belief in the hypotheses of the original space were transferred to hypotheses in the new space. Neither type of response adequately handles this problem.

Universal hypothesis space

Perhaps putative or apparent changes in the hypothesis space are illusions, and not actual changes to the space. That is, both the original and new hypothesis spaces might be just subspaces of some universal hypothesis space. However, the nature of such a universal hypothesis space is obscure. It cannot be the set of hypotheses (about the issue at hand) that one currently considers plausible since, as the earlier example illustrated, learning may introduce ways of understanding a problem that originally had not even been considered. Instead, a universal hypothesis space must contain, from the very beginning, every hypothesis we ever could possibly entertain about the domain (even if most have only minimal prior probability).

Learning (i.e., Bayesian updating) then occurs on this universal space, but we are only ever consciously aware of the small parts of that space corresponding to the hypotheses we currently entertain. Other hypotheses only come to light when enough data accumulates to push the posterior probability of the hypotheses into the radar of consciousness. While it is conceptually possible that a Bayesian model operates on a universal hypothesis space, such a model throws its explanatory power overboard on several fronts:

- 1) Most importantly, the universal hypothesis space is so enormous that any feasible computation on this space requires substantial additional assumptions (e.g., very low probability for vast areas of the space and almost-zero likelihood in these areas for most evidence). Models are rational for particular types of agents, and so one cannot ignore the empirical fact that we are (at least somewhat) computationally bounded agents. An “in principle” solution can thus only be provided if one assumes that the algorithm used to solve the problem exploits these assumptions. In the abstract, a computational-level model need not make any algorithmic commitments, but the assumptions necessary for computations on the universal hypothesis space imply constraints that cannot (and should not) be swept underneath the algorithmic rug.
- 2) In a universal hypothesis space, many quite different hypotheses will arguably have similar likelihoods for large portions of the possible evidence (e.g., Newtonian and relativistic mechanics have the same likelihoods for most common evidence). If two hypotheses have the same likelihoods for some evidence, then (mathematically) their posterior probabilities must be in the same ratio as their prior probabilities. Thus, much of the “learning” in the universal hypothesis space will actually be determined by the prior probability distribution, since that will determine which of the equally well-confirmed

models has significant posterior probability (and so is consciously considered). Many crucial questions about learning are thus pushed into questions about the structure of the prior probability distribution over the universal hypothesis space.

- 3) Most pro-Bayesian experimental evidence confirms Bayesian models that use constrained hypothesis spaces. If learning in fact occurs on a universal hypothesis space, then an argument is required to connect confirmation of a Bayesian model with a small hypothesis space with confirmation of Bayesian learning on a universal hypothesis space. Arguments that would justify learning on subspaces independently first, and then some simple form of “tacking together” such subspaces to form the whole, are not obvious, and would require independent motivation.
- 4) The universal-hypothesis-space-response requires that there be a common set of ontological primitives for all of the hypotheses. One of the lessons of the corresponding debate in philosophy of science (Feyerabend, 1975; Kuhn, 1962; Lakatos, 1970), however, is that changes in the hypothesis space generally involve some form of incommensurability, and often incommensurability between ontological primitives. It is doubtful, for example, that there are suitable primitives for every possible elementary particle hypothesis, including: earth, fire, air, and water; vs. indivisible atoms; vs. quanta in a field; vs. tightly coiled strings; vs. who knows what else for future highly confirmed hypotheses. Since science is a part of learning, the problem transfers to the psychological case: one must provide suitable primitives for a universal hypothesis space in psychological tasks. There have been many attempts to formalize a “universal” language, but all have turned out to be notoriously difficult. Moreover, representations of hypotheses using abstract mathematical structure (analogous with structural realist views

in philosophy of science, as in Worrall, 1989) is not sufficient to explain important parts of learning, such as how actual evidence confirms or disconfirms abstract hypotheses, and how abstract hypotheses can be used to guide real action. Abstract structures without real world interpretations cannot guide action.

Recent approaches that impose additional structure on the hypothesis space, such as hierarchical Bayesian models (e.g., Griffiths & Tenenbaum, 2007; Kemp, *et al.*, 2007; Tenenbaum, *et al.*, 2006; Tenenbaum & Niyogi, 2003), use that structure essentially to skew the probability distribution over hypotheses in a principled manner so that most of the probability is concentrated on small subspaces of hypotheses. These methods have conceptual appeal and enjoy much success in accounting for phenomena such as successful learning from very few samples.¹⁴ While intuitively it might seem that these approaches could account more generally for radical changes from the prior to posterior in different hypotheses (given the appropriate evidence and structure over hypotheses), this is only the case for hypothesis spaces for which the update over the entire space is easily computed. Such models do not address the central computational concern regarding a universal hypothesis space. More generally, structured procedures only provide an overall simplification if large swaths of the hypothesis space can be excluded. But such exclusion is precisely what the universal hypothesis space idea rejects, as any hypothesis might at some point gain weight. The whole point of the original response was to not constrain the hypothesis space, but then we need to keep track of what is happening even in parts of the space we do not deem very plausible at all.

It would be a mistake to conclude that this problem is equivalent to the frame problem and therefore applies to any model whatsoever. The frame problem typically focuses on knowing

¹⁴ Incidentally, this view has a corresponding predecessor in the philosophy of science (Reichenbach, 1949).

what variables or evidence can be safely ignored. The present problem arises for probability-based models because they require a pre-determined space of possibilities over which probabilities can then be specified and computed. Other models, such as logic-based models, do not have the same requirement that one give an *a priori* specification of all possibilities. Of course, those models are also based on some formal language, but their expressive content can be adjusted dynamically, with more limited impact on the existing corpus of beliefs or commitments. Sentences can be added without affecting the truth-values of existing sentences (though the new deductive closure must be computed). For example, Levi (1991; 2004) has tried to use models based on a Boolean algebra over a dynamic set of propositions to make sense of the idea of expansion and contraction of one's current hypothesis space. His account does not explain how competing hypotheses are introduced (the "context of discovery", in philosophical terms), but focuses on determining which hypotheses are included in the knowledge corpus based on decision theoretic considerations with respect to the consistency and informativeness¹⁵ of the current state of knowledge. Levi rejects the idea that all hypotheses are maintained only with probability. Rather, there is a set of hypotheses that is held with "full belief," but where those "fully believed" hypotheses can be revised in light of future evidence.

We mention Levi's theory for two reasons: First, it provides an account of rational belief change (i.e., satisfies weak Dutch book and convergence criteria) that does not involve Bayesian updating as its core feature. Second, his account shows that the problem of specifying a hypothesis space is not a problem that applies to every possible model of learning. Consequently, while the move to probabilistic models provides various advantages over logic-based models, probabilistic models should not be thought to constitute a superset of logic-based models with a

¹⁵ Informativeness on Levi's account is a measure on the expressive power of the Boolean algebra, not an information theoretic notion.

few extra free parameters. The move is made also at the expense of some conceptual aspects that can be handled more easily within logic-based models.

Bayesian change of hypothesis space

Instead of using a universal hypothesis space, a Bayesian could attempt to describe genuine hypothesis space change. Given two different hypothesis spaces HS_1 and HS_2 , the task is to explain how the probability distribution P_1 —the degrees of belief for the hypotheses in HS_1 —is transferred into a probability distribution P_2 over the hypotheses in HS_2 . If HS_2 is a subset of HS_1 , then one can simply use marginalization to transform P_1 into P_2 , though it is unclear whether marginalization constitutes a genuine *contraction* of a hypothesis space.¹⁶ The more interesting situations are those in which HS_2 contains hypotheses that are not contained in HS_1 , and so an expansion or shift in hypothesis space occurs. Intuitively, one might suggest that hypotheses that are in both HS_1 and HS_2 should preserve their relative probabilities, and all new hypotheses should receive equal probability at some low value. This would avoid (most) diachronic Dutch book type arguments, and therefore preserve a degree of rationality. But such an account seems descriptively inaccurate: hypothesis space change often involves placing significant probability on the *new* hypotheses (recall the shift to the deterministic sequence in the coin tossing example), and new hypotheses that are similar in nature to existing hypotheses (whatever that may mean in a particular circumstance) should presumably receive similar probabilities. For example, suppose that HS_1 contained two hypotheses, H_1 involving a car, and H_2 involving a bus. If a new hypothesis H_3 in HS_2 also involves a bus (that is different in some

¹⁶ Specifically, marginalization seems like just a sparser representation of the more general hypothesis space, rather than a denial of the possibility or adequacy of the hypotheses that are marginalized out.

way, of course), then there might be good reason to believe that there is some close relation between the original probability of H_2 and the new probability of H_3 . The basic point is that probabilities cannot simply be transferred across spaces; the transfer of probabilities must be in some sense rational, and the specific circumstances of a change in hypothesis space can involve probability transfers for which there is no obvious rational constraint. Again, we do not think it is impossible to provide an account, but much more work is needed, because no obvious Bayesian account is available, and the straightforward suggestions will not do.

Connecting rational Bayesian models with action

Our previous concerns were mainly foundational and conceptual, but we now turn to a methodological one. Even if we have a fully specified, rational, Bayesian model with solid foundations, we must somehow confirm the Bayesian model with empirical data. Since it is not a mechanistic model, we cannot correlate reaction times on tasks with the computational complexity of the update operation involved, and there need not be neurological evidence of Bayesian updating. The only evidence is correspondence in input and output behavior between humans and a Bayesian model, and this is what one standardly finds in the literature (Oaksford & Chater, 2007). In general, output behavior is given by a function that takes beliefs and utilities and determines a choice of output. Behavioral evidence therefore can only distinguish between <belief, utility, choice function> compounds that imply different behavior. The Bayesian model specifies the beliefs (and how they are updated) and proper experimental design (hopefully) fixes the utilities, but the choice function is rarely explicitly specified (though see, e.g., Körding & Wolpert, 2006; Oaksford, *et al.*, 1997). The challenge arises because the choice function that is

most appropriate conceptually (maximize subjective expected utility) is inconsistent with the one that seems to be most widely believed in psychology (probability matching).

If rationality is the motivating feature of a Bayesian model, then the conceptually most obvious choice function is the one used for the Dutch book arguments: namely, choose the option that maximizes (subjective) expected utility. In most psychological experiments, the utility that a participant receives for providing the right answer is the same for all hypotheses; the utility when she correctly answers “ H_1 ” is the same as when she correctly says “ H_2 .” In these situations, one maximizes expected utility by choosing the hypothesis with maximal posterior probability. This theory of rational choice thus predicts that the response distribution for the whole population of participants should be degenerate: all of the participants should respond with the hypothesis or choice with maximal posterior probability (or expected utility, for the rare cases in which the options happen to have non-equal utility).

This prediction is completely at odds with the empirical data. Many different psychological theories (both Bayesian and non-Bayesian) hold on empirical grounds that choices are made by probability matching: participants select hypothesis H with a probability corresponding to the posterior probability of H (assuming constant utilities); in other words, the behavioral response looks like a random sample by the individual participant from her posterior probability distribution. In practice, the content of a “probability matching” hypothesis is typically ambiguous between two claims: (A) the distribution of responses from a *population* of participants corresponds to the posterior distribution over hypotheses determined by the model; and (B) *each* participant in a population chooses an hypothesis by taking a random sample from her posterior distribution over hypotheses (which supposedly corresponds to the posterior of the model). Claim (B) implies (A), but not vice versa. Moreover, (B) is significantly harder to test:

one must collect repeated choices in the “same” situation from the same participant in order to determine whether her behavior corresponds to the appropriate posterior probability distribution.

Probability matching behavior is considered well-confirmed in numerous psychological experiments and enjoys theoretical support since, under fairly weak assumptions and the appropriate interpretation, it is derivable from the Luce choice axiom (Luce, 1959, 1977) that characterizes intuitive principles underlying human choice.¹⁷ Moreover, the standard confirmation method for Bayesian models is to compare the distribution of responses from a pool of participants with the posterior probability distribution of the model (e.g., Bonawitz, *et al.*, 2006; Kemp, *et al.*, 2007; Kemp & Tenenbaum, 2003; Sobel & Munro, 2006; Xu & Tenenbaum, 2005, 2007). Notice that, if each participant has the same hypothesis space, priors, and utilities, and performed an exact Bayesian update on the same data, then this evidence confirms Bayesian models only under the stronger hypothesis of probability matching (claim B) that each individual participant chooses her response hypothesis by sampling randomly from her posterior.

We thus have four propositions that are either explicitly stated in the literature on Bayesian models or are naturally implied by the constraints of rationality, but which seem to be mutually inconsistent:

- 1) People are Bayesian about learning and inference [explicit claim];

¹⁷ The Luce choice axiom states:

- (i) If options a and b are in a choice set S and a is never chosen over b in the binary choice situation, then a can be removed from S without affecting any choice probabilities; and
- (ii) If R is a subset of S , then the choice probabilities for the choice set R are identical to the choice probabilities for S conditional on R having been chosen (i.e. $P_R(a) = P_S(a | R)$ for all a in R).

The standard reading of the original Luce choice axiom assumes that the probability matching behavior occurs at the individual level, i.e. claim B, but subsequent literature goes both ways. Some authors take the Luce choice axiom to be only descriptive of observed behavior and not committed to the underlying choice mechanism in each individual. Consequently, they only take the Luce choice axiom to imply claim A.

- 2) People choose the option that maximizes expected utility given their beliefs [requirement of rationality];
- 3) Experiments successfully constrain participants' prior beliefs and utilities [methodological assumption]; and
- 4) The distribution of participant responses matches the model posterior [empirical data].

Given our focus on Bayesian models, we will not dwell on non-Bayesian resolutions of this puzzle (i.e., rejections of proposition 1), although there are plenty of suggestions (with their own problems) in the literature. There are also weaker versions of proposition 1 that try to resolve the apparent inconsistency by appeal to inaccuracies or approximations in the Bayesian computations. We have no doubt that approximations and errors take place in human inference, and that a perfectly degenerate distribution of responses cannot be expected. But this observation is only helpful for a rational model if the relevant algorithm or error distributions are specified and defended. One cannot claim that the rational model need not commit to an algorithm and then blame approximations in the algorithm when the data does not fit the model's prediction. We will also not debate proposition 4, since this is a purely empirical claim that is widely accepted in multiple scientific communities (e.g., economics, marketing, animal behavior, cognitive science, etc.).

There are both strong and weak ways of rejecting proposition 3. One might argue that priors and utilities differ widely across participants, and these differences actually fall outside the bounds of the purported (or plausible) measurement error reported in experiments (i.e., give a "strong" rejection). If there are widely varying priors or utilities in the population, then the match of empirical response distribution and Bayesian model posterior can arise simply from aggregation over the (varying) participant population. There may well be significant variation

between participants in the prior beliefs imported into the lab, the extent to which they are imported, and the participant’s utilities. But this strong rejection of proposition 3 calls into question many of the standard methods of experimental design in psychology and the procedures used to control for alternative explanations. It is a rather big pillar to start shaking. Moreover, most experiments cited here counterbalance relevant aspects of the cover story to control for (among other things) differing utilities, and use a control condition in which participants are asked to report the most likely hypothesis without seeing any evidence. Thus, any rejection of proposition 3 must argue not just that there is variation, but that the non-degenerate response distribution could arise from individual variation *that would not be picked up in these controls*. One could instead argue that the variation in the population is smaller than the plausible measurement error in the controls of the experiment, but is nonetheless sufficient to explain the empirical results (“weak” rejection of proposition 3). We turn now to one such proposal.¹⁸

¹⁸ We note, though, that neither the weak nor the strong rejection of proposition 3 seem like a plausible solution of the puzzle if the rejection focuses on the priors. For a given hypothesis space and arbitrary evidence, there is not a distribution of priors in the subject pool that would result in the necessary empirical matches between response distribution and model posterior, *and* between model prior and response distribution in a control condition. More formally, suppose we have N participants with priors: $P_1(H), \dots, P_N(H)$. Assume also that the likelihoods ($P(D | h)$ for all h in H) are the same for all individuals, and that all individuals choose optimally given their beliefs. The “population prior” (the initial aggregate response distribution) is the distribution over H of the number of participants for which h maximizes their prior, i.e. $P_{pop}(h) = \#_N \operatorname{argmax}_H P_i(h)$. Assume all individuals use Bayesian updating. Then the “population posterior” (the final aggregate response distribution) is $P_{pop}(H | D) = \#_N \operatorname{argmax}_H P_i(h | D)$.

The puzzle can be solved by appeal to variation in the participant priors only if the population posterior is distributionally equivalent to Bayesian updating on the population prior: $P_{pop}(H | D)$ must be distributionally equivalent to $P(D | H)P_{pop}(H)/P(D)$. Mathematically, this holds when: $\#_N \operatorname{argmax}_H P(D | h)P_i(h) \approx P(D | H) \#_N \operatorname{argmax}_H P_i(h)$. For arbitrary likelihoods, this condition is not satisfied for standard prior distributions (e.g., flat or Gaussian), although it may be satisfied for those priors and particular $P(D | H)$. (We know of no such analyses.) But satisfaction in special cases would only provide further support that the appearance of probability matching is largely accidental.

Random utility maximization

McFadden (1974) explores the circumstances under which a group of utility maximizers will exhibit a response distribution that looks like every individual reported a random sample from the same function. That is, McFadden attempts to explain the observed probability matching as a result of aggregation of optimal responses from a population (claim (A)).

Specifically, suppose that each individual's utility function, $U(H)$, is given by: $U(H) = V(H) + e(H)$, where H is a hypothesis in the space under consideration, $V(\cdot)$ is a non-stochastic function representing a utility function shared by every member of the population, and $e(\cdot)$ is stochastic, representing the individual's deviation from the population utility $V(\cdot)$. Individuals in McFadden's model choose the hypothesis that maximizes their own personal $U(H)$. Under a fairly weak set of assumptions on the distribution of individual variations¹⁹, McFadden shows that choice based on individual utility maximization leads to a population response distribution with the same maxima and minima as the shared, population-level utility function. Specifically, there are plausible conditions on $e(\cdot)$, and sensible monotonic functions $f(\cdot)$, such that the set of responses from a population of utility maximizers is given by: $f(U(H_i)) / \sum_j f(U(H_j))$.

The focus on utilities is irrelevant here. The basic point is: if a population of participants share a common "trend"-function over a space of options (e.g., choices, hypotheses, etc.) but have independent individual deviations (of a certain kind) from this trend, then the population-level response distribution when each individual maximizes her individual utility is a distribution that has the same maxima and minima as the trend-function. Moreover, these deviations can

¹⁹ If $e(\cdot)$ is i.i.d. with a Weibull distribution for each hypothesis in the space (and each participant), then the distribution of responses is given by $\exp(U(H_i)) / \sum_j \exp(U(H_j))$. The Weibull distribution is sufficient; weaker constraints on the distribution of $e(\cdot)$ can be given.

arguably be sufficiently small that they would not be easily observed in experimental settings. One could even remain agnostic as to the source of the individual deviations, as long as they satisfy the condition in McFadden's theorem (see previous footnote). In the Bayesian case, the trend-function can be the posterior probability, or the expected posterior utility, and individual deviations could arise in either the inference computation or the utilities.

A McFadden-style response preserves the rationality (if any) of both the update and choice procedure, and explains the appearance of probability matching at the population level as the result of individual differences within the population. The commitments of a McFadden-style response have testable implications: depending on assumptions one is willing to make about the precise nature of the distribution of individual discrepancies, the functional relation between the response distribution and the trend-function (the model prediction) can be determined and checked. In the specific case discussed in McFadden, the response distribution is an exponential transformation of the population trend (in the Bayesian case, the model posterior). The response distribution should thus share the location of maxima and minima with the model posterior, but not be identical with it. Identity is the standard confirmatory test for Bayesian models (as stated or implicit in the papers we cite with regard to our puzzle), and it is not known whether there is a plausible distribution for individual deviations that implies identity between the model posterior and the response distribution. Such a distribution is required unless the usual method of testing the Bayesian model is weakened or changed. Nevertheless, this proposal is of great interest, since the necessary sources of variation potentially exist in psychological experiments, and the proposal lends itself to other possible empirical predictions possible of empirical confirmation.²⁰

²⁰ If one had a way of influencing individual differences, then one could test predictions of a McFadden-style response. If one had two participant populations with different distributions of

Optimal betting in constrained environments

An alternative to McFadden's approach is to argue for the stronger claim about probability matching (B) by finding circumstances in which it is optimal or rational for an individual to take a random sample from her posterior.²¹ This response would amount to a rejection of proposition 2. There are several different avenues for trying to make this argument, but the most obvious candidates are circumstances with competition and resource constraints. For example, suppose each individual in a population believes that some resources are located at location X with probability .75, and at location Y with probability .25. If there is no competition for the resources, then—assuming more is better—the optimal behavior for an individual is to go to location X, since it is the more likely location. If resources are constrained, however, then the optimal strategy is to go to location X 75% of the time, and to location Y 25% of the time (assuming everyone has the same utilities for resources). That is, one should probability match (Stephens & Krebs, 1986).

While such behavior is clearly rational in this context, our present focus is on seemingly quite different circumstances that do not obviously involve competition or constraints on shared resources. A claim that such competitive considerations (conscious or not) are at work in the experimental circumstances of hypothesis learning implies that an experimenter's efforts to remove or control for aspects that might induce competitive behavior are futile, and so the

deviations, then these would provide predictions for a new task of the response distribution of the combined population.

²¹ For example, Fiorina (1971) argues that many experiments use non-random event probabilities that may appear non-constant to the participant. If event probabilities fluctuate, then choosing the hypothesis with maximal posterior probability is no longer optimal. This response is insufficient, though, since fluctuating probabilities do not automatically imply probability matching behavior is optimal. Moreover, probability matching occurs even in experiments in which participants appear to consider the event-probabilities stable.

proposal would be a rejection of both proposition 2 and 3. Of course, the success of experimental control will never be perfect, but in the case of learning hypotheses, one uses counter-balancing to control for differences in the specific utility of any hypothesis, and so the trigger of competitive behavior must derive from the posterior probability alone. Perhaps people exhibit competitive behavior for pure truth in real life, rather than just in psychological experiments, but there is little evidence for this. Moreover, it is quite unclear why there would be such competition, as true propositions are not a constrained resource: my knowledge of a true proposition does not preclude you from knowing it. Much more explanation is thus needed for why participants would import such competitive behavior in the first place, how exactly that behavior is imported, and in what sense psychological experiments resemble those circumstances of real life where such competitive behavior is used.

A seemingly different argument for the rationality of selecting a hypothesis by probability matching is based on an approach combining statistical learning theory with Bayesian inference procedures: PAC-Bayesian theory (see, e.g., Seeger, 2003 for a review). Probably Approximately Correct (PAC-) learning theory provides bounds that, with high probability, constrain an algorithm's worst-case generalization error; that is, it bounds (with high probability) the true error-rate of an algorithm given the algorithm's error-rate on a training sample. The appealing aspect of integrating features of PAC-learning with Bayesian inference is that PAC-bounds are robust even when the true hypothesis is not included in the hypothesis space. However, known PAC-Bayesian results only imply that probability matching on the posterior is optimal (i.e., provides the tightest PAC-Bayesian bounds) for tasks that already contain in their description some aspect of probability matching, such as estimating a distribution, selecting a hypothesis stochastically, or providing a weighted average (McAllester, 1999, 2003). In fact, for

the task of interest to us—selecting the true hypothesis—PAC-Bayesian considerations imply the same response as rational choice: select the hypothesis that maximizes the posterior of the model (McAllester, 1999, p. 165). We are thus in the same situation as with the arguments for competitive behavior: we need an independent explanation why subjects would import behavior that is suboptimal for the task they are asked to solve. In addition, PAC-Bayesian arguments further require motivation for why long-run, worst-case considerations are relevant.

One final response argues that problem solving and learning strategies are automated and optimized for everyday use to such an extent that the artificial settings of psychological experiments are unable to control and focus the subject on the optimal strategy for the task of the experiment. Despite an experimenter's best efforts at ecological validity of the experiment and despite debriefing reports by subjects that they understood the task and did their best, the underlying learning and inference machine might have been solving a different problem. If so, then a response is not an indication of how the subject solved the task at hand, but rather an indication of how an ingrained problem solving strategy renders this problem. For all we know, this strategy may involve Bayesian inference and be highly optimized for most of our everyday learning problems, just not for the specific task of the experiment. This response raises more questions than it answers:

- a) What is the nature of rationality employed for the automated cognitive processes?
- b) What is the underlying task that the behavior is rational for, and how can we determine such a task?
- c) Why do we solve a task other than the one the experiment poses?

Question (a) brings us full circle on the argument in this paper. As it stands, someone who takes probability matching to describe the actual choice procedure of an individual (claim

(B)) is committed to the claim that a participant's choice is not Bayes optimal. Rational choice (proposition 2) is mathematically much simpler than Bayesian updating, and it seems backwards to reject a simple theory of rational choice to preserve a complex theory of rational belief update.²² Moreover, it is unclear why *rational* belief updating is valuable if the choices made on the resulting beliefs are not rational. Pointing out that the Luce choice axiom implies (under particular assumptions) choice behavior in accordance with probability matching is not an argument, since the Luce choice axiom is not a normative principle, and therefore cannot be regarded as providing an account of rational behavior. Moreover, that axiom is itself ambiguous.

We repeatedly find that we do not get off the ground in testing Bayesian models as they currently stand in the literature. We require a weakening of (at least) one of the core claims—our four inconsistent propositions—that are currently used in the confirmation of Bayesian models, and in most cases such a weakening must go along with substantial additional commitments about some other aspect of the problem set-up, such as the utility constraints for the decision theoretic analysis. We have argued that many—but not all—attempts that weaken assumptions about the control of priors and utilities (proposition 3) are implausible. But if we cannot weaken proposition 3, then we are back at the challenge of providing a coherent notion of ‘rationality’ such that both Bayesian belief updating and probability matching choice behavior are rational.

²² There are also multiple arguments that probability matching really is irrational. Shanks, *et al.* (2002) argue that people must regard probability matching behavior as irrational since participants shift towards maximization behavior (though only with substantial effort) in experiments with (i) large financial incentives, (ii) feedback, and (iii) extensive training. Vulkan (2000) takes a similar view with regard to the difference in expected monetary payoff between probability matching and maximizing. West & Stanovich (2003) show that maximizing behavior correlates with indicators of high cognitive ability.

Conclusion

One of the most important desiderata of a scientific theory—perhaps *the* most important one—is explanatory power. Bayesian models are commonly presented as computational models describing how beliefs are updated, and so are taken by many (though certainly not all) in the community to have marginal or non-existent implications for the algorithmic and implementation level. If a Bayesian model describes only input-output relations, then it is instrumentalist and so has no substantial explanatory power, and simpler models are similarly successful with regard to pragmatic considerations of instrumentalist accounts. Thus, the explanatory power of a Bayesian model must derive from its rationality. As we argued in Section 3, however, Dutch book and asymptotic convergence arguments for the rationality of Bayesian updating fail for a number of reasons. At the current time, there does not seem to be a coherent, justifiable notion of ‘rationality’ that both implies that Bayesian updating is rational, and also can supply the explanatory strength needed for a computational-level model.

We do not advocate the use of logic-based models, but the strong dependence of a Bayesian model on a precise specification of the hypothesis space raises a variety of questions about the nature of the hypothesis space, and indicates that Bayesian models do not simply inherit the benefits of logic-based models. We argued that a Bayesian model must either (i) provide extra machinery to account for changes in the hypothesis space, or (ii) address an enormous algorithmic challenge whose solutions will depend on significant additional assumptions, some of which will imply (testable) constraints at the algorithmic level. Lastly, given a fully specified Bayesian model, one can only test the model if one provides bridge principles that connect it with observed behavior. Current empirical evidence directly conflicts with the most obvious and conceptually defensible principle of rational choice. Moreover, most

plausible alternative solutions import strong additional constraints or substantially weaken the tenets of experimental design in psychology.

We are thus left in the following state: Without a non-trivial account of rationality, and without a specification of the choice procedure, just about every empirical finding can be fit to confirm a Bayesian model. Bayesian models would thus be relatively untestable, and so explanatorily vacuous. We showed in the case of the discussion of McFadden's random utility maximization that a resolution of our puzzle is possible with commitments that are plausible, but that then also imply further empirical tests. We doubt that these are the only possible alternatives, but commitments of this type simply have not been openly suggested and defended in the psychological literature. Without such commitments, Bayesian models do not provide testable hypotheses. With such commitments, we can run experiments, but proponents of Bayesian models must be clear about which commitments they are, and are *not*, making.

Acknowledgments

Numerous conversations with Josh Tenenbaum, Tom Griffiths, Noah Goodman, and Chris Lucas helped shape the arguments and ideas in this paper, though we doubt that they would endorse many (or any) of our conclusions. Clark Glymour also provided valuable comments and critiques. The first author was supported by a grant from the James S. McDonnell Foundation Causal Learning Collaborative.

References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471-484.
- Bonawitz, E. B., Griffiths, T. L., & Schulz, L. E. (2006). Modeling cross-domain causal learning in preschoolers as Bayesian inference. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 89-94). Mahwah, NJ: Lawrence Erlbaum Associates.
- BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.
- Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*, 335-344.
- Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*, 93-131.
- Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford: Oxford University Press.
- Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*, 287-291.
- Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*, 294-300.
- Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 59-75). Oxford: Oxford University Press.

- de Finetti, B. (1937). La prevision: Ses lois logiques, se sources subjectives. *Annales de l'Institut Henri Poincare*, 7, 1-68.
- Douven, I. (1999). Inference to the best explanation made coherent. *Philosophy of Science*, 66, S424-S435.
- Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.). (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: The MIT Press.
- Feyerabend, P. (1975). *Against method*. London: New Left Books.
- Fiorina, M. P. (1971). A note on probability matching and rational choice. *Behavioral Science*, 16, 158-166.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 334-384.
- Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 323-345). Oxford: Oxford University Press.
- Grünwald, P., & Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66, 119-149.
- Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248-274). New York: Oxford University Press.
- Hild, M. (1998). The coherence argument against conditionalization. *Synthese*, 115, 229-258.
- Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, 17, 159-174.
- Jeffrey, R. (1965). *The logic of decision*. New York: McGraw-Hill.

- Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.
- Juhl, C. F. (1993). Bayesianism and reliable scientific inquiry. *Philosophy of Science*, 60, 302-319.
- Kelly, K. T., & Schulte, O. (1995). The computable testability of theories making uncomputable predictions. *Erkenntnis*, 43, 29-66.
- Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10, 307-321.
- Kemp, C., & Tenenbaum, J. B. (2003). Theory-based induction. In *Proceedings of the 25th annual conference of the cognitive science society*.
- Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, 55, 271-304.
- Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, 13, 1-9.
- Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10, 319-326.
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91-196). New York: Cambridge University Press.
- Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20, 1434-1448.
- Lehrer, K. (1988). Metaknowledge: Undefeated justification. *Synthese*, 74, 329-347.

- Levi, I. (1988). The demons of decision. *Monist*, 70, 193-211.
- Levi, I. (1991). *The fixation of belief and its undoing: Changing beliefs through inquiry*.
Cambridge: Cambridge University Press.
- Levi, I. (2002). Money pumps and diachronic books. *Philosophy of Science*, 69, S235-S247.
- Levi, I. (2004). *Mild contraction: Evaluating loss of information due to loss of belief*. Oxford:
Clarendon Press.
- Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, 59, 5-30.
- Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.
- Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*,
15, 215-233.
- Maher, P. (1992). Diachronic rationality. *Philosophy of Science*, 59, 120-141.
- Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.
- McAllester, D. A. (1999). PAC-Bayesian model averaging. In S. Ben-David & P. Long (Eds.),
Proceedings of the 12th annual conference on computational learning theory (pp. 164-170).
New York: ACM.
- McAllester, D. A. (2003). PAC-Bayesian stochastic model selection. *Machine Learning*, 51, 5-
21.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka
(Ed.), *Frontiers in econometrics* (pp. 105-142). New York: Academic Press.
- McKenzie, C. R. M. (2004). Framing effects in inference tasks-and why they are normatively
defensible. *Memory & Cognition*, 32, 874-885.
- McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment.
Cognitive Psychology, 54, 33-61.

- Mozer, M. C., Colagrosso, M. D., & Huber, D. E. (2002). A rational analysis of cognitive control in a speeded discrimination task. In T. Dietterich, S. Becker & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 51-57). Cambridge, MA: The MIT Press.
- Mozer, M. C., Colagrosso, M. D., & Huber, D. E. (2003). Mechanisms of long-term repetition priming and skill refinement: A probabilistic pathway model. In *Proceedings of the 25th annual conference of the cognitive science society*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Oaksford, M., & Chater, N. (1998). *Rational models of cognition*. New York: Oxford University Press.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
- Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *23*, 441-458.
- Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*, 883-899.
- Osherson, D. N., Stob, M., & Weinstein, S. (1988). Mechanical learners pay a price for Bayesianism. *Journal of Symbolic Logic*, *53*, 1245-1251.
- Price, H. (1986). Against causal decision theory. *Synthese*, *67*, 195-212.
- Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156-198). London: Harcourt, Brace, & Co.

- Rao, R. P. N. (2005). Bayesian inference and attentional modulation in the visual cortex. *Cognitive Neuroscience and Neuropsychology*, *16*, 1843-1848.
- Rao, R. P. N., Shon, A. P., & Meltzoff, A. N. (2004). A Bayesian model of imitation in infants and robots. In K. Dautenhahn & C. Nehaniv (Eds.), *Imitation and social learning in robots, humans, and animals: Behavioral, social, and communicative dimensions*. Cambridge: Cambridge University Press.
- Reichenbach, H. (1949). *Theory of probability*. Berkeley, CA: University of California Press.
- Richter, M. K. (1966). Revealed preference theory. *Econometrica*, *34*, 635-645.
- Samuelson, P. A. (1938). A note on the pure theory of consumer's behavior. *Economica*, *5*, 61-71.
- Savage, L. J. (1972). *Foundations of statistics*. New York: Dover.
- Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review*, *108*, 257-272.
- Schulte, O. (1999). The logic of reliable and efficient inquiry. *Journal of Philosophical Logic*, *28*, 399-438.
- Seeger, M. (2003). PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, *3*, 233-269.
- Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233-250.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Rem-retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

- Sobel, D. M., & Munro, S. (2006). When Mr. Blicket wants it, children are Bayesian. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual meeting of the cognitive science society* (pp. 810-815). Mahwah, NJ: Lawrence Erlbaum Associates.
- Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, 28, 303-333.
- Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.
- Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, 10, 327-334.
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27, 453-489.
- Teller, P. (1973). Conditionalization and observation. *Synthese*, 26, 218-238.
- Teller, P. (1976). Conditionalization, observation, and change of preference. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 1, pp. 205-259). Dordrecht: Reidel.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629-641.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10, 309-318.
- Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual conference of the cognitive science society* (pp. 1152-1157). Mahwah, NJ: Erlbaum.

- van Fraassen, B. C. (1984). Belief and the will. *Journal of Philosophy*, *81*, 235-256.
- Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101-118.
- West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*, 243-251.
- Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica*, *43*, 99-124.
- Xu, F., & Tenenbaum, J. B. (2005). Word learning as Bayesian inference: Evidence from preschoolers. In *Proceedings of the 27th annual conference of the cognitive science society*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*, 288-297.
- Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301-308.