Confirmation in the Cognitive Sciences: The Problematic Case of Bayesian Models

(pre-print, the final publication is available at springerlink.com)

Frederick Eberhardt
Philosophy-Neuroscience-Psychology,
Washington University in St Louis
eberhardt@wustl.edu

David Danks
Department of Philosophy,
Carnegie Mellon University
and Institute for Human and Machine Cognition

Abstract:
Bayesian models of human learning are becoming increasingly popular in cognitive science. We argue that their purported confirmation largely relies on a methodology that depends on premises that are *inconsistent* with the claim that people are Bayesian about learning and inference. Bayesian models in cognitive science derive their appeal from their normative claim that the modeled inference is in some sense rational. Standard accounts of the rationality of Bayesian inference imply predictions that an agent selects the option that maximizes the posterior expected utility. Experimental confirmation of the models, however, has been claimed because of groups of agents that "probability match" the posterior. Probability matching only constitutes support for the Bayesian claim if additional unobvious and untested (but testable) assumptions are invoked. The alternative strategy of weakening the underlying notion of rationality no longer distinguishes the Bayesian model uniquely. A new account of rationality—either for inference or for decision-making—is required to *successfully* confirm Bayesian models in cognitive science.

1. Introduction

The past fifteen years have witnessed a dramatic growth in Bayesian models in cognitive science, driven both by computational and algorithmic advances, as well as experimental findings. Bayesian models of learning and inference have been proposed for

just about every major phenomenon in cognitive science.[1] Although directed towards different problems in each case, these models share a common structure: each assumes that every individual has an initial ("prior") probability distribution over the space of possibilities. This prior is updated in response to new data by conditionalization to yield new ("posterior") probabilities for each possibility. The computation, also known as *Bayesian updating*, uses Bayes theorem to determine the probability of each hypothesis H_i in a set of mutually exclusive and exhaustive hypotheses $H_1,\ldots,H_n$ given the evidence E:

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{P(E)}.$$

$P(E \mid H_i)$ is 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), while $P(H_i)$ is the prior probability distribution over the space of hypotheses. Their product is divided by the *a priori* probability of the evidence, $P(E)$, which essentially acts as a normalization term.[2] Despite their superficial simplicity, Bayesian models can produce quite intricate behavior, depending on the particular likelihood functions and prior probability distribution.

Bayesian models are appealing as models of human inference 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; (iv) explain gradual transitions in belief between various

[1] For a partial list of phenomena and references see the Appendix.
[2] $P(E)$ need not be computed if we are interested only in the relative probabilities of the various hypotheses. It is necessary to model the impact of rare (i.e., surprising) data.

hypotheses by the use of probabilities to represent degrees of belief; and (v) are supported by arguments that Bayesianism is rational, and hence any learner that acts in a Bayesian manner is rational as well.

Unlike the majority of models in cognitive science, Bayesian models are generally not taken to provide a mechanistic explanation of how some cognition or behavior is performed. That is, with a few exceptions[3], Bayesian models are understood to be wholly agnostic about the neural or cognitive bases of the observed behaviors. Instead, they are offered as rational analyses: *computational* level models (in contrast to *implementational* or *algorithmic* level models) that explain the inferences they describe as rational behavior in the given environment (Marr, 1982; Anderson, 1990; Chater & Oaksford, 2000; Chater, Tenenbaum & Yuille, 2006; Oaksford & Chater, 1998).[4]

Confirmation of the claim that people are Bayesian with regard to learning and inference depends on how the Bayesian belief update is expressed in measurable behavior. Although not always explicitly stated, this connection is generally taken to be provided by the principles of rational choice: namely, that an agent will select the option that maximizes the (posterior) expected utility. Since experiments designed to test Bayesian models are almost always designed to eliminate the effect of utilities on the

[3] Examples of Bayesian models with mechanistic commitments include Rao (2005), Lee & Mumford (2003), and Doya, Ishii, Pouget, & Rao (2007).

[4] Bayesian models and rational analyses are not coextensive in principle (Danks, 2008); there can be non-Bayesian rational analyses and Bayesian models that are not rational analyses. For example, standard reinforcement learning models are generally not Bayesian, but are rational in a wide range of environments; Bayesian models using limited hypothesis spaces are (in the absence of arguments about memory or computational limits) not necessarily rational. In practice, though, almost all Bayesian models are rational analyses, and the nature of the normative claim of rationality in these models will turn out to be one of the main sticking points for their confirmation.

participant's belief, any choice is taken to be indicative of the participant's posterior degree of belief. Thus, if experimental control of prior beliefs is successful, we should expect participants within the same experimental condition to make the same choices. This is not the case. Instead, the *distribution* of the participants' choices is found to resemble a *random sample* of the *model posterior distribution*. Despite this inconsistency with the rational choice predictions, the results are still taken to confirm the Bayesian claim. In the following sections we attempt to reconstruct how such a conclusion could be reached, starting in Section 2 with a concrete example where such "probability matching" is used as confirmation.

If, as we suggest, no resolution of this deep tension is forthcoming without much stronger commitments concerning the account of rationality, then we will have successfully undermined the explanation of the observed behavior the Bayesian account was supposed to provide: "(i) Behavior *B* is rational or optimal *and* (ii) there is some process—either ontogenetic or phylogenetic—that leads the individual to engage in rational or optimal behavior. Therefore, the individual exhibits behavior *B*." The first *necessary* premise would no longer have a foundation.

2. Confirmation of Bayesian Models

In a standard experimental set-up used to confirm a Bayesian model, experimental participants are provided with a cover story about the evidence they are about to see. This cover story indicates (either implicitly or explicitly) the possible hypotheses that could explain the forthcoming data. Either the cover story or pre-training is used to induce in participants a prior probability distribution over this space. Eliciting participants' prior

probabilities over various hypotheses is notoriously difficult, and so the use of a novel cover story or pre-training helps ensure that every participant has the same hypothesis space and nearly the same prior distribution. In addition, cover stories are almost always designed so that each hypothesis has equal utility for the participants, and so the participant should care only about the correctness of her answer. In many experiments, an initial set of questions elicits the participant's beliefs to check whether she has extracted the appropriate information from the cover story. Participants are then presented with evidence relevant to the hypotheses under consideration. Typically, in at least one condition of the experiment, the evidence is intended to make a subset of the hypotheses more likely than the remaining hypotheses. After, or sometimes even during, the presentation of the evidence, subjects are asked to identify the most likely hypothesis in light of the new evidence. This identification can take many forms, including binary or *n*-ary forced choice, free response (e.g., for situations with infinitely many hypotheses), or the elicitation of numerical ratings (for a close-to-continuous hypothesis space, such as causal strength, or to assess the participant's confidence in their judgment that a specific hypothesis is correct). Any change over time in the responses is taken to indicate learning in light of evidence, and those changes are exactly what the Bayesian model aims to capture.

These experiments must be carefully designed so that the experimenter controls the prior probability distribution, the likelihood functions, and the evidence. This level of control ensures that we can confirm the predictions of the Bayesian model by directly comparing the participants' belief changes (as measured by the various elicitation methods) with the mathematically computed posterior probability distribution predicted

by the model. As is standard in experimental research, results are reported for a participant *population* (split over the experimental conditions) to control for any remaining individual variation. Since the model is supposed to provide an account of each participant in the population *individually*, experimental results must be compared to the predictions of an aggregate (or "population") of model predictions. A comparison at the population level inevitably complicates any inference to the individual level. The following example will illustrate this point.

We focus throughout on (the relevant parts of) experiment 1 in Schulz, Bonawitz, & Griffiths (2007), but there is nothing special about this experiment. The confirmation methodology is applied in many other papers, and we could have given a structurally identical description for many other experiments (e.g., Kemp, Perfors & Tenenbaum, 2007; Xu & Tenenbaum, 2005, 2007). The Schulz, *et al*. experiments examine whether children will infer a mental cause for some physical effect given appropriate evidence, despite their strong prior beliefs against such cross-domain causation. Children are read two story books, each featuring seven days of an animal's life, where each day constitutes a piece of evidence. In the *within-domain* book, a deer develops itchy spots (E) every morning after being exposed to two potential physical causes (A, B, C, and so on), where one of the causes (A: "running through cattails") occurs every day. That is, the seven days of evidence have the form:

Day 1: A and B then E

Day 2: A and C then E

Day 3: A and D then E

…

In the *cross-domain* book, a similar story is presented with two candidate causes each day: one is physical and varies from day-to-day; one is mental ("feeling scared") and occurs every day. That is, the cross-domain evidence is structurally identical with the within-domain evidence, except that the recurring cause (A) is a mental state.

After reading a book, children are asked (forced choice) which of the potential causes on the last day actually caused the itchy spots. In the within-domain condition, they are thus presented with a choice between two within-domain causes; in the cross-domain condition, they face a choice between one within-domain and one cross-domain cause. In a separate baseline control group, the children are not exposed to any evidence, but instead proceed directly to the final choice. This baseline provides an indication of children's prior beliefs about potential within- and cross-domain causes of itchy spots; as expected, it revealed children's indifference between within-domain causes, and strong bias for within-domain causes compared to a cross-domain cause. In contrast, children observing purely within-domain evidence showed a strong preference for the recurring cause (A), and children observing cross-domain evidence showed a weaker (but significant) shift to preference for the cross-domain cause. That is, the evidence seems to have led children to change their beliefs to favor the recurring cause, even if the recurring cause was in a different domain than the effect.

Schulz, *et al*. (2007) provide a Bayesian model of these belief shifts. The precise mathematical details are not important; the model can be paraphrased as follows: Children's prior probabilities are highly skewed in favor of within-domain potential causes as against cross-domain causes, but the strong evidence in favor of a cross-domain cause (i.e., that this single factor can explain all of the observations) is sufficient to overcome this initial bias. Within-domain causes still retain a significant posterior probability in the cross-domain condition, but only because of their high prior probability. Schulz, *et al*. (2007) situate their particular model in a more general hierarchical Bayesian model of framework theories (Tenenbaum & Niyogi, 2003; Tenenbaum, Griffiths & Niyogi, 2007), but their framework theory simply serves to specify prior probabilities.

In the experiment, each individual participant is asked (using various methods) to select the most likely hypothesis about which factor caused the outcome (here, the itchy spots); this feature is widespread among experiments testing Bayesian models. Consider, however, what prediction is actually made by the Bayesian model. The model predicts simply that the participant starts with a prior probability distribution, and updates that distribution by conditioning on the evidence. Strictly speaking, the model is silent about how the participant makes a choice based on her posterior probability distribution. A Bayesian model of inference must therefore be supplemented by a suitable choice principle to make any empirical predictions. Despite the importance of choice strategies, very few Bayesian models in cognitive science are accompanied by explicit choice principles (though see, e.g., Körding & Wolpert, 2006; Oaksford, Chater, Grainger & Larkin, 1997).

Given a probability distribution over some hypothesis space, and given the plausible utilities, the choice strategy that maximizes utility will select the hypothesis with greatest probability (or if there are multiple hypotheses with maximal probability, it chooses one of them). Consider the empirical prediction implied by that choice strategy. In particular, suppose (i) the cover story has the desired effect such that every participant has the same hypothesis space, same prior probability distribution, and same utilities; and (ii) the Bayesian model of learning and inference is actually correct. In this idealized case, every participant should have the same posterior probability distribution. Thus, if participants actually maximize utility, then they should all choose exactly the same hypothesis. That is, one should expect to discover little variance in individual responses, and ideally no variance at all. The distribution of participant responses should be a narrowly peaked function with virtually all weight on the hypothesis with greatest probability. More specifically, the distribution of responses should not match the predicted individual posterior probability distribution particularly well (except for certain, very special, posterior distributions). In this ideal case, it does not matter whether the hypothesis space is finite (e.g., a set of possible categories) or infinite (e.g., possible causal strength ratings). In either case, if people are perfect Bayesian learners and choose rationally (and the experiment is well-designed), then every participant should respond (approximately) identically. To state the problem more precisely:

> If the individual response provides the hypothesis that maximizes the posterior and every individual has the same posterior over the hypothesis space, then (assuming uniform utilities) the individual posterior will be

different from the distribution of participant responses unless the individual posterior places all probability weight on its maximum points.

This type of heavily-peaked response distribution is not, however, what is usually found. Instead, almost all experiments find (though do not always describe it in these terms) that the distribution of individual responses resembles the posterior that the model predicts an individual to have. In the case of Schulz, *et al*. (2007), this is precisely the way the data are presented: Bayesian model predictions of an *individual's posterior probabilities* are shown to match the *distribution of choices* across the participant population. That is, the model posterior probability for each hypothesis is compared with the proportion of children that chose that hypothesis when asked to identify the correct cause. [5] More importantly, this method of confirmation is widespread for Bayesian models (including all of the papers cited at the beginning of this section). These analyses thus imply the puzzling conclusion that the population *as a whole* acts as a rational Bayesian learner, but the individual learners do not.[6]

---

[5] For our purposes here the relevant point is that the match of the response distribution with the model prediction is considered the relevant criterion to assess fit. We do not intend to argue here about whether or not the model actually constitutes a good fit. Visually there are some quite clear discrepancies between model prediction and data in the cross domain condition, but the authors claim their "model accurately predicted […] with a Pearson product-moment correlation coefficient of r(9)=.85."

[6] There are some notable exceptions that do actually try to compare each individual participant's responses with the predictions of a Bayesian model. For example, Körding & Wolpert (2006) directly model each individual learner's prior probabilities and subsequent inferences. Steyvers, *et al*. (2003) model learning given participant-chosen interventions on an individual basis. Tenenbaum & Griffiths (2003) do not directly model individuals, but do obtain probability judgments that can be compared to the predicted posterior probability distribution (assumed to be the same for all individuals). Whether these analyses have provided more support for the Bayesian models is, we believe, open to question. The concern about the methodology of the more typical methods discussed here remains, independently.

As a concrete example of the problem, suppose we have three hypotheses and the predicted posterior probability distribution is $P(H_1) = 0.3$; $P(H_2) = 0.3$; and $P(H_3) = 0.4$. Hypothesis H_3 is slightly more likely than the other two. The rational forced choice for this posterior probability distribution (modulo our previous remarks about utilities) is "H_3". Thus, if the Bayesian model is correct and participants are choosing rationally, then we should expect everyone to respond H_3. But that is not what happens in these experiments; rather the common "confirmation" of the Bayesian model checks whether 30% of the participant population chooses H_1, 30% chooses H_2, and 40% chooses H_3.

The problem is that the following four propositions are jointly inconsistent:

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

3. Rationality of Bayesian Models

As we noted in the introduction, almost all Bayesian models can only contribute to explanations of observed behavior because of their (putative) status as rational models, and so must provide an account of the rationality of the behavior and inference. If a Bayesian model is not rational, then it cannot provide even the first premise in a rationality- or optimality-based explanation. The Bayesian model itself only constrains

the belief update in light of evidence, and so its empirical testability (and overall rationality) depends on an (often implicit) account of the connection between the updated beliefs, and the reported choice or exhibited behavior. In the previous section, we suggested that Bayesian models fit most naturally with a choice principle such as maximizing expected utility (proposition 2 above), but gave no particular justification for that claim. In this section, we explore in more detail the sense in which Bayesian models, and possible accompanying choice principles, can be understood as 'rational.'[7]

One standard defense of the rationality of Bayesian belief updating that is often mentioned in the psychological literature is diachronic probabilistic coherence, understood as avoidance of diachronic Dutch books. In general, a commonly accepted constraint on rationality is that a reasoner is rational only if there is no set of bets that she would accept for which she would be guaranteed to lose no matter what the outcome of the propositions bet upon; i.e., there is no "Dutch Book." Synchronic Dutch Book arguments show that only degrees of belief satisfying the axioms of probability provide betting odds against which no Dutch Book can be made at any single time (de Finetti, 1937; Ramsey, 1931). Diachronic Dutch Book arguments aim to show a similar result for changes of belief over time: namely, that only Bayesian updating, or some equivalent method, guarantees that there is no sequence of bets that guarantee a loss (Teller, 1973, 1976). The Dutch book defense of rationality connects in an obvious way the optimal belief update (Bayesian inference) with a choice principle based in standard decision

[7] Recall that 'Bayesian model' refers (in the cognitive science community) to a model of Bayesian belief updating. Throughout this section, we share this focus, and so will ignore the many arguments for synchronic Bayesianism: the theory that degrees of belief are (or should be) given by a coherent subjective probability distribution. Our issue here is only with the correct way to *change* one's beliefs over time, not whether synchronic Bayesianism is the correct way to understand degrees of belief at some moment in time.

theory: choose the action that maximizes subjective expected utility. The problem is, as we showed in the previous section, that its predictions do not fit the empirical data from cognitive science in any obvious way.

There are several independent reasons why one might view diachronic Dutch book arguments with suspicion, as they seem overly restrictive (Levi, 1988, 2002; Maher, 1992; van Fraassen, 1984). For example, diachronic Dutch book arguments require a reasoner to make firm and unchangeable *conditional* commitments (i.e., commitments about degrees of belief given any arbitrary evidence) at the outset of inquiry. While various proposals have been made to adjust the diachronic arguments in response to those concerns, it suffices for our purposes here to note that all of these weakenings of the diachronic Dutch book argument result in the conclusion that Bayesian belief updating is only one of many protections against irrationality. That is, if the conditions for diachronic rationality are weakened in plausible ways (e.g., if reasoners are permitted to "look ahead" before accepting or declining bets), then Bayesian updating no longer has a privileged position as *the* rational method of belief updating, since other (non-Bayesian) forms of belief update have similar normative grounds (e.g., Douven, 1999). Additional arguments would then be required to distinguish Bayesian updating as the *best* rational explanation.

One could instead try to justify the rationality of Bayesian updating by appeal to its long-run properties, though this reason is rarely cited in the cognitive science literature as a convincing argument (see Perfors et al (in press) for an exception). In particular, a plausible necessary condition on any rational belief change method is that it should converge to the truth when possible (though the method might also value, e.g., short-run

predictive accuracy). Given a resolution to some technical issues,[8] one can prove that Bayesian updating converges to the truth (when possible)[9] and is provably not dominated in speed of convergence by any other method.[10] This observation leaves unanswered the question of what choice of hypothesis is rational in situations in which the distribution has not converged, which is presumably the case in most experimental situations.[11] We still require an additional account of how a particular choice is made in light of a non-trivial posterior probability distribution. An obvious candidate would again be to choose the hypothesis that maximizes the posterior probability distribution. But as we found

[8] One natural assumption is that learners should only invoke computable functions, but this constraint is sometimes incompatible with the requirement (on synchronic Dutch book grounds) that a Bayesian reasoner know all (relevant) logical and mathematical implications of the various hypotheses that she entertains (Gaifman & Snir, 1982). Moreover, 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). In these circumstances the Bayesian reasoner only converges to the truth (whenever the truth can be learned) if she can sometimes "compute" uncomputable functions.

[9] If the true hypothesis H is empirically distinguishable from other hypotheses and $P(H) \neq 0$, then for all ε, P(Bayesian reasoner has degree of belief greater than $1-\varepsilon$ in the truth) $\rightarrow 1$ as the number of datapoints goes to infinity; see Savage (1972) for a canonical expression of this result.

[10] More precisely, Bayesian updating (for any non-dogmatic prior probability distribution) provably satisfies the condition that 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 methods that get to the truth faster in particular worlds, but none outperforms Bayesian updating in every world (Schulte, 1999). There are, however, methods that are similarly non-dominated by Bayesian updating, so any argument along these lines does not identify Bayesian inference as uniquely rational.

[11] Virtually no experiment that uses Bayesian models has a degenerate posterior distribution. If the Bayesian model is supposed to provide a normatively correct description of the belief update of an individual, then it follows that the experiment is explicitly considering circumstances in which the distribution over the hypotheses has not converged. Moreover, much of the appeal of Bayesian models (in contrast to logic-based models) results from the ability to describe shifts in degree of belief that are not complete, i.e. where uncertainty over the true hypothesis remains.

earlier, this combination of belief update and choice principle is inconsistent with the empirical results.

The two standard defenses of the rationality of Bayesian updating offered in the cognitive science literature both imply that the rational choice principle is to choose the hypothesis with maximal expected utility. In essentially all psychological experiments, that choice principle implies choosing the hypothesis with maximal probability, but that implication is inconsistent with empirical data, and so the "maximize expected utility" choice principle is unavailable to the proponent of Bayesian models in cognitive science.

In light of this tension, various weaker notions of rationality that explicitly deny the existence of an overarching general formal account of rationality have been suggested in the cognitive literature (e.g., in Oaksford & Chater, 2007). These accounts focus on rationality as successful, goal-directed action, where the constraints used to judge this particular 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 in the given situation. 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. Put crudely, such approaches claim that rationality of type A with normative principles P_1, …, P_k applies in situations of type S, and rationality of type B with normative principles Q_1, …, Q_n applies in situations of type S', and rationality A and rationality B need not

have anything more in common than identifying optimal behavior *relative to their respective* principles.

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' (see also Evans, 2009; Khalil, 2009).

The dilemma we face is that none of the standard accounts of the rationality of Bayesian belief update naturally predict empirical data in which the response distribution matches the model posterior. We thus consider the possibility that rationality requires something other than maximizing expected utility (i.e., rejecting proposition 2 above), or that experimental controls are perhaps less successful than is typically thought (i.e., amending proposition 3 above).

4. Alternative Reponses

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. The Luce choice axiom that characterizes a pair of constraints on human choice implies under fairly weak assumptions that an option is selected in proportion to its weight (Luce, 1959, 1977).[12] Assuming that experiments appropriately control the utilities and other saliencies across the available hypotheses, it then follows from the Luce choice axiom that a hypothesis is selected according to its posterior probability, i.e. it is probability matched. The axiom should not be mistaken for an axiom of *rational* behavior. It does not itself explain why probability matching constitutes the normatively correct behavior. Such an account would have to describe why the conditions of the Luce choice axiom are hallmarks of rationality. In particular, its second condition concerning "choice probabilities" effectively begs the question of the rationality of probability matching.

In practice, the content of the claim that hypotheses are probability matched 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 a hypothesis using a method that is equivalent to taking a random sample from her posterior

---

[12] 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. $P_R(a) = P_S(a \mid R)$ for all *a* in *R*).

Luce (1959) shows how the axiom implies that $P_S(x) = v(x) / \Sigma_{y \text{ in } S}\, v(y)$, where v(.) is a measure of value or weight over the options y in the choice set S.

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.

Claim (B) says that probability matching is, for reasons of fit-to-data, the proper choice principle to be tacked onto the Bayesian belief update. Without further explanation, however, this response removes (almost) all of the normative justification for the Bayesian model of inference. The claim being defended in this response is that humans use Bayesian updating for learning and inference, and then probability match their current posterior distribution when asked to make a choice (given equal utilities). But without an account of the rationality of probability matching, the proponent can no longer claim that people are approximately rational Bayesians. If the aim is to make the right choice as often as possible, then probability matching is provably sub-optimal and so usually thought to be "irrational" or a "bias" (e.g., Shanks, Tunney & McCarthy, 2002; Vulkan, 2000; West & Stanovich, 2003). For example, if we have a coin that is biased with probability 0.7 towards heads, we should *always* bet on heads, and not probability match (i.e., bet heads on 70% of trials). The former strategy can expect to win 70% of the time; a probability matching strategy can only expect to win in 58% of trials.

Thus, a model with a Bayesian belief update and a probability matching choice strategy implies that people are rational learners, but to no particular point, since they are irrational decision makers. Thus, a host of questions arise: In what sense is such a process rational? What efficiency or optimality could probability matching provide for a

learner?[13] Why would probability matching develop in the first place, rather than maximizing the posterior?

The "Bayesian inference + probability matching choice procedure" hypothesis might be descriptively correct, but without substantial additions to the account, there is no justification for describing such a cognitive agent as rational, and there are natural arguments that she actually is irrational. Moreover, there is a peculiar internal tension in models of this type, as they claim that people (approximately) compute the quite difficult Bayesian updates, but then fail to use a computationally quite simple choice strategy. If people can (approximately) carry out the first computations, it would be surprising (though not impossible) if they were unable to use the optimal choice method.

This leaves two alternatives. Either (I) one finds circumstances in which it is optimal or rational for an individual to take a random sample from her posterior, and then demonstrates that these circumstances apply in the experimental conditions. This response would amount to a rejection of proposition 2 (from our earlier list of four), that choices are made by maximizing expected utility. Or (II) one holds that probability matching is only descriptive of *population* behavior (claim A), and then shows how this population-level probability matching can arise from rational Bayesian learners. As we will later show, any argument endorsing option II maintains claim 2 at the cost of a weakening of claim 3, that experiments successfully constrain participants' prior beliefs and utilities. We start, however, with option I.

[13] This question is particularly pressing, since rational analyses must ultimately provide a developmental story (in phylogenetic time, ontogenetic time, or both) that exploits some optimality property of the model. Such a story seems much less plausible if people are probability matching.

4.1 Changing the circumstances

There are several different avenues for trying to characterize circumstances in which probability matching is rational, but the most obvious candidates are circumstances with competition and resource constraints (Stephens & Krebs, 1986).[14] 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, however, resources are constrained by competitors, then there are scenarios in which the optimal strategy is to go to location X 75% of the time, and to location Y 25% of the time. That is, one should probability match. In particular, when all individuals have the same utilities for the resources, the resource payoffs for each location are based on competition (i.e., more individuals at a location means fewer resources for everyone), and no communication is possible, then probability matching can form a Nash equilibrium: if everyone probability matches, then there is no incentive for any individual to unilaterally change her strategy.[15] Qualitatively, the idea is simply that randomizing can decrease the likelihood that we directly compete for a resource at a particular location (or at least,

[14] An alternative suggestion is made in Fiorina (1971), who 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.

[15] We have omitted many technical details, including some additional necessary assumptions. Interested readers should consult any standard book on optimal foraging theory (e.g., Stephens & Krebs, 1986) or see the appendix for a simple concrete example. We argue that this response is unsuccessful, and so these additional constraints on its scope are irrelevant for our present purposes.

make my success not depend on your strategy). To the extent that strategies implied by Nash equilibria are rationally justifiable (and there is debate about this), then probability matching is rational in these types of circumstances.

Our present focus, however, 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 inference and hypothesis learning implies that an experimenter's efforts to remove or control for aspects that might induce competitive behavior are futile, and so the 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, and if they do, why a suitable competitive payoff structure is appropriate.

A different argument (though to our knowledge not published in the cognitive literature) 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).[16] 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 error-rate of an algorithm on a test sample 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 participants 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.

We thus fall back on option II: probability matching results from other phenomena or inaccuracies in the individual computation (rather than actual individual probability matching) and is therefore only descriptive of the group-level behavior.

[16] Josh Tenenbaum (personal communication) suggested that PAC-Bayesian approaches provide a potential solution to the dilemma we point to and so we address the proposal here. A thorough description of PAC-Bayesian learning would go far beyond the scope of this article. We only aim here to indicate the reasons why we do not think this is a fruitful approach.

4.2 Explaining group-level probability matching

Since option II attributes probability matching to differences between participants, this proposal inevitably implies either a rejection that people are Bayesian, or a weakening of the assumption of successful control in the experiments (i.e., proposition 3). We assume the prior option is unavailable to the committed Bayesian. There are both strong and weak ways of rejecting proposition 3. A strong rejection 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.[17] 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

[17] We note that *any* rejection of proposition 3 that exclusively focuses on the prior faces a significant formal challenge. Suppose N participants have priors $P_1(H)$, …, $P_N(H)$. Assume also that all individuals have the same likelihoods ($P(D \mid h)$ for all h in H), and that they 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. $P_{pop}(h) = \#_N \; argmax_H \; P_i(h) \; / \; N$. If all individuals use Bayesian updating, then the "population posterior" (the final aggregate response distribution) is $P_{pop}(H \mid D) = \#_N \; argmax_H \; P_i(h \mid D) \; / \; N$.

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: $P_{pop}(H \mid D)$ must be distributionally equivalent to $P(D \mid H)P_{pop}(H)/P(D)$. Mathematically, this holds when: $\#_N \; argmax_H \; P(D \mid h)P_i(h) \approx P(D \mid H) \; \#_N \; argmax_H \; P_i(h)$. For arbitrary likelihoods, this condition is not satisfied for standard prior distributions (e.g., flat or Gaussian), although it may be satisfied for such priors given particular likelihoods $P(D \mid H)$. (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.

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.

A weak rejection of proposition 3 would 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. Random utility maximization provides one such proposal. 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 observed probability matching as a result of aggregation of optimal responses from a population. 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, $V(.)$ is a non-stochastic function representing a utility function shared by every member of the population, and $e(.)$ is stochastic, representing the individual's deviation from the population utility $V(.)$. 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 variations[18], McFadden shows that choice

[18] If e(.) 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 $\exp(U(H_i)) / \sum_j \exp(U(H_j))$.

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*(.), and sensible monotonic functions *f*(.), 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 function (whether utility or probability) over a space of options (including hypotheses) but have independent individual deviations (of a certain kind) from this trend, then the population-level response distribution when each individual selects the maximum of her individual function is a distribution that has the same maxima and minima as the population function. Moreover, the individual deviations can 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 potentially explains the appearance of probability matching at the population level as the result of individual differences within the population. In a McFadden-style response, the functional relation between the response distribution and

---

The Weibull distribution is sufficient; weaker constraints on the distribution of e(.) can be given.

the population function (the model prediction) depends on the precise distribution of individual discrepancies. For the specific distribution of individual variations in McFadden (1974), 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. Since identity is the standard confirmatory test for Bayesian models (as stated or implicit in the papers we have cited[19]), that individual variation distribution is not appropriate. It is unknown whether there is a plausible distribution for individual deviations that implies identity between the model posterior and the response distribution. More generally, any use of a McFadden-style response depends crucially on establishing, on *independent* grounds, either that the individuals in the population exhibit the suitable variation, or that two populations exhibit specific differences in variation. Such testing has, to our knowledge, never been done in the psychological literature on Bayesian models. Nevertheless, this proposal is of great interest, since the necessary sources of variation potentially exist in psychological experiments, and the proposal lends itself to precise empirical predictions.

4.3 Solving a different task

An alternative response—which could also be combined with the optimal betting or PAC-Bayesian response—argues that problem solving and learning strategies are automated and optimized for everyday use to such an extent that the artificial settings of

[19] For example, Schulz, *et al*. (2007), which we discussed in Section 2, claim in their conclusion: "The Bayesian model presented in the Appendix [in their paper] … show[s] that the judgments of the four- and five-year-olds in our experiments are close to the probabilities entertained by an ideal Bayesian learner using a particular causal theory."

psychological experiments are unable to control and focus the subject on the best strategy for the task of the experiment. Despite an experimenter's best efforts at developing an ecologically valid experiment, and despite debriefing reports by participants 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 the participant response is not an indication of how she solved the task at hand, but rather an indication of which ingrained problem solving strategy was triggered by this task. 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 is ultimately unsatisfactory, however, as it 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?

In particular, question (a) brings us full circle back to where we started: the standard of rationality.

4.4 A return to behaviorism?

Without an account of the rationality of the observed input-output relation, the computational level models provide a summary of the observed data, but no rational explanation for the behavior. That is, their models begin to look very similar to those advocated by (methodological) behaviorists: the model happens to capture some important behavioral regularities, but implies no commitments beyond that. Such an

interpretation has been (informally) offered for Bayesian models. A "data summary" Bayesian model is principally (but perhaps not solely) of interest when it is the most compact or efficient representation of some empirical phenomena. In many cases, however, comparable levels of empirical fit can be achieved by simpler models with less conceptual baggage (e.g., Chater, Oaksford, Nakisa & Redington, 2003; Gigerenzer, Czerlinski & Martignon, 1999; Nelson, 2009). Thus, some alternative argument must be provided for their use.

The most plausible such justification is the widespread applicability of Bayesian models: the existence of Bayesian models for a wide range of domains provides some measure of unification to those disparate domains (assuming the method of confirmation is more successful in these domains). This argument, however, only has non-pragmatic force when the widespread applicability arises because of some similarity in the underlying mechanisms that generate the phenomena, or the function/role of the phenomena in a containing system. In either case, stronger commitments than simple unification can (and should) be made. And if there is no such shared mechanism or function/optimality, then there is no particular reason to prefer Bayesian models as data summaries. Moreover, the idea that Bayesian models *should* be only data summaries is 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. The suggested restriction would amount to giving up on that hard-won right.

5. Conclusion

Our assessment of the standard confirmation methods for Bayesian models in cognitive science has been largely negative. But we are explicitly not concluding that Bayesian models are not testable, or that the use of Bayesian models in cognitive science is inevitably misguided; such claims would require an exhaustive exploration of the space of possible models, and would constitute, in light of evidence from several other experimental techniques that appear to support parts of the Bayesian paradigm, a surprising rejection of the whole methodology. Our conclusion is more conservative, and largely methodological. All of the problems we have raised can ultimately be traced back to insufficient specification of the Bayesian models, either in the justification that they are 'rational,' or in the choice mechanism that maps a posterior probability distribution to a behavior. Full specification of the models is, in a certain sense, risky: such specified models are more readily disconfirmed and rejected. Such specification is, however, absolutely necessary if the models are to help us understand the nature of cognition. Commitment to a non-trivial account of rationality with a fully specified choice procedure would turn experiments like the one described in Section 2 into proper tests of Bayesian models. An alternative solution, such as random utility maximization (McFadden, 1974), could provide a resolution of our puzzle that maintains the standard view of rationality in terms of Bayesian belief update combined with rational choice. And just as it adds plausible commitments, it also provides further constraints for new types of tests for Bayesian models. Alternatively, an account of rationality that is weaker than Dutch Book rationality, but is sufficiently concrete to satisfy the explanatory demands of a computational level model (i.e., does not imply a large number of "rational" inference

methods), would almost certainly be of interest well beyond the confirmation of Bayesian models. 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.

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 has no substantial explanatory power. Thus, a Bayesian model must be rational for it to play a role in any substantive explanation at all. As we argued in Section 3, however, the standard accounts of rationality fail to account for the behavior in the empirical data. We thus reached our four claims that are jointly inconsistent:

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

The question that remains open is: If one is committed to claim 1, what can be given up?

Appendix

Recent books on Bayesian models include: Oaksford & Chater (2007) with Open Peer Commentary (Oaksford & Chater, 2009); Doya, Ishii, Pouget, & Rao (2007); and Chater & Oaksford (2008). Bayesian models were also the focus of a 2006 special issue of *Trends in Cognitive Science* (Chater et al., 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 & Tenenbaum, 2003; Kemp et al., 2007; Tenenbaum & Griffiths, 2001)
- Causal learning and reasoning (Bonawitz, Griffiths & Schulz, 2006; Griffiths & Tenenbaum, 2005; Schulz et al., 2007; Sobel & Kushnir, 2006; Sobel, Tenenbaum & Gopnik, 2004; Steyvers, Tenenbaum, Wagenmakers & Blum, 2003)
- Inference about conditionals (Oaksford & Chater, 2007; Oaksford, Chater & Larkin, 2000)
- Covariation assessment (McKenzie & Mikkelsen, 2007)
- Imitation (Rao, Shon & Meltzoff, 2004)
- Information selection (Oaksford et al., 1997)
- Framing effects (McKenzie, 2004)
- Memory effects (Schooler, Shiffrin & Raaijmakers, 2001; Shiffrin & Steyvers, 1997; Steyvers & Griffiths, 2008)
- Object perception (Kersten & Yuille, 2003; Kersten, Mamassian & Yuille, 2004)
- Repetition effects and priming (Mozer, Colagrosso & Huber, 2002, 2003)

- Word learning (Xu & Tenenbaum, 2005, 2007)

Example of optimality of probability matching strategy in competitive circumstances

Suppose that a resource occurs with probability 0.75 at location X and with probability 0.25 at locations Y. Further, suppose that for two competitors the pay-off is structured such that a competitor only obtains resources if (i) resources are present at the chosen location *and* (ii) the other competitor did not select the same location. The pay-off matrices for the two possible locations of resources are shown below:

Resource is at location X (75% of cases)	Competitor 2 selected location X	Competitor 2 selected location Y
Competitor 1 selected location X	0 / 0	1 / 0
Competitor 1 selected location Y	0 / 1	0 / 0

Resource is at location Y (25% of cases)	Competitor 2 selected location X	Competitor 2 selected location Y
Competitor 1 selected location X	0 / 0	0 / 1
Competitor 1 selected location Y	1 / 0	0 / 0

In foraging theory the optimal strategy is generally taken to be the one that maximizes the average pay-off over several foraging trials (Stephens & Krebs, 1986). We can thus combine the two possible payoff structures by weighting them according to the rate of occurrence of the resource at each location:

Combined pay-off structure	Competitor 2 selected location X	Competitor 2 selected location Y
Competitor 1 selected location X	0 / 0	0.75 / 0.25
Competitor 1 selected location Y	0.25 / 0.75	0 / 0

If both competitors select location X with probability 0.75 and location Y with probability 0.25 then no unilateral change of strategy by a competitor would improve that competitor's payoff, that is, these two strategies that "match the probabilities" of the occurrence of the resource at the two locations specify a Nash equilibrium.[20] Note, however, that the optimality of the probability matching strategy crucially depends on the combination of the resource distribution and the *specific* resource payoff structure. In particular, if competitors can split resources when they both select the same location, then probability matching no longer constitutes a Nash equilibrium.

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. We also received valuable feedback from several anonymous reviewers. The first author was partially supported by a grant from the James S. McDonnell Foundation Causal Learning Collaborative. The second author was partially supported by a James S. McDonnell Foundation Scholar Award.

[20] If competitor *C1* selects location *X* with probability *p* and competitor *C2* selects location *Y* with probability *q*, then *C1* receives 0.75*(1-*q*) for location *X* and 0.25*q* for location *Y*. At equilibrium these have to be equal, hence *q* = 0.75. The argument is the same for *p* since the payoff structure is symmetric.

References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

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.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*, 93-131.

Chater, N., & Oaksford, M. (eds.). (2008). *The probabilistic mind: Prospects for Bayesian cognitive science.* Oxford: Oxford University Press.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes*, *90*, 63-86.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 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*(7), 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.

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. (2007). *Bayesian brain: Probabilistic approaches to neural coding.* Cambridge, MA: The MIT Press.

Evans, J. S. B. T. (2009). Does rational analysis stand up to rational analysis? *Behavioral and Brain Sciences*, *32*(1), 88-89.

de Finetti, B. (1937). La prevision: Ses lois logiques, se sources subjectives. *Annales De L'institut Henri Poincare*, *7*, 1-68.

Fiorina, M. P. (1971). A note on probability matching and rational choice. *Behavioral Science*, *16*, 158-166.

Gaifman, H., & Snir, M. (1982). Probabilities over rich languages, testing, and randomness. *Journal of Symbolic Logic, 47*, 495-548.

Gigerenzer, G., Czerlinski, J., & Martignon, L. (1999). How good are fast and frugal heuristics? In J. Shanteau, B. Mellers, & D. Schum (Eds.), *Decision research from Bayesian approaches to normative systems.* (pp. 81-103). Norwell, MA: Kluwer.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334-384.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford, & N. Chater (Eds.), *Rational models of cognition.* (pp. 248-74). New York: Oxford 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., & Tenenbaum, J. B. (2003). Theory-based induction. In *Proceedings of the 25th annual conference of the cognitive science society.*

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307-321.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*, 1-9.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271-304.

Khalil, E. L. (2009). Are stomachs rational? *Behavioral and Brain Sciences*, *32*(1), 91-92.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319-326.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434-1448.

Levi, I. (1988). The demons of decision. *Monist*, *70*, 193-211.

Levi, I. (2002). Money pumps and diachronic books. *Philosophy of Science*, *69*, S235-S247.

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*(3), 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. 164-70). New York: ACM.

McAllester, D. A. (2003). PAC-Bayesian stochastic model selection. *Machine Learning*, *51*(1), 5-21.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics.* (pp. 105-42). New York: Academic Press.

McKenzie, C. R. M. (2004). Framing effects in inference tasks-and why they are normatively defensible. *Memory & Cognition*, *32*(6), 874-885.

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 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. 51-7). 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.

Nelson, J. D. (2009). Naive optimality: Subjects' heuristics can be better motivated than experiments' optimal models. *Behavioral and Brain Sciences*, *32*(1), 94-95.

Oaksford, M., & Chater, N. (eds.). (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. (2009). Open peer commentary and authors' response. *Behavioral and Brain Sciences*, *32*(1), 85-120.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*(4), 883-899.

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*(2), 441-458.

Osherson, D. N., Stob, M., & Weinstein, S. (1988). Mechanical learners pay a price for Bayesianism. *Journal of Symbolic Logic, 53*, 1245-1251.

Perfors, A., Tenenbaum, J., Griffiths, T.L., Xu, F. (in press) A tutorial introduction to Bayesian models of cognitive development. *Cognition*.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays.* (pp. 156-98). London: Harcourt, Brace & Co.

Rao, R. P. N. (2005). Bayesian inference and attentional modulation in the visual cortex. *Cognitive Neuroscience and Neuropsychology*, *16*(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.

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*(1), 257-272.

Schroyens, W. (2009). On is and ought: Levels of analysis and the descriptive versus normative analysis of human reasoning. *Behavioral and Brain Sciences*, *32*(1), 101-102.

Schulte, O. (1999). The logic of reliable and efficient inquiry. *Journal of Philosophical Logic*, *28*, 399-438.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared make your tummy ache? Naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology*, *43*(5), 1124-1139.

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., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition*, *34*(2), 411-419.

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. (2008). Rational analysis as a link between human memory and information retrieval. In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science.* (pp. 329-49). Oxford: Oxford University Press.

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*(7), 327-334.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & 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.* (pp. 205-59). Dordrecht: Reidel.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.

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. 1152-7). Mahwah, NJ: Erlbaum.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-Based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309-318.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik, & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* (pp. 301-22). Oxford: Oxford University Press.

van Fraassen, B. C. (1984). Belief and the will. *Journal of Philosophy*, *81*(5), 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*(2), 243-251.

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*(3), 288-297.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301-308.