Bayesian Modeling of the Mind: From Norms to Neurons

Michael Rescorla

Abstract: Bayesian decision theory is a mathematical framework that models reasoning and decision-making under uncertain conditions. The past few decades have witnessed an explosion of Bayesian modeling within cognitive science. Bayesian models are explanatorily successful for an array of psychological domains. This article gives an opinionated survey of foundational issues raised by Bayesian cognitive science, focusing primarily on Bayesian modeling of perception and motor control. Issues discussed include: the normative basis of Bayesian decision theory; explanatory achievements of Bayesian cognitive science; intractability of Bayesian computation; realist versus instrumentalist interpretation of Bayesian models; and neural implementation of Bayesian inference.

1. INTRODUCTION

Bayesian decision theory is a mathematical framework that models reasoning and decision-making under uncertain conditions. The framework --- initiated by Bayes (1763), systematically articulated by Laplace (1814/1902), modernized by Ramsey (1931) and de Finetti (1937/1980), equipped with a secure mathematical foundation by Kolmogorov (1933/1956), and further elaborated by Jeffreys (1961) and Savage (1974) --- figures prominently across a range of scientific disciplines. Bayesian decision theory is a *normative* enterprise: it addresses how people *should* reason and make decisions, not how they *actually* reason and make decisions.

Nevertheless, many authors contend that it describes some mental activity with at least some degree of accuracy (e.g. Arrow, 1971; Davidson, 1980; Luce & Suppes, 1965). Over the past few decades, Bayesian theorizing has flourished within cognitive science. This research program uses the Bayesian framework to build and test precise mathematical models of the mind. Researchers have offered Bayesian models for an array of psychological domains, achieving particularly notable success with perception and motor control. Ongoing neuroscientific work investigates the mechanisms through which the brain (approximately) implements Bayesian inference.

2. BAYESIAN DECISION THEORY

The core notion of Bayesian decision theory is *credence*, or *subjective probability* --- a quantitative measure of the degree to which an agent believes an hypothesis. I may have low credence that a meteor shower occurred five days ago, higher credence that Seabiscuit will win the race tomorrow, and even higher credence that Emmanuel Macron is French. An agent's credence in hypothesis h is notated as $p(h)$ and is assumed to fall between 0 and 1. Credences are psychological facets of the individual agent, not objective chances or frequencies out in the world. The agent's credences need not track any *objective probabilities* that inhere in mind-independent reality. To illustrate, suppose that a biased coin has objective chance .3 of landing heads. I may mistakenly believe that the coin is fair and therefore assign subjective probability .5 to the hypothesis that it will land heads. Then my credence departs dramatically from the objective chance of heads.

Bayesian decision theory hinges upon *synchronic norms* governing how an agent should allocate credences over hypotheses. The probability calculus axioms codify these norms (Box 1). Also crucial for the Bayesian framework is the notion of *conditional probability* $p(h \mid e)$: the

probability of *h* given *e*. When $p(e) > 0$, we can explicitly define conditional probability in terms of unconditional probability. When $p(e) = 0$, no explicit definition is in general possible. To handle such cases, we must take conditional probability as a primitive notion subject to further axiomatic constraints (Box 2).

INSERT BOX 1 AND BOX 2 ABOUT HERE

Credences evolve over time. For example, if I learn that Seabiscuit is recovering from an illness, then I may lower my credence that he will win the race. *Conditionalization* is a norm that governs how credences should evolve over time. The basic idea behind Conditionalization is that, upon learning *e*, you should replace your former credence $p(h)$ with $p(h \mid e)$. Thus, your old conditional credence $p(h \mid e)$ becomes your new *unconditional* credence in *h*. $p(h)$ is called the *prior probability* and $p(h \mid e)$ is called the *posterior probability*. The literature offers various subtly different ways of formulating Conditionalization more precisely (Meacham, 2016; Rescorla, forthcoming a).

Bayes's theorem is an extremely useful tool for computing the posterior $p(h \mid e)$. The theorem states that:

$$p(h \mid e) = \frac{p(h)\, p(e \mid h)}{p(e)},$$

when $p(e) > 0$. This formula expresses the posterior in terms of the prior probabilities $p(h)$ and $p(e)$ and the *prior likelihood* $p(e \mid h)$. In most applications, $p(e)$ figures only as a normalization constant to ensure that probabilities sum to 1, so it is common to write the theorem as:

$$p(h \mid e) = \eta\, p(h)\, p(e \mid h),$$

where $\eta = \frac{1}{p(e)}$. A generalized analogue to Bayes's theorem obtains in many cases where $p(e) = 0$ (Schervish, 1995, pp. 16-17). There are also cases (including some cases that arise in Bayesian cognitive science) where $p(e) = 0$ and no generalized analogue to Bayes's theorem is available (Ghosal & van der Vaart, 2017, pp. 7-8).[1]

Bayes's theorem must be sharply distinguished from Conditionalization. Bayes's theorem is a direct consequence of the probability calculus axioms. As such, it is purely *synchronic*: it governs the relation between an agent's current conditional and unconditional credences. In contrast, Conditionalization is a *diachronic* norm. It governs how the agent's credences at an earlier time relate to her credences at a later time. Any agent who conforms to the probability calculus axioms also conforms to Bayes's theorem, but an agent who conforms to the probability calculus axioms at each moment may violate Conditionalization. Thus, one cannot derive Conditionalization from Bayes's theorem or from the probability calculus axioms. One must articulate Conditionalization as an additional constraint upon credal evolution.[2]

The final key notion of Bayesian decision theory is *utility*: a numerical measure of how much an agent desires an outcome. According to Bayesians, an agent should choose actions that maximize *expected utility*. The expected utility of action *a* is a weighted average of the utilities assigned to possible outcomes, where the weights are probabilities contingent upon performance of action *a*. How to formulate expected utility maximization more rigorously is a topic of extended debate (Steele & Stefánsson, 2016).

[1] For example, *Dirichlet process priors* do not admit a formula analogous to Bayes's theorem (Ghosal and van der Vaart, 2017, pp. 59-101). Dirichlet process priors play an important role in nonparametric Bayesian statistics, and they also figure in Bayesian cognitive science (Navarro et al., 2005).

[2] In the scientific literature, the phrase "Bayes's Rule" is used sometimes to denote Conditionalization, sometimes to denote Bayes's theorem, and sometime to denote an admixture of the two.

Bayesian decision theory (comprising the probability calculus axioms, Conditionalization, and expected utility maximization) has found fruitful application within statistics (Gelman et al., 2014), physics (Trotta, 2008), robotics (Thrun, et al., 2005), medical science (Ashby, 2006), and many other disciplines. There are other mathematical frameworks for modeling uncertainty (Colombo et al., forthcoming), but no other framework approaches the Bayesian framework's sustained record of achievement across so many disciplines.

Box 1. The probability calculus

In Kolmogorov's (1933/1956) codification of the probability calculus, probabilities are assigned to sets of possible outcomes. For example, suppose we seek to define probabilities over possible results of a horse race. We can specify an outcome by describing the order in which the horses finish. The hypothesis that Seabiscuit wins the race is the set of outcomes in which Seabiscuit finishes before every other horse. The hypothesis that Seabiscuit does not win the race is the set of outcomes in which Seabiscuit does not finish before every other horse. Three axioms govern the assignment of probabilities to hypotheses:

- Probabilities fall between 0 and 1.
- The set containing all possible outcomes receives probability 1. (Intuitively: this hypothesis exhausts the relevant possibilities, so it must be true.)
- An additivity axiom.

To illustrate the additivity axiom, suppose that h_1 and h_2 are mutually exclusive hypotheses (i.e. disjoint sets of possible outcomes). More concretely, let h_1 be the hypothesis that Seabiscuit wins the race and h_2 the hypothesis that War Admiral wins the race. Consider the union $h_1 \cup h_2$: the set of possible outcomes in which Seabiscuit finishes before every other horse *or* War Admiral

finishes before every other horse. This is the hypothesis that Seabiscuit wins the race or War Admiral wins the race. Additivity requires that:

$$p(h_1 \cup h_2) = p(h_1) + p(h_2).$$

More generally, *finite additivity* demands that:

$$p(h_1 \cup h_2 \cup \ldots \cup h_n) = p(h_1) + p(h_2) + \ldots + p(h_n)$$

when h_1, h_2, …, h_n is a finite list of mutually exclusive hypotheses. More generally still, consider a potentially infinite list of mutually exclusive hypotheses h_1, h_2, …, h_i, …. *Countable additivity* demands that:

$$p(\bigcup_i h_i) = \sum_i p(h_i),$$

where $\bigcup_i h_i$ is the union of the h_i. Most Bayesians endorse countable additivity (e.g. Easwaran, 2013b), but some endorse only finite additivity (e.g. de Finetti, 1972). Cognitive science applications typically presuppose countable additivity.

Box 2. Conditional probability

When $p(e) > 0$, we may define the conditional probability $p(h \mid e)$ using the *ratio formula*:

$$p(h \mid e) = \frac{p(h \cap e)}{p(e)}.$$

Here hypotheses are sets of possible outcomes (see Box 1), and $h \cap e$ is the intersection of *h* and *e*: the set of outcomes contained in both *h* and *e*. The ratio formula restricts attention to outcomes contained in *e* and then selects the proportion of those outcomes also contained in *h*. When $p(e) = 0$, the ratio formula is ill-defined. However, scenarios where $p(e) = 0$ arise in virtually all scientific fields that employ Bayesian decision theory, including Bayesian cognitive science. For

example, it is common to update credences upon learning that a continuous random variable X has value x. This requires conditional probabilities $p(h \mid X = x)$, where $X = x$ is the set of outcomes in which X has value x. The probability calculus axioms entail that $p(X = x) = 0$ for all but countably many values x (Billingsley, 1995, p. 162, p. 188), so the ratio formula does not supply the needed conditional probabilities. There are several alternative theories that supply conditional probabilities for cases where $p(e) = 0$ (Easwaran, 2019). All of these theories abandon any attempt at explicitly *defining* conditional probability and instead impose axiomatic constraints that conditional probabilities should satisfy. The most popular approach uses *regular conditional distributions*, introduced by Kolmogorov in the same treatise that codified the probability calculus (1933/1956). Regular conditional distributions figure crucially in probability theory (Billingsley, 1995) and in many scientific applications of the Bayesian framework, including within statistics (Ghosal & van der Vaart, 2017), economics (Feldman, 1987), and cognitive science (Bennett et al., 1996).

3. JUSTIFYING CREDAL NORMS

Bayesian decision theory governs the rational allocation, evolution, and employment of credence. For example, someone whose credences satisfy the probability calculus axioms is alleged to be more rational than someone whose credences violate the axioms. Cognitive scientists often motivate Bayesian modeling by invoking the rationally privileged status of Bayesian norms. But why should we ascribe a privileged status to Bayesian norms? Why is someone who satisfies Bayesian norms any more rational than someone who violates them? What is so special about *these* norms as opposed to alternative norms one might follow? A vibrant tradition stretching over the past century aims to establish that Bayesian norms are

rationally privileged. The goal is to justify Bayesian norms over alternative possible norms. Attempts at systematic justification have focused mainly on the probability calculus axioms and Conditionalization, rather than expected utility maximization.

One justificatory strategy, pursued by Cox (1946) and Jaynes (2003), advances constraints upon rational credence and then derives the probability calculus axioms from those constraints. Unfortunately, the Cox-Jaynes constraints do not seem any more antecedently plausible than the axioms themselves (Weisberg, 2009). Thus, it is debatable how much extra justification the Cox-Jaynes constraints confer upon the already very plausible axioms. Moreover, the Cox-Jaynes constraints are *synchronic*, so they cannot possibly justify the *diachronic* norm Conditionalization. Even if the Cox-Jaynes constraints justify the probability calculus axioms, the problem of justifying Conditionalization would remain.

A second justificatory strategy, stretching back to Ramsey (1931) and de Finetti (1937/1980), emphasizes the connection between probability and *gambling*. Your credences influence which bets you are willing to accept. For example, you will accept different odds on Seabiscuit winning the race depending on whether you have a high or low credence that Seabiscuit will win the race. Ramsey and de Finetti exploit the connection with gambling to justify the probability calculus axioms. To illustrate, suppose that John's credences violate the probability calculus axioms. Ramsey and de Finetti show that (under ancillary assumptions) a devious bookie can offer John a collection of bets that inflict a *sure monetary loss*. John will accept each bet as fair, yet he will lose money overall no matter how the individual bets turn out. A collection of bets with this property is called a *Dutch book*. In contrast, the bookie cannot mount a Dutch book against an agent whose credences obey the probability calculus axioms. So agents who violate the axioms can be pumped for money in a way that agents who obey the

axioms cannot. Ramsey and de Finetti argue on this basis that credences should obey the probability calculus axioms.[3]

Lewis (1999) and Skyrms (1987) extend Dutch book argumentation from the synchronic realm to the diachronic realm. To illustrate, suppose Mary learns *e* but does not replace her former credence $p(h)$ with new credence $p(h \mid e)$, i.e. she violates Conditionalization. Lewis shows that (under ancillary assumptions) a clever bookie can inflict a sure loss upon Mary by offering her a series of bets that she regards as fair: some bets offered *before* she learns *e* and a final bet offered *after* she learns *e*. These bets comprise a *diachronic Dutch book*. Skyrms shows that an agent who conforms to the probability calculus axioms and to Conditionalization is not similarly vulnerable to a diachronic Dutch book. So agents who violate Conditionalization can be pumped for money in a way that agents who obey Conditionalization cannot. Lewis and Skyrms argue on this basis that agents should obey Conditionalization. Although Lewis and Skyrms assume that $p(e) > 0$, one can prove generalized Dutch book results for many cases where $p(e) = 0$, including all or virtually all cases likely to arise in realistic applications (Rescorla, 2018).

There is a large literature on Dutch book arguments, much of it quite critical. One popular critique is that vulnerability to a Dutch book does not on its own entail irrationality. *From a practical perspective*, there is something undesirable about credences that leave you vulnerable to a Dutch book. But why conclude that your credences are *irrational*? By analogy, suppose you believe you are going to perform badly in an upcoming job interview. Your belief is quite rational: you have ample evidence based on past experience that you are likely to perform badly in the interview. Your belief makes you nervous, and as a result you do even worse in the job interview than you would have had you formed an irrationally overconfident belief that you

[3] The Dutch book argument for countable additivity (see Box 1) posits a book containing countably many bets. For worries about this posit, see (Arntzenius, Elga, and Hawthorne, 2003) and (Rescorla, 2018, p. 717, fn. 9).

would do well. The job interview example shows that *beliefs* with negative practical consequences may be quite rational. Why, then, must *credences* with negative practical consequences be irrational? See (Hájek, 2009) for discussion of this objection, along with many other objections to Dutch book arguments. And see (Eberhardt & Danks, 2011) for critical discussion of diachronic Dutch book arguments in connection with Bayesian cognitive science.

A third justificatory strategy centers on the notion of *accuracy*. Here "accuracy" measures how much your credences differ from actual truth-values: if you assign high credence to false hypotheses and low credence to true hypotheses, then your credences are relatively inaccurate; if you assign high credence to true hypotheses and low credence to false hypotheses, then your credences are relatively accurate. There are various possible "scoring rules" that quantify accuracy more precisely. Suppose that your credences violate the probability calculus axioms. Joyce (1998, 2009) shows that, for a large class of scoring rules, there exists an alternative credal allocation that satisfies the axioms and that is guaranteed to improve your accuracy score. Intuitively: if your credences violate the axioms, then (for a large class of scoring rules) you can ensure that your credences are more accurate by emending them so as to obey the axioms. Credences that already conform to the axioms, by contrast, are not so improvable. Joyce argues on this basis that it is irrational to violate the axioms.

One problem with the accuracy-based argument is that it only works for certain scoring rules. If we allow a sufficiently broad class of scoring rules, then you cannot always ensure an improved accuracy score by correcting violations of the probability calculus axioms (Hájek, 2008; Maher 2002). Thus, the accuracy-based argument only goes through if we restrict the class of admissible scoring rules. Do the needed restrictions on admissible scoring rules beg the

question in favor of the probability calculus axioms? Or can one independently motivate the needed restrictions? These questions are subject to ongoing debate (Pettigrew, 2019).

Some authors extend accuracy-based argumentation from the synchronic to the diachronic realm (Easwaran, 2013a; Greaves & Wallace, 2006; Leitgeb & Pettigrew, 2010). Consider *expected accuracy*: the accuracy score you expect for your future credences. Different update rules will yield different future credences, so expected accuracy depends on how you plan to update your credences in light of new evidence. Greaves and Wallace show that (under ancillary assumptions) Conditionalization maximizes expected accuracy. In other words, you maximize the expected accuracy of your future credences by using Conditionalization as your update rule. Greaves and Wallace argue that this result privileges Conditionalization over alternative update rules one might follow. However, Schoenfield (2017) shows that the expected accuracy argument for Conditionalization rests upon a crucial assumption: namely, that you are certain you will update your credences based upon an *e* that is *true*. What if the *e* upon which you update is false? People make mistakes all the time, so it seems entirely possible that you will update your credences based upon a false *e*. Schoenfield shows that, once you allow for this possibility, Conditionalization no longer maximizes the expected accuracy of your future credences.[4] Schoenfield's analysis casts doubt upon the expected accuracy argument for Conditionalization.[5]

Justification of credal norms is a very active research area. Several notable developments have occurred quite recently, so the landscape is likely to shift in the coming years.

---

[4] In contrast, the Lewis-Skyrms Dutch book results for Conditionalization extend smoothly to cases where *e* may be false (Rescorla, forthcoming c).

[5] Briggs and Pettigrew (2020) give an *accuracy-dominance* argument for Conditionalization. They show that (under ancillary assumptions) an agent who violates Conditionalization could have guaranteed an improved accuracy score by instead obeying Conditionalization. However, their argument assumes away the possibility that the agent updates based upon a false *e*. So the argument leaves open whether Conditionalization guarantees an improved accuracy score in situations where *e* may be false.

4. BAYESIAN COGNITIVE SCIENCE

A major trend in cognitive science over the past few decades has been the rise of Bayesian modeling. Adherents pursue the following methodology when studying a psychological task: first, build an idealized Bayesian model of the task; second, fit the Bayesian model to experimental data as well as possible by fixing any free parameters; third, evaluate how well the model fits the data with all free parameters specified.

Bayesian perceptual psychology is a particularly successful branch of Bayesian cognitive science. How does the perceptual system estimate distal conditions based upon proximal sensory input? For example, how does it estimate the shapes, sizes, colors, and locations of nearby objects based upon retinal stimulations? Helmholtz (1867/1925) proposed that the perceptual system estimates distal conditions through an *unconscious inference*. Bayesian perceptual psychology develops Helmholtz's proposal, postulating unconscious *Bayesian* inferences executed by the perceptual system (Knill & Richards, 1996). On the Bayesian approach, perception has prior probabilities $p(h)$ over distal conditions (e.g. possible shapes; possible lighting conditions) and prior likelihoods $p(e \mid h)$ relating distal conditions to sensory input (e.g. the likelihood of retinal input e given that a certain shape is present under certain lighting conditions). The perceptual system responds to input e by computing the posterior $p(h \mid e)$, in accord with Conditionalization. Based on the posterior, the perceptual system selects a privileged hypothesis h regarding distal conditions. In many models, though not all, the selected hypothesis h is the one that maximizes the posterior.

Bayesian perceptual psychologists have constructed empirically successful models of numerous phenomena, including:

- *Perceptual constancies*. The perceptual system can estimate the same distal property in response to diverse proximal sensory stimulations. For example, you can recognize that a surface has a particular color shade across diverse lighting conditions, despite radically different retinal input. This is *color constancy*. How is color constancy achieved? Helmholtz conjectured that the perceptual system estimates lighting conditions and then uses the lighting estimate to infer surface color from retinal input. Bayesian models formalize Helmholtz's conjecture, generating predictions that fit actual human performance quite well (Brainard, 2009).
- *Perceptual illusions.* Perceptual estimation deploys a prior probability over possible distal conditions. The prior helps the perceptual system decide between rival hypotheses that are, in principle, equally compatible with current proximal input. Perceptual priors often match the statistics of the environment quite well. For that reason, they often induce accurate perceptual estimates. But sometimes the prior assigns low probability to actual distal conditions, inducing a *perceptual illusion*. For example, the perceptual system has high prior probability that the light source is overhead (Stone, 2011). The light-from-overhead prior informs perceptual estimation of shape from shading. Since the light source is almost always located overhead, the resulting shape estimates tend be very accurate. When the light source happens to be located elsewhere, the shape estimates are inaccurate. Bayesian models can similarly explain a host of other perceptual illusions, such as motion illusions (Weiss, Simoncelli, & Adelson, 2002; Kwon, Tadin, & Knill, 2015).

- *Cue combination.* If you hold an apple while looking at it, then you receive both visual and haptic information regarding the apple's size. A similar phenomenon occurs within modalities, as when the visual system estimates depth based upon convergence, retinal disparity, motion parallax, and other cues. How does the perceptual system integrate multiple cues into a unified estimate? The Bayesian framework helps us model sensory cue combination in a principled way (Ernst and Banks, 2002). For instance, estimation based upon two distinct cues might feature prior likelihoods $p(e_1 \mid h)$ and $p(e_2 \mid h)$, where e_1 and e_2 correspond to the two cues. Inputs e_1 and e_2 yield a posterior $p(h \mid e_1, e_2)$, which supports a unified estimate. Bayesian models of cue combination have achieved notable success for many intermodal and intramodal estimation tasks (Trommershäuser et al., 2011). A good example is the causal inference model given by (Körding et al., 2007), which estimates location by assessing whether a visual cue and an auditory cue have a common cause or two distinct causes. The model successfully explains several multisensory effects, such as the ventriloquism illusion.

Overall, the Bayesian framework generates principled, quantitatively precise explanations for diverse perceptual phenomena (Rescorla, 2015; Rescorla, forthcoming b).

Priors employed by the perceptual system are highly mutable. They can change quite rapidly in response to changing environmental statistics. For instance, the light-from-overhead prior can rapidly change in response to visual-haptic input indicating an altered lighting direction (Adams et al., 2004). Mutability of priors is key to accurate perception: as just noted, a prior that is not well-tuned to the environment will tend to induce inaccurate perceptual estimates.

Motor control is another area where Bayesian modeling has proved highly successful. How does the motor system select motor commands that promote one's goals? Even a mundane task such as picking up a cup is a complex undertaking with multiple degrees of freedom (Bernstein, 1967). On a Bayesian approach, the motor system selects appropriate motor commands through unconscious Bayesian inference and decision-making (Wolpert, 2007). The motor system maintains a running estimate of relevant environmental conditions, including both distal state and bodily state. For example, when you reach towards a target, your motor system maintains a running estimate of the target's location along with the location and velocity of your hand. Credences are updated based upon incoming proximal input, through iterated application of Conditionalization. As the task unfolds, your motor system uses its updated credences to select motor commands that maximize expected utility (Wolpert & Landy, 2012). The utility function reflects the task goal (e.g. reaching the target with your hand) along with other task-invariant desiderata (e.g. minimizing energetic expenditure). Detailed Bayesian models successfully explain human motor performance in a range of tasks (Haith & Krakauer, 2013; Rescorla, 2016).

A basic fact about human motor performance is that it responds differentially to external perturbations depending on how those perturbations impact the task goal. Some perturbations are *task-relevant*: they affect execution of the task. Other perturbations are *task-irrelevant*: they do not affect execution of the task. For example, if the task is to maintain a certain hand position, then perturbations that affect hand position are task-relevant while perturbations that affect joint angles but not hand position are task-irrelevant. It is well-established that the motor system preferentially corrects task-relevant perturbations over task-irrelevant perturbations (Nashed et al., 2012; Todorov, 2004). This experimentally observed disparity is readily explicable within a

Bayesian framework: correcting a perturbation expends energy, so an ideal Bayesian decision-maker whose utility function penalizes energetic expenditure will only correct perturbations that impede the task (Todorov and Jordan, 2002). In contrast, rival frameworks have great difficulty explaining the experimentally observed disparity between task-relevant and task-irrelevant perturbations. Rival frameworks typically enshrine *the desired trajectory hypothesis*, according to which the motor system crafts a detailed plan for motor organs and then seeks to execute that plan (e.g. Friston, 2011; Latash, 2010). The desired trajectory hypothesis incorrectly predicts that the motor system will correct task-irrelevant perturbations along with task-relevant perturbations (Braun & Wolpert, 2007; Scott, 2012).

In addition to perception and motor control, researchers have offered Bayesian models for many other psychological domains (Chater et al., 2010), including *intuitive physics* (Battaglia et al., 2013), *social cognition* (Baker & Tennenbaum 2014), *causal reasoning* (Griffiths & Tenenbaum, 2009), *child development* (Gopnik & Bonawitz, 2015), *human and non-human navigation* (Madl et al. 2014), and *natural language parsing* (Hale, 2011; Levy, 2008; Narayan & Jurafsky, 1998).

4.1 Approximate Bayesian inference

Bayesian inference computes the posterior from the priors. Unfortunately, computation of the posterior is in general *intractable*, consuming resources of time or memory beyond those available to a limited physical system (Kwisthout et al., 2011). The reason is that calculating the normalization constant η requires summing (or integrating) over the hypothesis space, which usually requires substantial computational resources. The brain only has limited resources at its disposal, so it cannot execute intractable Bayesian inferences.

In some special cases, one can circumvent computational intractability by positing priors that allow tractable inference. When the prior probability and the prior likelihood are Gaussian, for example, the posterior is easily computable (Gelman et al., 2014, pp. 39-42). Gaussian priors work well for some modeling purposes, such as Bayesian modeling of certain simple perceptual tasks. However, human priors need not be precisely Gaussian even in the special case of perception (Stocker & Simoncelli, 2006). As a general matter, we should not expect the priors that figure in human mental activity to support tractable Bayesian inference (Pouget et al., 2013).

The standard response is to explore algorithms that *approximately implement* idealized Bayesian inference. A huge literature in statistics and machine learning investigates algorithms that can tractably approximate intractable Bayesian inference (Murphy, 2012). There are two main implementation strategies:

- *Variational approximation* approximates the posterior with a probability distribution drawn from a "nice" family of distributions, such as Gaussians. The goal is to choose a nice distribution as "close" as possible to the true posterior. This is often a tractable task even if precise computation of the posterior is intractable.
- *Sampling approximation* approximates the posterior by drawing *samples* from the hypothesis space. In a sampling model, there is an *objective* probability that the system draws a given hypothesis. This objective probability serves as the *subjective* probability assigned by the system to the hypothesis (Icard, 2016). A good sampling algorithm draws samples in a way that approximates computation of the posterior. More precisely: a system that implements the algorithm will, given sufficient computational resources, asymptotically converge to objective

> sampling probabilities that match the true posterior probabilities (Murphy, 2012, pp. 817-876).

Variational and sampling approximation schemes have both been used to model mental activity (Sanborn, 2017). For example, Gershman et al. (2012) hypothesize that multistable perception arises from a tractable sampling approximation to intractable Bayesian computation. Their sampling model explains a variety of phenomena, such as the distribution of switching times between percepts.

An approximate Bayesian inference will, by definition, deviate from the rational ideal codified by Conditionalization. The question therefore arises whether approximate Bayesian inference has a rationally privileged status (Danks & Eberhardt, 2011; Eberhard & Danks, 2011). Why does an algorithm that merely *approximates* Bayesian inference deserve praise as any more rational than an algorithm that is radically anti-Bayesian? Why is it better to *come close* to the Bayesian ideal than to depart from that ideal quite dramatically?

In response, one might simply concede that approximate Bayesian inference does not have a rationally privileged status. If our goal is to delineate well-confirmed models of mental activity, then we can achieve this goal without attributing a rationally privileged status to our models. The mind might execute approximate Bayesian inference *whether or not* approximate Bayesian inference is rationally privileged. Another possible response is to argue that approximate Bayesian inferences are somehow rationally privileged, given the mind's limited computational resources. Icard (2018) develops this viewpoint, drawing on the notion of *bounded rationality* (Simon, 1956). Icard considers an agent who makes predictions based upon incoming evidence. The agent has limited computational resources at her disposal, so she must

balance "rational" use of those resources against the quality of her predictions.[6] Icard shows that (under ancillary assumptions) the optimal solution to this problem yields a sampling approximation to Bayesian inference. Further research in the spirit of Icard's analysis is needed. Such research would help us assess the extent, if any, to which approximate Bayesian inference has a rationally privileged status.[7]

4.2 Violation of Bayesian norms?

A recurring worry about Bayesian cognitive science is that people often seem to violate Bayesian norms (Colombo et al., forthcoming; Eberhardt & Danks, 2011; Mandelbaum et al., forthcoming). Kahneman and Tversky argue that human reasoning systematically violates the probability calculus axioms and that human decision-making systematically violates expected utility maximization (Kahneman & Tversky, 1979; Tversky & Kahneman, 1983). Similarly, Rahnev and Denison (2018) argue that some perceptual phenomena are anti-Bayesian.

Proponents of Bayesian modeling respond that many apparently anti-Bayesian phenomena are explicable in Bayesian terms (Stocker, 2018). Consider the *size-weight illusion*: when you lift two objects of equal weight but different size, the smaller object feels heavier. At first, the illusion looks anti-Bayesian because it flouts a prior expectation that larger objects are heavier. However, the illusion turns out to be explicable by a Bayesian model that estimates relative *densities* (Peters et al., 2016). As this example illustrates, it is often possible for

[6] See (Lieder & Griffiths, 2020) for general discussion of how humans make rational use of limited cognitive resources.

[7] De Bona and Staffel (2018) study an agent who violates the probability calculus axioms. They show that (under ancillary assumptions) the agent can improve her accuracy score and also reduce her potential Dutch book losses by moving closer towards conformity to the axioms. Building on these and other similar results, Staffel (2019) argues that an agent is more rational to the extent that her credences conform to the probability calculus axioms. It remains unclear whether Staffel's argumentative strategy can extend to the diachronic norm Conditionalization.

Bayesians to accommodate apparently anti-Bayesian phenomena by constructing a more sophisticated Bayesian model (Wei & Stocker, 2015).

In any event, Bayesian cognitive science does not aim to establish that *all* or *most* mental activity conforms to Bayesian norms (Griffiths et al., 2012). It aims to *investigate* how well mental activity conforms to Bayesian norms. Obviously, the enterprise would be a waste of time if *no* mental activity even approximately conformed to Bayesian norms. However, Bayesian cognitive science does not presuppose that any particular mental process conforms to the norms. Bayesian norms are not intended to be universal psychological laws. It may be that some mental activity conforms well to the norms while other mental activity violates the norms, perhaps very dramatically. For example, even if personal-level reasoning and decision-making flagrantly violate Bayesian norms, perception may still conform quite closely. Even if *people* are very irrational, *the perceptual system* may make rational or near-rational inferences. Similarly, it may be that *certain* perceptual processes conform quite well to Bayesian norms while *others* do not. One must investigate on a case-by-basis how well and in what circumstances a given mental process conforms to the norms.

Even when a mental process violates Bayesian norms, it may still implement a tractable algorithm that approximates idealized Bayesian inference. Consider *order effects*: the order in which evidence is received impacts human judgments. Order effects straightforwardly violate Conditionalization, which does not take the order of evidence into account. However, Sanborn et al. (2010) argue that order effects in human categorization are explicable by a sampling model that approximates idealized Bayesian inference. Thus, idealized Bayesian modeling can sometimes shed light upon a mental phenomenon that violates Bayesian norms. In such a case,

the Bayesian model articulates an ideal benchmark that human performance approximates but fails to meet due to limited computational resources.

5. REALISM VERSUS INSTRUMENTALISM

Realists about Bayesian cognitive science hold that, when a Bayesian model is explanatorily successful, we have reason to regard the model as approximately true. For example, perceptual psychology offers Bayesian models that successfully explain constancies, illusions, and cue combination. Realists hold that we have good reason to regard these models as approximately true descriptions of how the perceptual system works (Rescorla, 2015). Similarly for successful models of other psychological domains, such as motor control (Rescorla, 2016).

What does it mean to say that a Bayesian model is "approximately true"? A Bayesian model posits *credal states*: assignments of credences to hypotheses. The model also posits *transitions* among credal states (e.g. the transition from priors to posterior). If the model is approximately true, then at a minimum there must be credal states and credal transitions roughly like those posited by the model. The mind must instantiate priors roughly like those posited by the model, and it must transition from the priors to a credal state roughly like the posited posterior.

How does a physical system "assign" a credence to a hypothesis? There are several options here:

- *Explicit enumeration*. A system can explicitly enumerate the credence assigned to each hypothesis. Explicit enumeration is not an option when the hypothesis space is infinite, as it is in virtually all scientific applications.

- *Parametric encoding*. A system can sometimes encode a probability distribution using finitely many parameters. For example, a system can encode a Gaussian distribution by recording the distribution's mean and variance. Parametric encoding is more general than explicit enumeration. It still has relatively limited applicability because many naturally arising probability distributions are not finitely parametrizable.
- *Sampling encoding*. When a system samples stochastically from the hypothesis space, there is an objective probability that the system draws a given hypothesis. As noted in section 4.1, this objective probability can serve as the subjective probability assigned by the system to the hypothesis. Thus, a system's sampling behavior can encode a probability distribution.

Applications of the Bayesian framework (e.g. in statistics or robotics) virtually always use parametric or sampling encoding rather than explicit enumeration. Likewise, realists about Bayesian cognitive science can happily allow that the mind sometimes uses parametric or sampling encoding of credal states (Rescorla, 2020).

Instrumentalists about Bayesian cognitive science reject realism. They regard a Bayesian model as nothing more than a useful tool for making predictions about human behavior (Block, 2018; Colombo & Seriès, 2012). From an instrumentalist viewpoint, the empirical success of a Bayesian model provides no reason to accept that there are credal states and transitions remotely like those posited by the model. For example, a Bayesian perceptual model may correctly predict various experimental results, but we should not conclude that the perceptual system even approximates a Bayesian inference like that posited by the model. We should only say that the perceptual system behaves *as if* it executes a Bayesian inference. Talk about priors and posteriors

is just a useful fiction --- a helpful way of summarizing observed human behavior. We should attribute no psychological reality to the core theoretical posits of Bayesian models.

The debate between realism and instrumentalism about Bayesian cognitive science relates to a more general debate within philosophy of science. *Scientific realists* say that the explanatory success of a scientific theory gives us reason to adopt a positive attitude towards the approximate truth of the theory (Chakravartty, 2007; Putnam, 1975). *Instrumentalists* disagree (van Fraassen, 1980). They regard scientific theories simply as useful tools for making predictions. For example, scientific realists say that the explanatory success of modern physics gives us reason to believe in subatomic particles, while instrumentalists say that we have no reason to believe in subatomic particles. Those inclined towards instrumentalism in general will presumably incline towards instrumentalism regarding Bayesian cognitive science. However, one might be a scientific realist regarding most sciences (physics, chemistry, biology, etc.) and an instrumentalist regarding Bayesian cognitive science. One might believe in the subatomic particles posited by physics but not the credal states and transitions posited by Bayesian cognitive science. Moreover, one might be a realist about certain Bayesian models (e.g. Bayesian models of perception and motor control) but an instrumentalist about other Bayesian models (e.g. Bayesian models of high-level cognition). See (Rescorla, 2020) for a defense of realism regarding at least some Bayesian models.

6. NEURAL IMPLEMENTATION OF BAYESIAN INFERENCE

If we accept that the mind executes (or approximately executes) Bayesian inferences, the question arises how those inferences are physically implemented. How does the brain encode priors and posteriors? Through what neural mechanisms does the brain implement, or

approximately implement, the transition from priors to posterior? To tackle these questions, neuroscientists construct *neural network* models that encode credal states and that approximately implement Bayesian computations. Researchers have proposed two main approaches (Fiser at al. 2010): parametric and sampling.

The most popular version of the parametric approach uses *probabilistic population codes* (PPCs), in which the firing pattern over a neural population encodes parameters of a probability distribution. Consider a population of neurons tuned to some observable variable (e.g. orientation of a bar), where each neuron's tuning curve peaks at a different value of the variable (e.g. a different possible orientation of the bar). Under biologically plausible assumptions about neural noise, firing activity in the population can encode the mean and variance of a Gaussian distribution for the variable (Knill and Pouget, 2004). PPCs can also encode certain non-Gaussian finitely parametrizable distributions (Ma et al., 2006). PPCs support a number of Bayesian operations, including computation of the posterior (Orhan & Ma, 2017; Pouget et al. 2013). When computation of the posterior is intractable, PPCs can sometimes approximate Bayesian inference through variational methods (Beck, Pouget, & Heller, 2012).

The second main approach to neural implementation uses sampling (Fiser et al., 2010). On a sampling approach, each neuron's activity encodes a sample from the hypothesis space. Specifically, a neuron's membrane potential can encode a possible value of a variable. By sampling from the hypothesis space, the brain approximates computation of the posterior. There are several neural network models that develop this intuitive idea (Aitchison & Lengyel, 2016; Buesing et al., 2011; Hoyer & Hyvärinen, 2002; Orbán et al., 2016). Some of the models successfully replicate neurophysiological phenomena that existing PPC models do not capture. For example, experimentally observed neural responses have a complex dynamics, including

oscillatory response to a constant stimulus (Roberts et al., 2013) and large transient increases at stimulus onset (Ray & Maunsell, 2010). These dynamical properties are predicted by the (Aitchison & Lengyel, 2016) sampling model but not by existing PPC models.

Research into neural implementation of Bayesian inference is still at an early stage. We do not know whether the brain uses a PPC implementation scheme, a sampling implementation scheme, or some other scheme. Note that PPC and sampling implementations need not be mutually exclusive: Shivkumar et al. (2018) give a sampling-based model of neural responses in V1, and they show that the model can be interpreted as employing a PPC.

Some research into neural implementation emphasizes *predictive coding*. On a predictive coding approach, the neural network generates a prediction about expected input and compares this prediction with actual input, computing a *prediction error* term. Prediction error then propagates back through the network as an impetus to further computation. Prediction error is usually the difference between prediction and input (Rao & Ballard, 1999) but is sometimes the ratio between them (Spratling, 2016). Predictive coding can be combined with a parametric scheme, such as a PPC (Spratling, 2016), or with a sampling scheme (Lee & Mumford, 2003). Clarke (2015) and Hohwy (2014) promote a version of predictive coding, developed by Friston (2005, 2010), that computes a variational approximation to the posterior. Proponents of predictive coding argue that it can explain a range of neurophysiological phenomena, such as extra-classical receptive field effects in the visual cortex. Critics retort that those same phenomena can be explained equally well without recourse to predictive coding (Aitchison & Lengyel, 2016; Aitchison & Lengyel, 2017; Heeger, 2017). To illustrate, consider the extra-classical effect known as *surround suppression*, in which neural response diminishes when the stimulus extends beyond the neuron's receptive field. An example would be a cell that

preferentially responds to a bar with a certain orientation but that exhibits a reduced response when the bar extends beyond the cell's receptive field. Predictive coding can explain surround suppression. For instance, longer bars are more predictable than short bars because they are more common in the natural environment, so a longer bar induces lower prediction error and hence reduced response by the error-detecting cell (Rao & Ballard, 1999). However, alternative models can explain surround suppression without any appeal to predictive coding (e.g. Coen-Cagli, Kohn, & Schwartz, 2015). These models also posit that activity in a neural population modulates cellular response --- just not through computation of prediction error.

Notably, research into the neural basis of approximate Bayesian inference presupposes a realist rather instrumentalist perspective on credal states and transitions (Ma, 2019). This research program aims to discover how credal states are physically realized in the brain and how the brain transits between credal states in approximate accord with Bayesian norms. So the research program presupposes that there *are* credal states and that the brain transits between them in rough accord with Bayesian norms. In other words, it presupposes realism. Searching for the neural mechanisms that underlie approximate Bayesian inference would be pointless if the brain only behaves *as if* it approximately implements Bayesian inference. Thus, the dispute between realism and instrumentalism has major ramifications for neuroscience. If realism is correct, then research into neural implementation of Bayesian computation is a vitally important endeavor. If instrumentalism is correct, then that research is a waste of time.

Debate continues as to which of these two viewpoints is correct. However, there is some suggestive neurological evidence for the realist perspective. On a sampling approach, and also on certain versions of the PPC approach (Ganguli & Simoncelli, 2014), neural activity evoked by sensory input encodes the posterior probability, while spontaneous activity absent any sensory

input encodes the prior. If spontaneous activity encodes the prior, then spontaneous activity should change during development as the prior comes to match environmental statistics. More specifically: a prior that matches environmental statistics should closely resemble the average posterior evoked by external input; so one predicts a close match between spontaneous activity and average evoked activity in adults but a less close match at earlier stages of development. A study of the ferret visual cortex confirmed that prediction, revealing spontaneous activity that grew closer to average evoked activity during development (Berkes et al., 2011). The change in spontaneous activity arguably reflects gradual acquisition of a prior fitted to environment statistics. This interpretation presupposes that the prior is a genuine mental state physically realized in the brain. See also (Walker et al., 2020), which provides neurophysiological evidence for PPC implementation of Bayesian inference in the rhesus macaque visual cortex.

7. CONCLUSION

Bayesian decision theory provides an organizing framework for studying normative, computational, and neural aspects of mental activity. There is good evidence that human performance across a range of psychological domains closely matches either the Bayesian ideal or some tractable approximation thereof. If we accept that a Bayesian model (or a tractable approximation model) accurately describes a mental process, then we can take the model as a guide towards discovery of underlying neural mechanisms.

Which mental processes approximately conform to Bayesian norms, and which do not? Which circumstances promote conformity to Bayesian norms, and which circumstances impede conformity? How do limited computational resources impact the mind’s ability to conform to Bayesian norms? When a mental process approximately conforms to Bayesian norms, how did

evolutionary and developmental pressures help instill this approximate conformity? How do nature and nurture jointly shape the priors employed during Bayesian computation? Continued interdisciplinary investigations into these and other questions raised by the Bayesian program promise to illuminate how the mind grapples with an uncertain world.

Acknowledgments

I presented some of this material at the 2019 Norwegian Summer Institute on Language and Mind. I thank all the participants, especially Steven Gross and Georges Rey, for their feedback. I am also grateful to Wayne Wu and two anonymous referees for comments that greatly improved the paper and to Thomas Icard and Julia Staffel for helpful discussion.

Works Cited

Adams, W., Graf, E., & Ernst, M. (2004). Experience can change the "light-from-above" prior. *Nature Neuroscience, 7*, 1057-1058.

Aitchison, L., & Lengyel, M. (2016). The Hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS Computational Biology*, *12*, e1005186.

---. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219-227.

Arntzenius, F., Elga, A., & Hawthorne, J. (2004). Bayesianism, infinite decisions, and binding. *Mind, 113*, 251–283.

Arrow, K. (1971). *Essays on the theory of risk-bearing*. Chicago: Markham.

Ashby, D. (2006). Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine*, *25*, 3589-3631.

Baker, C., & Tenenbaum, J. (2014). Modeling human plan recognition using Bayesian theory of mind. In G. Sukthankar, R. P. Goldman, C. Geib, D. Pynadath, D., & H. Bui (Eds.), *Plan, activity, and intent recognition: Theory and Practice* (pp. 177-204). Waltham: Morgan Kaufmann.

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

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 470-418.

Beck, J., Heller, K. & Pouget, A. (2012). Complex inference in neural circuits with probabilistic population codes and topic models. In P. Bartlett (Ed.), *Advances in Neural Information Processing Systems* (pp. 3068–3076). Cambridge: MIT Press.

Bennett, B., Hoffman, D., Prakash, C., & Richman, S. (1996). Observer theory, Bayes theory, and psychophysics. In D. Knill and W. Richards (Eds.), *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an internal model of the environment. *Science, 331*, 83-7.

Bernstein, N. (1967). *The coordination and regulation of movements*. Oxford: Pergamon.

Billingsley, P. (1995). *Probability and measure*. 3rd ed. New York: Wiley.

Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic?. *Philosophical Transactions of the Royal Society B, 373*, 20170341.

Brainard, D. (2009). Bayesian approaches to color vision. In M. Gazzaniga (Ed.), *The Visual Neurosciences*, 4th ed. (pp. 395-408). Cambridge MIT Press.

Braun, D., & Wolpert, D. (2007). Optimal control: When redundancy matters. *Current Biology, 17*, R973-R975.

Briggs, R., & Pettigrew, R. (2020). An accuracy-dominance argument for conditionalization. *Nous*, *54*, 162-181.

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PloS Computational Biology, 7*, e1002211.

Chakravartty, A. (2007). *A metaphysics for scientific realism*. Cambridge: Cambridge University Press.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *WIREs Cognitive Science, 1*, 811-823.

Clark, A. (2015). *Surfing uncertainty*. Oxford: Oxford University Press.

Coen-Gagli, R., Kohn, A., & Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience, 18*, 1648-1655.

Colombo, M., Elkin, L., & Hartmann, S. (Forthcoming). Being realist about Bayes and the predictive processing theory of mind. *The British Journal for the Philosophy of Science*.

Colombo, M., & Seriès, P. (2012). Bayes on the brain --- on Bayesian modeling in neuroscience. *The British Journal for the Philosophy of Science, 63*, 697-723

Cox, R. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics, 14*, 1–13.

de Finetti, B. (1937/1980). Foresight. Its logical laws, its subjective sources. Rpt. in H. E. Kyburg, Jr. & H. E. Smokler (Eds.), *Studies in subjective probability*. Huntington: Robert E. Krieger.

---. (1972). *Probability, induction, and statistics*. New York: Wiley.

Danks, D., & Eberhardt, F. (2011). Keeping Bayesian models rational: the need for an account of algorithmic rationality. *Behavioral and Brain Sciences, 34*, 197.

De Bona, G., & Staffel, J. (2018). Why be (approximately) coherent? *Analysis, 78*, 405-415.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Easwaran, K. (2013a). Expected accuracy supports conditionalization --- and conglomerability and reflection. *Philosophy of Science, 80*, 119-142.

---. (2013b). Why countable additivity?. *Thought*, *2*, 53-61.
---. (2019). Conditional probabilities. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.
Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: the problematic case of Bayesian models. *Minds and Machines*, *21*, 389-410.
Ernst, M., & Banks, M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429-433.
Feldman, M. (1987). Bayesian learning and convergence to rational expectations. *Journal of Mathematical Economics*, *16*, 292-313.
Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*, 119-130.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, *360*, 815-836.
---. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127-138.
---. (2011). What is optimal about motor control?. *Neuron*, *72*, 488-498.
Ganguli, D., & Simoncelli, E. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation*, *26*, 2103-2134.
Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehatri, A., & Rubin, D. (2014). *Bayesian data analysis*, 3rd ed. New York: CRC Press.
Gershman, S., Vul, E., & Tenenbaum, J. (2012). Multistability and perceptual inference. *Neural Computation*, *24*, 1-24.
Ghosal, S., & van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge: Cambridge University Press.
Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *WIREs Cognitive Science*, *6*, 75-86.
Greaves, H., & Wallace, D. (2006). Justifying Conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, *115*, 607-632.
Griffiths, T., Chater, N., Norris, D., & Pouget, A. (2012). How Bayesians got their beliefs (and what those beliefs really are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*, 415-422.
Griffiths, T., & Tenenbaum, J. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661-716.
Haith, A., & Krakauer, J. (2013). Theoretical models of motor control and motor learning. In M. Richardson, M. Riley, & K. Shockley (Eds.), *Progress in motor control VII: neural computational and dynamic approaches* (pp. 7-28). Springer: New York.
Hájek, A. (2008). Arguments for --- or against --- probabilism?. *British Journal for the Philosophy of Science*, *59*, 793-819.
---. (2009). Dutch book arguments. In P. Anand, P. Pattanaik, and C. Puppe (Eds.), *The handbook of rationality and social choice*. Oxford: Oxford University Press.
Hale, J. (2011). What a rational parser would do. *Cognitive Science*, *35*, 399-443.
Heeger, D. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences*, *114*, 1773-1782.
Helmholtz, H. von. (1867/1925). *Treatise on physiological optics*, trans. and ed. J. Southall. Manasha: George Banta Publishing Company.

Hohwy, J. (2014). *The predictive mind*. Oxford: Oxford University Press.
Hoyer, P., & Hyvärinen, A. (2002). Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in Neural Information Processing Systems, 155,* 277-284.
Icard, T. (2016). Subjective probability as sampling propensity. *The Review of Philosophy and Psychology, 7,* 863-903.
---. (2018). Bayes, bounds, and rational analysis. *Philosophy of Science, 85,* 79-101.
Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.
Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Oxford University Press.
Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science, 65,* 575-603.
---. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of Belief*. Springer.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263-291.
Knill, D., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience, 27,* 712–719.
Knill, D. & Richards, W., eds. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.
Kolmogorov, A. N. (1933/1956). *Foundations of the theory of probability*. 2nd English ed. Trans. N. Morrison. New York: Chelsea.
Körding, K., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J., & Shams, L. (2007). Causal inference in multisensory perception. *PloS One, 9,* e943.
Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science, 35,* 779-784.
Kwon, O.-S., Tadin, D., & Knill, D. (2015). Unifying account of visual motion and position perception. *Proceedings of the National Academy of Sciences, 112,* 8142–8147.
Laplace, P.-S. (1814/1902). *A Philosophical Essay on Probabilities*, trans. F. Truscott and F. Emory. New York: Wiley.
Latash, M. (2010). Motor control: In search of physics of the living systems. *Journal of Human Kinetics, 24,* 7-18.
Lewis, D. (1999). Why conditionalize?. In *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.
Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America*, 20, 1434-1448.
Lieder, F., & Griffiths, T. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43,* 1-60.
Leitgeb, H., & Pettigrew, R. (2010). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science, 77,* 236-72.
Levy, R. (2008). Expectation-based syntactic processing. *Cognition, 106,* 1126-1177.
Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology*, vol. iii (pp. 249-410). New York: Wiley.
Ma, W. J. (2019). Bayesian decision models: A primer. *Neuron, 104,* 164-175.

Ma, W. J., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432-1438.
Madl, T., Franklin, S., Chen, K., Montaldi, D., & Trappl, R. (2014). Bayesian integration of information in hippocampal place cells. *PloS One*, *9*, e89762.
Maher, P. (2002). Joyce's argument for probabilism. *Philosophy of Science*, *69*, 73-81.
Mandelbaum, E., Won, I., Gross, S., & Firestone, C. (2020). Can resources save rationality? "Anti-Bayesian" updating in cognition and perception. *Behavioral and Brain Sciences*, *43*, 31-32.
Meacham, C. (2016). Understanding Conditionalization. *Canadian Journal of Philosophy*, *45*, 767-797.
Murphy, K. (2012). *Machine learning: A probabilistic perspective*. Cambridge: MIT Press.
Narayanan, S. & Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of the Twelfth Annual Meeting of the Cognitive Science Society*.
Nashed, J., Crevecoeur, F. & Scott, S. (2012). Influence of the behavioral goal and environmental obstacles on rapid feedback Responses. *Journal of Neurophysiology*, *108*, 999-1009.
Navarro, D., Griffiths, T., Steyvers, M., & Lee, M. (2005). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101-122.
Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, *92*, 530-542.
Orhan, A. E., & Ma, W. J. (2017). Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. *Nature Communications*, *8*, 1-14.
Peters, M., Ma, W. J., & Shams, L. (2016). The size-weight illusion is not anti-Bayesian after all. *PeerJ*, *4*, e2124.
Pettigrew, R. 2019. Epistemic utility arguments for probabilism. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019). https://plato.stanford.edu/archives/win2019/entries/epistemic-utility/.
Pouget, A., Beck, J., Ma., W. J., & Latham, P. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*, 1170–1178.
Putnam, H. (1975). *Mathematics, matter, and method*. Cambridge: Cambridge University Press.
Ramsey, F. P. (1931). Truth and probability. In R. Braithwaite (Ed.), *The foundations of mathematics and other logical essays*. London: Routledge and Kegan.
Rahnev, D., & Denisov, R. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, e223.
Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79-87.
Ray, S., & Maunsell, J. (2010). Differences in gamma frequencies across visual cortex restrict their possible use in computation. *Neuron*, *67*, 885–896.
Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (Ed.), *The Oxford handbook of the philosophy of perception*. Oxford: Oxford University Press.
---. (2016). Bayesian sensorimotor psychology. *Mind and Language*, *31*, 3-36.
---. (2018). A Dutch book theorem and converse Dutch book theorem for Kolmogorov Conditionalization. *The Review of Symbolic Logic*, *11*, 705-735.
---. (2020). A realist perspective on Bayesian cognitive science. In A. Nes & T. Chan (Eds.), *Inference and Conscious*. New York: Routledge.

---. (Forthcoming a). On the proper formulation of Conditionalization. *Synthese*.
---. (Forthcoming b). Perceptual co-reference. *Review of Philosophy and Psychology*.
---. (Forthcoming c). An improved Dutch book theorem for Conditionalization. *Erkenntnis*.
Roberts, M., Lowet, E., Brunet, N., Wall, M. T., Tiesigna, P., Fries, P., & Weerd, P. (2013). Robust gamma coherence between macaque V1 and V2 by dynamic frequency matching. *Neuron, 78,* 523-536.
Sanborn, A. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition, 112,* 98-101.
Sanborn, A., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117,* 1144-1167.
Savage, L. (1974). *Theory of statistics,* 2nd ed. New York: Dover.
Schervish, M. (1995). *Theory of statistics*. New York: Springer.
Schoenfield, M. (2017). Conditionalization does not (in general) maximize expected accuracy. *Mind, 504,* 1155-1187.
Scott, S. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience, 5,* 532-546.
---. (2012). The computational and neural basis of voluntary motor control and planning. *Trends in Cognitive Sciences, 16,* 541-549.
Shivkumar, S, Lange, R., Chattoraj, A., & Haefner, R. (2018). A probabilistic population code based on neural samples. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Biachi & R. Garnett (Eds.), *Advances in neural information processing systems, 31*, 1-10.
Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review, 63,* 129-138.
Skyrms, B. (1987). Dynamic coherence and probability kinematics. *Philosophy of Science, 54,* 1–20.
Spratling, M. (2016). A neural implementation of Bayesian inference based on predictive coding. *Connection Science, 28,* 346-383.
Staffel, J. (2019). *Unsettled thoughts*. Oxford: Oxford University Press.
Steele, K., & Stefánsson, H. (2016). Decision theory. In E. Zalta (Ed.), *Stanford encyclopedia of Philosophy* (Winter 2016). <https://plato.stanford.edu/archives/win2016/entries/decision-theory/>.
Stocker, A. (2018). Credo for optimality. *Behavioral and Brain Sciences, 41,* e244.
Stocker, A., & Simoncelli, E. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 4,* 578-585.
Stone, J. (2011). Footprints sticking out of the sand, part 2: Children's Bayesian priors for shape and Lighting direction. *Perception, 40,* 175–190
Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. Cambridge: MIT Press.
Todorov, E., & Jordan, M. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience, 5,* 1226-1235.
Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience, 7,* 907-915.
Trommershäuser, J., Körding, K., & Landy, M, eds. (2011). *Sensory cue combination*. Oxford: Oxford University Press.
Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics, 49,* 71-104.

Weisberg, J. (2009). Varieties of Bayesianism. In D. Gabbay, S. Hartman, and J. Woods (Eds.), *Handbook of the history of logic*, vol. 10. New York: Elsevier.

Tversky, A., & Kahneman, D. (1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.

van Fraassen, B. (1980). *The scientific image*, Oxford: Oxford University Press.

Walker, E., Cotton, R. J., Ma, W. J., & Tolias, A. (2020). A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, *23*, 122-129.

Wei, X.-X., & Stocker, A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience*, *18*, 1509-1517.

Weiss, Y., Simoncelli, E., & Adelson, E. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*, 598–604.

Wolpert, D. (2007). Probabilistic models in human sensorimotor control. *Human Movement Science*, *26*, 511-524.

Wolpert, D., & Landy, M. (2012). Motor control is decision-making. *Current Opinion in Neurobiology*, *22*, 1-8.