CONTRIBUTED PAPER

Can Bayesian Models of Cognition Show That We Are (Epistemically) Rational?

Arnon Levy

The Hebrew University of Jerusalem, Jerusalem, Israel
Email: arnon.levy@mail.huji.ac.il

Abstract

"According to [Bayesian] models" in cognitive neuroscience, says a recent textbook, "the human mind behaves like a capable data scientist." Do they? That is, do such models show we are rational? I argue that Bayesian models of cognition, perhaps surprisingly, don't and indeed can't show that we are Bayes-rational. The key reason is that they appeal to approximations, a fact that carries significant implications. After outlining the argument, I critique two responses, seen in recent cognitive neuroscience. One says that the mind can be seen as *approximately* Bayes-rational, while the other reconceives norms of rationality.

1. Introduction

The Bayesian approach plays a central role in present-day cognitive neuroscience. A current textbook presentation says that "according to [Bayesian] models, the human mind behaves like a capable data scientist (or crime scene investigator, or diagnosing physician, . . .) when dealing with noisy and ambiguous data" (Ma et al. 2023, 15). Given this vivid description and given that many philosophers view Bayesian inference as a pillar of rationality, especially in contexts involving "noisy and ambiguous data," it would seem that epistemology and science may be converging on a similar message. Or to put the matter more bluntly, cognitive science appears to show that we are (epistemically) rational.

But appearances are misleading, or so I will argue. Bayesian cognitive science does not tell in favor of the idea that we are Bayes-rational. Indeed, I will make a somewhat stronger claim: Bayesian models, in their present form, are *unable* to show such a thing. More specifically, Bayesian modelers, in most contexts, assume that the mind doesn't carry out full-blown Bayesian computations. Instead, they posit algorithms that *approximate* such computations. But this, even under the assumption that such models succeed admirably, is a far cry from showing that the brain "behaves like a capable data scientist."

After making the argument, I discuss two potential responses, both extracted from recent cognitive neuroscience. The first—the idea that the mind can be viewed as *approximating Bayesian rationality*—seems to me to be simply mistaken, given a reasonable understanding of the notion of approximation. The second—a retreat to a view on which the brain is *rational, given resource constraints and performance limitations*—may well be cogent from a methodological standpoint. But read as an attempt to reconceive the relevant (normative) notion of epistemic rationality, it appears undermotivated and, inasmuch as the present article's topic is concerned, somewhat beside the point.

Before I delve in, let me make a few comments to further clarify the main question and its significance. As the title states, my question is whether current Bayesian models *can* show that we are Bayes-rational? I will not commit to any very specific sense of possibility—but I am asking about the Bayesian program's *potential*: Assuming it succeeds, in more or less its current form, will it buttress the idea that our cognition displays a good match with Bayesian epistemic norms? Second, while the question has a general flavor, clearly it can be addressed at different cognitive capacities and may receive different responses, depending on the case. I cannot cover a wide range of cases in this short article. But I discuss one central case—intuitive physics—and I think many of the lessons generalize. Third, and perhaps most importantly, is my question a live question? Do Bayesian modelers of cognition even claim to show that humans exhibit Bayesian rationality, or am I tackling a strawman? To be sure, a direct, unqualified claim to the effect that Bayesian models show us to be rational is rare. That said, and as the quote from Ma, Körding, and Goldreich attests, statements in this spirit can be located. More generally, I think framing matters in these terms can help us ascertain the overall message stemming from Bayesian work in cognitive neuroscience—specifically, whether and to what extent it matches normative views in epistemology.

2. Bayesianism, philosophical and scientific

In philosophy of science and epistemology, Bayesianism is the view that a rational agent has degrees of belief (aka credences) that conform to the axioms of probability. Further, such an agent responds to evidence by updating her credences according to Bayes's formula[1]:

$$P(h|e) = \frac{P(e|h) * P(h)}{P(e)}$$

Where $P(x)$ denotes the probability of x and $P(x|y)$ denotes the conditional probability of x given y. Usually, $P(h)$ is termed the *prior*, $P(e|h)$ the *likelihood* and $P(h|e)$ the *posterior*. Updating consists in computing the posterior probability given the likelihood and the setting of $P(e)$ by incoming evidence.

Bayesianism, thus understood, embodies a *normative* claim about *epistemic* rationality: It suggests a standard for adjusting one's beliefs in the face of evidence. Importantly, epistemic rationality should be distinguished from instrumental

---

[1] This synchronic claim is sometimes labeled "probabilism," whereas "Bayesianism" often denotes the further (diachronic) claim concerning updating.

rationality, which concerns how one ought to act given one's aims—means-end reasoning. Speaking roughly, epistemic reasoning aims at figuring out *what is the case* while instrumental reasoning aims at how to *achieve some goal.* (This is not meant as an exhaustive distinction, but these are the only two kinds of rationality that will be discussed here.) Epistemic and instrumental reasoning are connected but distinct. I come back to this later; for now, note that Bayesianism—both in philosophy and cognitive neuroscience—is primarily addressed at epistemic rationality.

Bayesianism is widely accepted within philosophy of science and epistemology. Here, I too will assume it as a standard for epistemic rationality. This is primarily because I'll be examining work in cognitive science that seems to make this assumption. But it's also worth noting that there exist various philosophical arguments for Bayesianism (on which I can only briefly comment—see note 6) and that it holds attraction inasmuch as it is a formal, and therefore a clear and precise, theory of rationality.

With this in mind, consider now the Bayesian approach in cognitive neuroscience. This work is scientific and empirically oriented: It aims to account for cognition in actual human beings. Its primary tenet is that many cognitive capacities can be modeled as a process of Bayesian updating. In the past two decades, a wide range of phenomena have received such a treatment—from early perception (e.g., Bialek 2012), through concept and word learning (e.g., Lake et al. 2015) and including explicit reasoning tasks such as syllogistic thinking (e.g., Tessler et al. 2022) and "intuitive" physical reasoning (discussed further below).

The question I want to address is whether or not such models have the potential to show that humans, at least in some domains and under some contexts, are rational in a sense that parallels the normative claims made by Bayesians in philosophy. In other words, can the scientific models show that humans are, at least sometimes and in some domains, rational in the sense of conforming with Bayesian norms? To be clear, my question isn't about the empirical standing of Bayesian models, nor about whether we should be realists about them (Rescorla 2020; Colombo et al. 2021). Rather, as stated at the outset, I'm addressing the approach's *potential*: Will such models show that we are rational, if taken to describe cognition and assuming they are empirically successful? I will only discuss models that target personal-level reasoning, as those are the central and most straightforward locus of claims concerning rationality. Let me clarify that personal reasoning need not be explicit or conscious, and I do not make any assumptions on this score. I think many of the following points apply, without major modifications, to Bayesian models aimed at subpersonal cognition, and possibly to certain aspects of perception.[2] But to extend the argument in these ways would require more space than I have here.

On the Bayesian picture, the brain encodes a set of priors and likelihoods and reasoning consists in adjusting the posterior in the light of incoming (typically perceptual) evidence. However, this general idea has undergone a significant evolution over the last 15 years or so. Early models tended to assume relatively simple priors and regarded updating as a matter of computing Bayes's formula as such, given estimates of relevant priors and likelihoods (e.g., Körding and Wolpert 2004; Griffiths

---

[2] Perception is not typically understood in term of rationality. But it is often described and modeled in terms of *optimality*, and there are important parallels between these notions.

and Tenenbaum 2006). Over time, these early models faced criticism, partly alleging that the models were insufficiently grounded in underlying mechanisms (Jones and Love 2011; Bowers and Davis 2012). Concomitantly, there was a growing recognition amongst Bayesian cognitive scientists that in many contexts computing the strictly "true" Bayesian posterior is infeasible: The space of priors is often highly complex, a fact compounded by the need to continuously update in the face of incoming perceptual information (Sanborn and Griffiths 2010; Icard 2014; Griffiths et al. 2015). These developments led to the specification and investigation of a range of approximation algorithms, that is, computationally "cheaper" methods of calculating posterior probabilities. In the next section I illustrate these claims with some examples of concrete models and say more about approximation, as it is central to my eventual argument. At the moment let me simply highlight the overall structure of current Bayesian models: They describe reasoning as a probabilistic inference problem, the optimal solution to which is computing a posterior probability using Bayes's formula. But partly because carrying out the Bayesian computation in its full-blown form is computationally intractable, they posit that the brain computes an approximation to Bayes's formula. It is such approximation-based models that are then explored in further detail and tested against experimental data.

3. Approximations and their significance

Let me zoom in on approximations. To this end I will discuss models of "intuitive physics," that is, our capacity to make inferences about the properties of physical objects and the outcomes of physical scenarios. When these phenomena were first explored, more than a generation ago, the focus was, as a recent review puts it, "on misconceptions that people demonstrate when reasoning about the attributes and movements of objects and substances in the world" (Kubricht et al. 2017, 74). Explanations of such misconceptions tended to portray our intuitive physical reasoning as based on simple heuristics, guiding us relatively well in many cases, but also liable to frequent and systematic errors. More recently, Bayesian modelers have revisited these findings and have offered a rosier, so-called Noisy Newtonian picture, according to which "people's judgments are based on optimal statistical inference over a Newtonian physical model that incorporates sensory noise and intrinsic uncertainty about the physical properties of the objects being viewed" (Sanborn et al. 2013, 411).

   Let us look more closely at such Noisy Newtonian models. They combine two key ideas. First, that people make judgments—at least in the domain of mechanics—by assuming that the physical world behaves according to Newtonian principles. Second, that inferences are drawn in a probabilistic fashion: They presume that perception supplies uncertain information. Thus, suppose a subject observes a tower of bricks (as in the game *Jenga*) and is then asked, "will the tower remain stable or will it collapse?," according to the Noisy Newtonian picture, the subject proceeds by estimating the masses and relative positions of the bricks and then simulates, on the basis of Newtonian principles of mechanics, its stability to see whether it is likely to fall. This is done while assuming that input is only imperfectly correlated with the actual goings-on (this is the "noisy" part). Given this assessment, the subject then

provides an answer—in effect, an estimate of the posterior probability of the tower's collapsing.

Intuitive physical reasoning involves sophisticated computations if the Noisy Newtonian account is correct. But it is important to see that even so it falls far short of "truly" solving the probabilistic-physical problem. This is well illustrated by Battaglia et al.'s (2013) work, perhaps the best known of the Noisy Newton papers. For one thing, these authors do not presume that the cognitive system solves—in an analytical sense—Newton's equations. It runs a discrete simulation instead: The model appeals to the Open Dynamics Engine—a simulator of rigid body dynamics, which makes multiple simplifications. For another thing, it does not compute the actual Bayesian posterior, instead sampling from it multiple times, a form of Monte Carlo process. Indeed, even this is done in a very partial way—whereas ordinarily Monte Carlo simulations are run many times, Battaglia et al. assume that people run "only one or a few samples" (ibid., 18238).[3] Overall, then, the picture is of an agent that performs an inference with the rough form of a probabilistic physical computation, albeit with significant deviations from the full, "true" computation. And in this, Noisy Newtonian models of intuitive physical reasoning are not unusual. Indeed, they are a case in point: While current models of other phenomena will vary in the underlying computations, depending on what capacity is being modeled, for the most part they employ significant approximations, including limited sampling and related "short cuts" (e.g., Sanborn and Griffiths 2010; Leider et al. 2013; Vul et al. 2014).

Let us now return to the overall question—does a model of this sort vindicate a view on which our cognition, or the specific cognitive capacity of intuitive physical reasoning, is Bayes-rational, in the sense philosophers have in mind when discussing epistemic rationality? I think the answer is a rather definite no. For the Bayesian epistemologist advocates *exact conformity with* Bayes's formula, and not approximating it. Approximating Bayes is consistent with quite significant deviations from rationality, including classic probabilistic fallacies.[4]

The worry isn't mild nor is it merely abstract. To see this, consider the following example, drawn from very recent work on intuitive physical reasoning and that explicitly addresses a probabilistic fallacy. Critical of the Noisy Newtonian models described above, Ludwin-Peery et al. (2020) performed an experiment showing that subjects in such settings are prone to a conjunction fallacy, wherein they judge the conjunction of two events (e.g., the tower collapsing and the red bricks landing in front of the blue ones) to be more probable than one of the conjuncts (the tower collapsing). Follow-up work by advocates of the Noisy Newtonian viewpoint has attempted to explain this by assuming that subjects sometimes simulate only part of the physical scene—leading to greater efficiency at the expense of probabilistic consistency (Bass et al. 2022).

Now, as before, I am not concerned with empirical adequacy, namely with whether partial simulation is empirically well supported. Perhaps a Noisy Newtonian model

---

[3] In supplementary materials to their paper, Battaglia et al. (2013) estimate that actual subject's performance is consistent with 3–7 simulation runs.

[4] Moreover, approximations to Bayes do not, in general, meet the conditions assumed by most arguments for Bayesianism, such Dutch Book and Accuracy arguments (as noted by Williams 2021, §4.1.)

assuming partial simulation explains our facility for intuitive physics, warts (i.e., fallacies) and all. But it cannot be regarded as rational, in the sense that epistemologists who advocate Bayesianism have in mind. The conjunction fallacy is, as the name attests, a fallacy. Moreover, this isn't a matter of my choice of example. Recent work demonstrates that a process that approximates Bayesian reasoning, under conditions relevant in modeling human cognition, is liable to result in a variety of fallacies and biases—such as the unpacking fallacy, base rate neglect, and anchoring (Leider et al. 2013; Sanborn and Chater 2016).

Thus, while Bayesian models in cognitive science may appear, at a course-grained level, to portray the mind as Bayes-rational, in practice they model it in terms of approximation algorithms, and this turns out to leave them quite a ways off from what epistemologists have in mind in endorsing Bayesian updating as a standard of epistemic rationality.

4. Salvaging rationality?

I now want to consider two responses to the foregoing argument. Both appear, often implicitly, in recent cognitive neuroscience. The first suggests that while Bayesian modeling may not portray the mind as conforming strictly to cannons of Bayesian rationality, it does show it to be *approximately Bayes-rational*—and that is no small matter.[5] (And, perhaps, no more than one can expect, given a mind that is "housed" in a material, limited, fallible product of evolution by natural selection such as our brain.)

Such a line of thinking raises thorny conceptual issues: What exactly is mental approximation? Under what conditions does a cognitive process count as approximating a given computation? I do not know of any general, detailed accounts of cognitive approximation, and developing one lies beyond the scope of this article (but see Levy under review). But I want to suggest that at least one of the following two conditions are typically met when the term "approximation" is appropriately applied. First, "approximation" often refers to an attempt by an agent to solve a problem in a way that falls short of the fully correct solution, but is cost-effective, given the agent's purposes. Second, an approximation is a procedure or method that comes close to the fully correct solution—typically one that comes very close, or even as close as one pleases, under well-specified conditions.

Often, both conditions hold. Suppose a physicist is considering an *n*-body system in a Newtonian context and wants to know what the orbits of one or more of the bodies is (or maybe just whether the bodies have stable orbits). Because the problem is hard, perhaps even impossible, to solve analytically, she proposes an approximate solution: a way of computing the problem that is computationally tractable and known to provide a result that is arbitrarily close to the target solution. She might turn to a Taylor expansion, for instance, or run a Monte Carlo simulation. Indeed, many of the approximation algorithms appealed to by Bayesian models of the brain—including in the work on intuitive physics discussed in the preceding text—are drawn from engineering, physics, and computer science. It seems that at least some of the

---

[5] Sanborn and Griffiths (2010), in a paper that provides an extensive treatment of Monte Carlo approximations to Bayesian inference, describe their topic as pertaining to "the processes by which *human minds might approximate* optimal solutions to computational problems" (1145, emphasis added).

motivation for appealing to them resides precisely in the fact that they serve as approximations—meeting both conditions—in these "home" areas.

But notice that when such approximation methods are imported into the context of Bayesian cognitive neuroscience, neither of the conditions mentioned is typically met: Usually, modelers do not envision that subjects (i.e., those whose cognitive processes are being modeled) are making a deliberate attempt to cheaply solve a problem.[6] Nor is it the case that the method being employed comes arbitrarily close to the correct solution—to be precise, the conditions under which the mind is thought to execute many of the relevant approximations are substantially different from those under which the approximation is known to provide solutions that are close to the target solution. In the example discussed in the preceding text, for instance, the cognitive system is presumed to run "only one or a few samples"—orders of magnitude less than any acceptable simulation in physics would run. Indeed, some argue that this fact—that cognitive approximations to Bayes fall significantly short of Bayes proper—can explain various biases and systematic errors to which humans are known to be prone (Sanborn and Chater 2016; Gershman 2021).

A worrying possibility is that what the approximations response is really grounded in a relatively simple but seductive mistake: moving from the claim that the cognitive process at issue is well captured by *a model that includes an approximation* to Bayesian inference to the claim that the *mind approximates* Bayesian inference. This is an erroneous inference: The mind doesn't—except perhaps in rare cases—approximate anything; it simply works as it does. Nor need it be the case, according to Bayesian models, that the mind comes especially close to Bayesian inference, as the number-of-samples example just noted attests. The modeler's description of the model as involving approximations-to-Bayes is justified inasmuch as the algorithms appealed to are used as approximations (in the sense that they meet the two conditions specified previously) in the context from which the algorithm is drawn—be it in physics, statistics, or computer programming. But this does not license the claim that our mind is *approximating* Bayes.

The second response I'll discuss can also be discerned in recent cognitive science. It involves an adjustment of the notion of rationality. A recent paper by Leider and Griffiths (2020) nicely illustrates the idea. These authors highlight accounts of cognitive phenomena, akin with those discussed in the preceding text, that appeal to approximations and other shortcuts, and suggest that "the rational use of limited resources ... provid[es] a unifying framework for explaining the corresponding phenomena" (ibid., 2). This so-called resource rationality approach construes cognition as aiming to maximize its use of computational and other resources, given the information at its disposal and taking account of the agent's learned experience, goals, and opportunities. Leider and Griffiths explicitly contrast this with the "classic notion of rationality, according to which people ... handle uncertainty according to

---

[6] One of the papers described in the preceding text may be an exception: Bass et al. (2022) explain the conjunction fallacy in intuitive physics as involving partial mental simulation of the physical scenario, wherein subjects simulate only some of the objects in the scene. This, they suggest, is "key to efficient implementations of useful commonsense physical reasoning" (ibid., 4–5) and refer to this at one point as a "useful approximation" (ibid., 16). But it is unclear whether they think of this as a deliberate approximation employed by the agent. And, in any event, this is an outlier and most appeals to approximation do not seem to involve explicit shortcuts in reasoning.

probability theory" (ibid.). As they note (ibid., 4), a number of recent authors in the field have made similar suggestions.

Now, the notion of resource rationality, like any notion of rationality that takes into account the availability of resources and performance limitations, is by its very nature an instrumental notion: It centers on the best use to which the agent's finite cognitive means should be put, relative to a set of goals. This is evident in Leider and Griffiths's treatment, as well as in others, inasmuch as they posit that utility is one of the constitutive determinants of resource rationality (ibid., section 2). So, the appeal to resource rationality, *as opposed to* the "classic" notion of rationality, is in effect a rather dramatic shift in focus, from epistemic to instrumental rationality.

Such a suggestion can be understood as entirely descriptive, that is, as saying that the category of humans in general, or some cognitive system in particular, is best modeled in terms of maximizing the use of relevant resources, given various constraints, opportunities, and the like. I will not attempt to evaluate this descriptive-methodological suggestion. It may be, for all we know at present, that some such notion as resource rationality can serve as a useful umbrella under which many cognitive phenomena can be studied. Be that as it may, clearly such a research program does not aim for, and will not result in, a vindication of the idea that human cognition conforms with Bayesian norms of rationality, in the narrower epistemic sense.

Instead, what I discuss now is a different tendency—seen at several points in Leider and Griffiths, as well as, *mutatis mutandis*, in other authors—to treat resource rationality as a *normative* notion, and, as such, as a candidate to *replace* the "classic notion of rationality." As they put it toward the end of their paper: "Research is now revisiting the debate about human rationality with resource rationality as a more realistic normative standard" (ibid., 13).

It seems to me that this suggestion can be read in two ways, and I'd like to offer a few comments on both. The first reading has it that the "classic," Bayesian standard is still, from a purely epistemic standpoint, appropriate but that the everyday cognition of humans is rarely able to live up to this standard and should therefore be judged according to a more relaxed, pragmatic standard. In response to such a reading, I think we should at least say the following: While there may be considerations, based in cognitive science, that merit such a "forgiving" attitude in everyday epistemic contexts, there are still many contexts in which the full-blown, Bayesian notion of rationality is needed and appropriate. One such context, and indeed an important one, is scientific reasoning. There may well be others. The key point is that even if cognitive science can contribute to our understanding of when a more forgiving epistemic attitude is warranted, this does not involve an *abandonment* of Bayesian rationality in favor of a notion such as a resource rationality. Rather, it would amount to the claim that latter notion is the standard against which to judge a well-specified subset of human performance given relevant conditions and appropriate expectations.

The second reading is stronger and has it that we should *replace* Bayesian rationality with resource rationality, or more generally with a pragmatic-instrumental notion that, presumably, embodies an appropriate trade-off between epistemically good outcomes and feasibility. Relatives of this proposal have appeared, over several decades, in the literature on bounded rationality (e.g., Gigerenzer 2008). It suggests that traditional epistemology is premised on an inappropriate notion of rationality, one that doesn't offer a plausible picture of how real people, in the real

world, ought to think. Notice that the claim isn't (only) one about how real people in the real world in fact think, nor is it a claim about how real people can reasonably be expected to perform, epistemically speaking. It is a *normative* claim about how they *ought* to think.

For my own part, I am doubtful of this line of thought—it seems to me that we should retain the traditional, "unbounded," notion of rationality, if only as a bar for optimal epistemic performance. The only plausible arguments for adopting an alternative notion of rationality, it seems to me, depend on conflating epistemic rationality with instrumental rationality, a conflation we have independent reasons to resist (Kelly 2003; Christiansen 2021). But I will not elaborate on that here. I only note that, when viewed from the standpoint of our initial question, the suggested shift from the Bayesian standard toward resource rationality is somewhat beside the point. That question, to recall, concerned the match between Bayesian modeling and a commonly accepted standard of *epistemic* rationality, whereas the present suggestion is premised on a shift in the notion of rationality, effectively abandoning a purely epistemic standard for a notion that is pragmatic and instrumental in character. Put differently, we began by asking about the extent to which Bayesian cognitive science vindicates the thought that we are *epistemically* rational, *sensu* epistemic Bayesians. Advocates of resource rationality do not attempt to provide an answer, but rather to change the question.

References

Bass, Illona, Kavin A. Smith, Elizabeth Banowitz, and Tomer D. Ullman. 2022. "Partial Mental Simulation Explains Fallacies in Physical Reasoning." *Cognitive Neuropsychology* 38 (7–8):413–24.

Battaglia, P., Jessica B. Hamrick, and Joshua B. Tenenbaum. 2013. "Simulation as an Engine of Physical Scene Understanding." *PNAS* 110(45):18327–32.

Bialek, William. 2012. *Biophysics: Searching for Principles*. Princeton, NJ: Princeton University Press.

Bowers, Jeffrey S., and Colin J. Davis. 2012. "Bayesian Just-So Stories in Psychology and Neuroscience." *Psychological Bulletin* 138 (3):389–414.

Christiansen, David. 2021. "The Ineliminability of Epistemic Rationality." *Philosophy and Phenomenological Research* 103 (3):501–17.

Colombo, Matteo, Lee Elkin, and Stephan Hartmann. 2021. "Being Realist about Bayes, and the Predictive Processing Theory of Mind." *British Journal For Philosophy of Science* 72 (1):185–220.

Gershman, Samuel. 2021. *What Makes Us Smart: The Computational Logic of Human Cognition*. Princeton, NJ: Princeton University Press.

Gigerenzer, Gerd. 2008. *Rationality for Mortals: How People Cope with Uncertainty*. New York: Oxford University Press.

Griffiths, Tom, and Joshua B. Tenenbaum. 2006. "Optimal Predictions in Everyday Cognition." *Psychological Science* 17 (9):767–73.

Griffiths, Tom L., Falk Leider, and Noah D. Goodman. 2015. "Rational Use of Cognitive Resources: Levels of Analysis between the Computational and the Algorithmic." *Topics in Cognitive Science* 7(2) :217–29.

Icard, Thomas. 2014. "Toward Boundedly Rational Analysis." *Proceedings of the Annual Meeting of the Cognitive Science Society* 36. https://cogsci.mindmodeling.org/2014/papers/118/

Jones, Matt, and Bradley C. Love. 2011. "Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition." *Behavioral and Brain Sciences* 64:169–231.

Kelly, Thomas. 2003. "Epistemic Rationality as Instrumental Rationality: A Critique." *Philosophy and Phenomenological Research* 66 (3):612–40.

Körding, Konrad P., and Daniel Wolpert. 2004. "Bayesian Integration in Sensorimotor Learning." *Nature* 427:244–47.

Kubricht, James R., Keith J. Holyoak, and Hongjing Lu. 2017. "Intuitive Physics: Current Research and Controversies." *Trends in Cognitive Science* 21 (10):749–59.

Lake, Brendan, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. "Human-Level Concept Learning through Probabilistic Program Induction." *Science* 350:1332–38.

Leider, Falk, and Tom Griffiths. 2020. "Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *Behavioral and Brain Sciences* 43 (e1):1–60.

Leider, Falk, Tom Griffiths, and Noah Goodman. 2013. "Burn-In, Bias, and the Rationality of Anchoring." In P. Bartlett, F.C.N. Pereira, Leon Bottou, Chris J.C. Burges, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*. Lake Tahoe, NV: Conference on Neural Information Processing Systems.

Levy, Arnon. under review. "Approximating Bayes? On the Notion of Approximation in Bayesian Cognitive Science."

Ludwin-Peery, Eithan, Nieil Bramley, Ernst Davis, and Todd M. Gureckis. 2020. "Broken Physics: A Conjunction-Fallacy Effect in Intuitive Physical Reasoning." *Psychological Science* 31 (12):1602–11.

Ma, Weiji, Konrad Kording, and Daniel Goldreich. 2023. *Bayesian Models of Perception and Action*. Cambridge, MA: MIT Press.

Rescorla, Michael. 2020. "A Realist Perspective on Bayesian Cognitive Science." In *Inference and Consciousness*, edited by A. Nes and T. Chan, pp. 40–73. New York: Routledge.

Sanborn, Adam and Nick Chater. 2016. "Bayesian Brains without Probabilities." *Trends in Cognitive Science* 20:883–93.

Sanborn, Adam N., and Thomas L. Griffiths. 2010. "Rational Approximations to Rational Models: Alternative Algorithms for Category Learning." *Psychological Review* 117 (4):1144–67.

Sanborn, Adam N., Vikash K. Mansinghka, and Thomas L. Griffiths. 2013. "Reconciling Intuitive Physics and Newtonian Mechanics for Colliding Objects." *Psychological Review* 120 (2):411–37.

Tessler, Michael Henry, Joshua B. Tenenbaum, and Thomas L. Griffiths. 2022. "Logic, Probability, and Pragmatics in Syllogistic Reasoning." *Topics in Cognitive Science* 14(3):574–601.

Vul, Edward, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2014. "One and Done? Optimal Decisions from Very Few Samples." *Cognitive Science* 38:599–637.

Williams, Daniel. 2021. "Epistemic Irrationality in the Bayesian Brain." *British Journal for Philosophy of Science* 72(4):913–938.