**Metaphysics of The Bayesian Mind**

 Recent years have seen a Bayesian revolution in cognitive science, a surge of work that models human learning and inference as following principles of Bayesian probabilistic inference. This work should be of interest to metaphysicians of science, whose naturalist project involves working out the metaphysical implications of our leading scientific accounts, and in advancing our understanding of those accounts by drawing on the metaphysical frameworks developed by philosophers (Ladyman & Ross, 2009; Sider, 2020). Toward these ends, in what follows I articulate and defend a metaphysics of the Bayesian mind.

My central claim is that the Bayesian approach supports a novel empirical argument for *normativism*, the thesis that belief has a normative essence. I begin by reviewing the recent Bayesian turn in cognitive science (§1). I then explain how I understand the normativist thesis (§2). On the version of normativism that is my focus, the norms in question are those of rationality. This gives the view a kind of resonance with positions defended by Davidson (1980) and Dennett (1987), given their emphasis on rationality. Other authors have noted an affinity between Bayesian cognitive science and Davidson- and Dennett-style views in the philosophy of mind (Rescorla, 2013; Rescorla, 2019; Oaksford, 2014), but here I develop a particular metaphysical understanding of the connection.

 After clarifying the normativist thesis, I present my argument in its favor (§3). Central to my argument is the use that Bayesian cognitive scientists make of Marr’s (1982) levels of analysis, which brings with it a commitment to *multiple realizability*. Relying on the *causal powers subset account* of multiple realization (Wilson, 1999; Shoemaker, 2001), I explain how this form of multiple realizability supports the normativist conclusion. In the final section (§4), I address how my metaphysical account bears on questions in the philosophy of cognitive science, including how to understand empirical findings of human irrationality that are often thought to pose a serious problem for both Bayesian cognitive science and normativism (Tversky & Kahneman, 1974; Kahneman et al. 1982), and how to think about the explanatory role of normativity within cognitive science (Danks, 2008, 2018).

1. Bayesian Models of Cognition

We begin with a review of Bayesian cognitive science. The Bayesian approach is quite general and has been applied to a wide variety of cognitive domains. It is generally thought to have been most successful in shedding light on perception (Knill & Richards, 1996; Rescorla, 2015), but throughout this paper I will take as my focus central cognitive processes involving the fixation of belief. Examples include Bayesian models of causal learning and reasoning (Griffiths & Tenenbaum, 2005; Holyoak & Cheng, 2011), categorization (Anderson, 1991; Sanborn et al., 2010), memory (Anderson & Milson, 1989; Shiffrin & Steyvers, 1997), language acquisition and processing (Chater and Manning, 2006; Xu & Tenenbaum, 2007), and intuitive physics (Sanborn et al., 2013; Gerstenberg & Tanenbaum, 2017).

A standard way to motivate the Bayesian approach is to observe that various cognitive tasks involve a kind of inherent uncertainty and inductive inference that goes beyond what is directly observed (Oaksford & Chater, 2007; Tenenbaum et al., 2011). In that case, the same sorts of considerations that have led many philosophers to adopt Bayesian confirmation theory as the correct normative view in the philosophy of science can be used to motivate Bayesian approaches to cognitive modeling, at least if we adopt an assumption of “optimality” regarding the domains being modeled—an assumption to be discussed momentarily.

 Now, the idea that causal learning, for instance, involves inductive inference is hardly new, as readers of Hume will recognize. But advances in computer science, statistics, artificial intelligence, and aligned fields in understanding Bayesian updating procedures and approximations thereof have paved the way for the development of more sophisticated and rigorous psychological models than were previously available (Gershman et al., 2015). And so Holyoak and Cheng (2011) describe the introduction of Bayesian approaches as the most important methodological advancement in the creation of a “new synthesis” in work on causal learning and inference. A new synthesis that, incidentally, has surpassed an earlier associative framework inspired by Hume (Holyoak & Cheng, 2011: 136-137).

It will be helpful to have a case study that we can consider in some detail and return to throughout the paper, so consider Anderson’s (1990, 1991) *Rational Model of Categorization*, an early paradigm of the Bayesian approach, and a model that has inspired later research to be discussed below. Categorization is the process by which subjects partition objects into categories or clusters, like categorizing Rover as a Labrador Retriever, or dividing a set of stimuli with various observable features into distinct clusters in an unsupervised categorization task. Anderson’s model falls under the broader research program of *Rational Analysis* (Anderson, 1990; Chater & Oaksford, 1999; Lieder & Griffiths, 2019), the methodology for which consists of six steps.

1. *Goals*: specify the goals of the cognitive system.
2. *Environment*: develop a formal model of the environment to which the system is adapted.
3. *Computational limitations*: make minimal assumptions about computational limitations.
4. *Optimization*: derive the optimal behavior function, given (1)-(3).
5. *Data*: examine the empirical evidence to see whether the predictions of the behavior function are confirmed.
6. *Iteration*: repeat, iteratively refining the theory.

Applying this program to the case at hand, Anderson proposes that the goal of categorization is the optimal prediction of unobserved features of stimuli (step 1). By categorizing Rover as a Lab, for example, you can predict that he is likely to be friendly. Making certain assumptions about the stimuli to be categorized (step 2), such as that their features are independent of one another, and making minimal assumptions about people’s computational limitations (step 3), Anderson proposes that the optimal way to achieve this goal involves a form of Bayesian belief updating (step 4).

I pause here because this is a point of contact with the normativist thesis to be discussed in §2. On the Rational Analysis approach, normative models of how subjects *should* reason can guide psychological research. They can double as descriptive models of how they *do* reason, given the empirical premise that human cognition is adapted to perform at optimal or near-optimal levels, whether this is the result of evolution or learning. If we help ourselves to the assumption that Bayesian normative theory is correct, this fourth step of the program leads directly to Bayesian cognitive science.[[1]](#footnote-1)

Of course, whether Bayesian normative theory is correct is a matter of dispute (Glymour, 1980). In the present paper, I do not want to engage this debate directly; it would distract from my main goals. Instead, I will simply assume that some sort of Bayesian normative perspective is correct—a controversial but fairly widely shared view—and then work out the metaphysical consequences of Bayesian cognitive science in light of this unargued assumption. In his discussions of Bayesian cognitive science, Danks (2014, 2018) warns against conflating the psychological and the normative, noting for instance that an opponent of Bayesian normative theory could still endorse Bayesian models like Anderson’s. I grant Danks’ point but insist I am not guilty of such a conflation: I am explicit about what I am assuming.

 Turning next to the empirical evidence (step 5), the Rational Model correctly predicts a number of classic experimental findings observed in human categorization; I will limit myself to two that will serve as running examples. First, humans are sensitive to the *central tendency* of a category, so that the reliability with which a stimulus is classified as belonging to a category drops off as a function of its distance from the average or central tendency of that category (Medin & Schaffer, 1978). For instance, subjects are more likely to categorize a cartoon face as belonging to a given cluster of faces the closer it is to the average or central tendency of that cluster with respect to observable features like nose length, forehead height, etc. (Reed, 1972). The Rational Model correctly predicts this result (Anderson, 1991: 415-417).

Second, some categories are *linearly separable* in that a straight line in a multidimensional feature space can be used to divide instances from non-instances. Other categories cannot be so divided, and so are *linearly* *nonseparable*. The Rational Model correctly predicts that in certain circumstances, linearly nonseparable categories are easier for subjects to learn than linearly separable ones, in contrast with standard prototype models of categorization which entail that linearly separable categories should be inherently easier to learn (Anderson, 1991: 418-419).[[2]](#footnote-2)

 We will get to the sixth and final step of Anderson’s methodology in §§3-4 when we consider how the account might be refined in response to empirical findings the Rational Model fails to predict. At this point, however, I want to bring normativism into the discussion.

2. The Normativist Thesis

 The idea that the mental is in some important sense normative has been explored by a number of leading philosophers (Sellars, 1956; Brandom, 1994; Wedgwood 2007a; Boghossian, 2008; McHugh & Whiting, 2014). There are different ways of developing the thought, but on the version defended here the governing norms are those of rationality.[[3]](#footnote-3) This gives the resulting normativist view a kind of resonance with the position defended by Davidson (1970), who writes of the “constitutive ideal of rationality” governing the mental, and with that of Dennett (1987), who takes the intentional stance to operate according to an “assumption of rationality.” There is also another point of connection with these authors worth mentioning here at the outset.

 Some philosophers have taken normativism to ensure a kind of irreducibility of the mental. Davidson (1970: 146), when he invokes the constitutive ideal of rationality, does so in the course of arguing that mental types cannot be identified with or otherwise reduced to physical types, even as mental tokens are identical with physical tokens—a version of nonreductive physicalism. Dennett (1981, 1991) sees the intentional stance, with its assumption of rationality, as picking up on a higher-level *real pattern* that is missed by the lower-level design and physical stances. As we will see in the sections that follow, a similar nonreductive element will be central to my position.

 In the present section my aim is just to clarify how I understand the normativist thesis, leaving the argument in its favor for §3. Toward this end, I draw on the work of Wedgwood (2007a, 2007b, 2017). My argument in favor of normativism in §3 will be very different from Wedgwood’s, but my understanding of the position is inspired by and similar to his (while differing in a few details).

 Normativism asserts that belief has a normative essence. The discussion in this paper does not require any very specific view of the nature of essences, but for the sake of concreteness suppose they are provided by *real definitions* that come in the form of Russellian propositions, and so are composed of worldly entities like objects, properties, and relations (Rosen, 2015). We could then say that belief has a normative essence just in case its real definition includes a normative property as a constituent. If we take rational *ought* properties to be paradigmatic examples of such normative properties, we can add that normativism would be true if, for example, belief’s real definition includes the proposition that a subject of belief is capable of reasoning in the way they rationally *ought* to reason. There is an important debate over which normative notion is most basic—reasons, oughts, something else (Broome, 2015)—but here I will stay officially neutral on the matter while using *ought* properties as my example.[[4]](#footnote-4)

 To see how belief might have a normative essence, consider that properties or state types in general have various different *causal powers*. I use this term expansively to cover both powers to cause effects of a given type, as well as powers (“liabilities”) to be effects of causes of a given type. Perhaps powers can be modeled as functions from circumstances to effects (Shoemaker, 1981). At any rate, I am neutral on their underlying metaphysics, and leave open whether powers are fundamental or analyzable in other terms (counterfactuals, structural equations, etc.).

 Consider then belief as a property or state type. Its various causal powers can be divided into three mutually exclusive and exhaustive categories. There are the *rational powers*, which are those exercised when some norm of theoretical or practical rationality is satisfied.[[5]](#footnote-5) There are the *irrational powers*, which are those exercised when some norm of theoretical or practical rationality is violated.[[6]](#footnote-6) And finally there are the *arational powers*, which are those causal powers of belief that are neither rational nor irrational.

For the purposes of this taxonomy, I take the relevant norms of theoretical rationality to include those of Bayesian epistemology or confirmation theory, and the relevant norms of practical rationality to include those of Bayesian decision theory. I say “include” to allow that there may be additional norms beyond these, but how such additional norms might fit into the present argument is not something I will try to address. To give examples of the Bayesian norms: I assume that a subject’s beliefs at a given time should satisfy the axioms of the probability calculus, I endorse some sort of conditionalization principle prescribing how to transform beliefs in response to new evidence, and I accept some form of expected utility principle saying that agents should choose the action that would maximize their expected utility.[[7]](#footnote-7) This of course leaves many normative questions open, but my aim here is not to settle such questions.

I have spoken of belief in an undifferentiated sense up to this point, but there is a familiar distinction between *outright belief*, like believing that it will rain, and *degrees of belief* or *credences*, like having a credence of 0.83 that it will rain. The normativist position defended in this paper concerns degrees of belief or credences, for it is they that figure in Bayesian models in cognitive science. It may be possible to extend my argument to outright belief, but doing so is no part of the present project. I will continue to use the term “belief” at various points in what follows, but this should always be understood as shorthand for degrees of belief or credences.

With this as setup, here is the central claim in next section’s argument.

**Essential Powers**: Possessing rational powers is essential to belief (i.e., credence state types), while possessing irrational or arational powers is not.

In the remainder of this section, I want to try to provide an intuitive gloss on this thesis. Human beings fall prey to various forms of irrationality. We engage in wishful thinking and self-deception, commit the conjunction fallacy, neglect base rates, have a tendency toward confirmation bias, and more. But are the irrational powers manifested in these various effects *essential* to belief? No they are not, if **Essential Powers** is true.

 Consider a familiar sort of multiple realizability thought experiment. Imagine an alien species or artificial intelligence that has an internal state occupying a causal role much like that of belief in humans, but lacking our irrational powers. The beings are immune to our all-too-human irrational foibles: they do not commit the conjunction fallacy, neglect base rates, etc. Would this disqualify them from having beliefs at all? Intuitively not. No *Star Trek* viewer regards Spock as being *too logical* to believe anything. But then the irrational powers that our human belief tokens possess must not be essential to belief as a type. Or at least this is what **Essential Powers** says.[[8]](#footnote-8)

A similar conclusion applies to arational powers. Humans take longer on average to recognize valid modus ponens arguments than valid modus tollens ones (Rips, 1994: 177), where this is neither to our rational credit nor discredit. More generally, cognitive science has discovered all sorts of timing effects along these lines. But an alien species or artificial intelligence that was wired differently so as to be faster at modus tollens, or to show different patterns of timing effects in general, would not thereby be disqualified from having beliefs, for arational powers are not essential to belief. Or again at least this is what **Essential Powers** says.

But finally, consider the rational powers. Perhaps no single, very fine-grained rational power is essential for belief, and so perhaps the aliens or artificial intelligences could have beliefs even if they lacked the power to recognize the validity of the Celarent syllogism, for example.[[9]](#footnote-9) But when sufficiently many rational powers are missing, and so a being is incapable of drawing almost any of the inferences it rationally ought to draw, this does disqualify it from having beliefs at all, if **Essential Powers** is right. “If we cannot find a way to interpret the utterances and other behavior of a creature as revealing a set of beliefs largely consistent… we have no reason to count that creature as rational, [or] as having beliefs,” writes Davidson (1973: 137).

 It will be helpful to have a schematic illustration of the claim of **Essential Powers** that we can return to in what follows. Consider a token of a certain credence type *H* (for *h*ypothesis). Suppose the causal profile of this *H*-token is {1R, 2R, 3I, 4I, 5A, 6A}, where different numerals represent different causal powers and the subscripted letters indicate whether a power is rational (‘R’), irrational (‘I’), or arational (‘A’). Next, suppose in accordance with **Essential Powers** that the powers essential to *H* as a type are just the subset of rational powers, {1R, 2R}. It then follows that a distinct *H*-token with a distinct causal profile, of say {1R, 2R, 38I, 45I, 51A, 64A}, could still belong to *H* as a type since it has

the essential subset {1R, 2R}, even though it differs in its irrational and arational powers.

 I will say more to clarify the normativist thesis further as we go along. But at this point we are ready for the main argument.

3. Subset Realization & The Normativist Argument

 In a debate with Wedgwood, Rey (2007) complains that normativists too often ignore empirical psychology and focus instead on things like a priori reflection on radical interpretation (Davidson, 1984) or conceptual analysis of the notion of “belief” (Wedgwood, 2007b). Against this tendency, in this section I present the first empirical defense of normativism that I am aware of.[[10]](#footnote-10) My argument says we should accept normativism because it follows from Bayesian cognitive science, which in turn we should accept because of its empirical success.[[11]](#footnote-11) Spelling out the reasoning a bit more, the argument says the use that Bayesians make of Marr’s (1982) levels of analysis, when taken together with our best philosophical understanding of the multiple realizability claim at the heart of Marr’s account, entails **Essential Powers**, and so supports normativism. The burden of the section is to make this case.

 We start with Marr’s levels. Bayesians typically pitch their accounts at Marr’s highest, *computational level*, which concerns what an information-processing system is doing and why (Anderson, 1990; Chater & Oaksford, 2007; Tenenbaum et al., 2011). In general, a computational model will specify a mathematical function characterizing the type of system being modeled; in the Bayesian case, this function will describe a kind of Bayesian updating. Anderson’s Rational Model of Categorization (§1) is a paradigmatic example—it is explicitly offered at Marr’s computational level.

Marr’s intermediate, *algorithmic level* concerns how the task specified at the computational level is accomplished, including how inputs and outputs are represented by the system, and the algorithm used. While some Bayesians have made a point of ignoring algorithmic realization, especially in the early years of the research program, more recently there has been a push to develop accounts at this level by looking to algorithms that approximate ideal Bayesian inference (Tenenbaum et al., 2011; Sanborn & Chater, 2016; Lieder & Griffiths, 2019). This work will become our focus momentarily. Finally, Marr’s bottom, *implementational level* concerns how this is all realized in the hardware of the brain (or other physical system). The discussion that follows will often lump the algorithmic and implementational levels together while focusing on examples taken from the algorithmic level.

 An attraction of Marr’s levels is that they allow for multiple realizability. A single computational model might be realized by different algorithms in different systems, while in turn the same algorithm might be realized in different forms of hardware. Because the nature of the realization relation has been a topic of intense philosophical scrutiny (Melnyk, 2003; Polger & Shapiro, 2016), this is a promising candidate area where metaphysicians of science might hope to shed light. Toward this end, I will draw on the *causal powers* *subset account* of realization developed by Wilson (1999, 2011), Shoemaker (2001, 2007), and others. I have defended the subset account at some length previously (omitted). Here I will mostly just make use of it rather than rehash my defense, although I note that how well the account fits with the Bayesian approach counts in its favor.

 On my preferred way of understanding the subset account, it can be divided into two steps. First, start with a causal theory of some higher-level type *H*. Using the Ramsey-Lewis method (Lewis, 1970), you derive from this causal theory a *functional definition* of *H*, understood as a specification of a set of causal powers the possession of which is metaphysically necessary and sufficient for being in *H*.[[12]](#footnote-12) Applying this first step to the case at hand, my proposal is to regard computational-level Bayesian models as functionally defining the credence states they range over. For instance, Anderson’s Rational Model might be used to functionally define credences about categorization. This can be understood as a version of *psychofunctionalism* about credences (Block, 1980). Psychofunctionalists in general take mental states to be functionally defined by empirical psychological theories; my particular version here takes those theories to be Bayesian.[[13]](#footnote-13) Given the rational nature of Bayesian computational models, the functional definitions that are derived from them will specify *exclusively rational powers*.[[14]](#footnote-14)

 Proceeding to the second step of the account, a lower-level state type *L* realizes *H* just in case *L*’s overall set of causal powers includes as a proper subset those powers specified in *H*’s functional definition.[[15]](#footnote-15) Applying this second step to the case at hand, I propose that an algorithmic or neural state type *L* realizes a credence state type *H*, as it is functionally defined by a Bayesian computational model,just in case *L*’s causal profile includes as a proper subset all the causal powers that the Bayesian computational model attributes to *H*.[[16]](#footnote-16) This captures in a straightforward, perhaps even flatfooted way the idea that higher-level states are more “abstract” than their lower-level realizers: if you take *L*’s causal profile and abstract certain causal powers away, you end up with just those powers specified in *H*’s functional definition. Illustrating the idea schematically, suppose a Bayesian computational model functionally defines *H* in terms of the set of rational powers {1R, 2R}. Suppose in addition that the lower level state type *L* belonging to the algorithmic-level has a causal profile of {1R, 2R, 3I, 4I, 5A, 6A}. Then *L* realizes *H*.

 A crucial feature of my version of the subset account is that it does *not* say that tokens of a higher-level type *H* possess *only* those causal powers described by *H*’s defining causal theory (omitted). Rather, what the defining causal theory tells you is which causal powers are *essential* to *H* as a type. Any particular token of *H* will in addition inherit from its specific lower-level realizer various causal powers that are not essential to *H* as a type.[[17]](#footnote-17) One way of motivating this element of the view is by connecting it to the sort of token identity theory embraced by various nonreductive physicalists (Davidson, 1970; Fodor, 1974).[[18]](#footnote-18)

 Suppose *H* is multiply realized by *L* and *L*\*. If the token identity theory is true, then some *H*-tokens will be identical with *L*-tokens, and so will have the same set of causal powers as those *L*-tokens, {1R, 2R, 3I, 4I, 5A, 6A}. Meanwhile, other *H*-tokens will be identical with *L*\*-tokens, and so will have the same set of causal powers as those *L*\*-tokens, which we can suppose is {1R, 2R, 38I, 45I, 51A, 64A}. In that scenario, the variously realized *H*-tokens will differ in their irrational and arational powers: some *H*-instances will possess the irrational power 3I, others won’t; some will possess the arational power 64A, others won’t; etc. However, these irrational and arational idiosyncrasies can be abstracted away to get at what unites the various *H*-instances together, to get at what is essential to *H* as a type: the subset of rational powers {1R, 2R}.[[19]](#footnote-19)

You can think of the view in terms of Dennett’s (1991) notion of *real patterns*. A central part of what science does is abstract away from the endless dissimilarities between entities to focus on their underlying similarities, thereby picking up on the real patterns that unite them. In studying cognition, just what is to be abstracted away and what is to be left in place as part of the pattern? According to the present proposal, it is the rational powers that are essential to credence state types, it is the rational powers that comprise the higher-level real patterns that Bayesian computational models are picking up on, while the various idiosyncratic irrational and arational powers that specific credence tokens possess, because of how they happen to be realized, are the dissimilarities to be abstracted away.[[20]](#footnote-20)

 Applying the account to our example of the Rational Model of Categorization, recall that the model is able to account for a wide range of experimental findings, including those pertaining to central tendencies and linear nonseparability (§1). One thing the model fails to predict, however, is that human categorization is subject to an order effect (Medin & Bettger, 1994). How people partition a set of stimuli depends on the order in which they are presented, with early trials leading subjects down a path to certain partition choices that get locked in as later stimuli arrive. This is in violation of Bayesian norms. An ideal Bayesian categorizer would be insensitive to the order in which stimuli are presented, and so this is an example of an irrational power.[[21]](#footnote-21)

While continuing to embrace Anderson’s Rational Model at the computational level, Sanborn and colleagues (2010) propose that the key to explaining the observed order effect is to descend to the algorithmic level. They explore Monte Carlo algorithms developed in computer science and statistics, like Gibbs algorithms and particle filters, in which ideal Bayesian inference is approximated by repeatedly sampling from a given probability distribution as more data are observed (cf. Tenenbaum et al., 2011; Sanborn & Chater, 2016; Lieder & Griffiths, 2019).

Framing their work in terms of the subset account, what Sanborn and colleagues show is that first, the states described by the different algorithms all possess the various rational powers that Anderson’s computational model attributes to credence states, regarding for instance central tendencies and linear nonseparability. After all, if a given algorithm did not possess such rational powers, it would not provide a promising candidate realizer for the Rational Model. But then second, Sanborn and colleagues show that the algorithms differ with respect to the *additional* causal powers they possess—in particular, they differ with respect to the order effects they generate. One form of particle filter matches human performance especially well, while other algorithms do not. For instance, the algorithm that Anderson (1991) himself proposed, dubbed the “local MAP,” overpredicts the effect and so is more sensitive to stimulus order than human beings are.

This work represents a general trend, an emerging approach that Bayesians are pursuing across a wide range of cognitive domains (Levy et al., 2009; Vul et al., 2009; Lieder et al. 2012; Denison et al. 2013; Sanborn & Chater, 2016; Lieder & Griffiths, 2019). That is, they start with a Bayesian computational model that accounts for human performance within a given domain very well in general; a model that describes a certain subset of essential rational powers, as I would put it. But then they look for additional causal powers, beyond this rational subset, to find clues about algorithmic-level realization.[[22]](#footnote-22) Rescorla (2019: 51-52) cites this work to make the methodological point that Bayesian models can guide research even when it comes to explaining effects that are on their face anti-Bayesian. In contrast, I cite it to make the metaphysical point that we can regard such anti-Bayesian effects as evidence for how beliefs are realized at lower levels while still taking Bayesian computational models to functionally define beliefs purely in terms of rational causal powers.

Discovering that human categorization is subject to this order effect may be especially psychologically revealing, it may show that human cognition makes use of a particle filter algorithm, for example. But as revealing as it may be, the irrational power is not essential to our categorization credences, it is not part of the higher-level real pattern. A being could have the same type of credence state that humans have, and be just as well described by Anderson’s Rational Model as humans are, while making use of a different algorithm that does not give rise to the same order effect. Mr. Spock could have credences about how to categorize cartoon faces, for instance, even if Vulcan cognition evolved differently and so makes use of the local MAP. After all, this is why Anderson could offer the local MAP in connection with his Rational Model.

In the schematic example, if you think of 3I as the irrational power that is manifested in the order effect, it will follow that *H*-instances that are token identical with *L*-instances, and so have the overall causal profile of {1R, 2R, 3I, 4I, 5A, 6A}, will possess this power and so demonstrate the order effect, while *H*-instances that are token identical with *L*\*-instances, and so have the overall causal profile of {1R, 2R, 38I, 45I, 51A, 64A}, will not (although they will have other irrational powers). The irrational power manifested in the order effect will flow not from the essence of belief itself, but from how certain belief tokens (not all) happen to be realized.

 In §4 I expand on this line of thought by connecting it to well-known psychological findings of human irrationality. To close out the present section, though, I want to gesture at how it might connect to an example of arationality (while acknowledging this discussion will be brief and more could be said). In an overview of Bayesian models of cognition, Griffiths and colleagues discuss arational timing effects of the sort mentioned in §2. They write that there are

certain limits on the phenomena that are valuable to study within a Bayesian paradigm. Some phenomena will surely be more satisfying to address at an algorithmic or neurocomputational [implementational] level. For example, that a certain behavior takes people an average of 450 milliseconds to produce, measured from the onset of a visual stimulus, or that this reaction time increases when the stimulus is moved to a different part of the visual field or decreases when the same information content is presented auditorily, are not facts that a rational computational theory is likely to predict (Griffiths, et al., 2008: 3).

Again, the authors make a methodological point—about which Marrian level is most “satisfying” for studying such effects—but I want to metaphysicalize the issue. On my account, timing effects are not essential to credence states that are functionally defined by a Bayesian computational model, but instead are the result of how such states happen to be realized at the algorithmic or implementational levels. This allows that different lower-level realizations of Bayesian credence states might give rise to different timing effects. In accordance with the claim made by **Essential Powers**, such arational powers simply are not essential to credence state types as they are functionally defined by Bayesian computational models.

4. Implications for Cognitive Science

 A metaphysics of science that is worth the trouble of writing about should not only be informed by our best science, it should in turn seek to inform that science. It should aim to advance our understanding of science by drawing on the metaphysical accounts being developed. In this final section I pursue this aim by drawing out the implications of my view for two potential problems facing Bayesian accounts. First, some critics contend that Bayesian accounts of central cognitive processes are untenable given empirical findings of systematic human irrationality, especially including those from the *heuristics and biases* research program (Tversky & Kahneman, 1974; Kahneman et al., 1982; Kahneman & Tversky, 2000).

 Now, there is not and need not be a single, unified Bayesian response to the sundry findings of human irrationality. However, one line that has emerged follows the approach described in the previous section (Chater & Sanborn, 2016; Icard, 2018; Lieder & Griffiths, 2019). It proceeds in a top-down fashion by starting with a Bayesian computational model of some cognitive domain. It then regards findings of irrationality that deviate from Bayesian ideals not as reason to throw out the model, but as evidence for how things are realized at the algorithmic level. In that case, alternative algorithmic realizations of the same Bayesian computational model might give rise to different forms of systematic irrationality. If so, the irrational powers discovered in heuristics and biases research might be especially psychologically important for what they reveal about algorithmic realization, but they are not part of the computational-level real pattern, and so they are not essential to credence state types as they are functionally defined by Bayesian computational models.

To illustrate, consider *anchoring* (Tversky & Kahneman, 1974; Bahník et al., 2017). In the best-known demonstration of the effect, Tversky and Kahneman asked subjects to estimate the percentage of African countries in the United Nations after having them observe a roulette wheel that landed on either 65 or 10. Subjects who observed a 65 spin had a median guess of 45% on the question, while those who observed a 10 spin had a median guess of 25%. To explain why the evidentially irrelevant factor of a roulette spin seemed to be influencing guesses, Tversky and Kahneman proposed an anchoring heuristic, where subjects arrive at their probability estimates by using the number given by the roulette spin as an initial focal point and then adjusting a bit from there. The effect is robust and has been proposed to play a role in negotiations, consumer behavior, and courtroom sentencing (Bahník et al., 2017).

 Lieder and colleagues contend that anchoring is consistent with a Bayesian approach (Lieder et al., 2012; Lieder et al., 2017). While embracing a Bayesian computational model, they show that a Metropolis-Hastings algorithm can account for much of the empirical data on anchoring by using a form of sampling from a posterior distribution that approximates Bayesian inference. Similar Bayesian treatments have been advanced for other famous cognitive biases as well, including the conjunction fallacy (Sanborn & Chater, 2016; Zhu et al., 2020), preference reversals (Howes et al., 2016), risk aversion (Khaw et al., 2017), and more.

While there is no guarantee in advance that cognitive biases should be understood in these terms, attempting to do so now constitutes a burgeoning research project (Sanborn & Chater, 2016; Icard, 2018; Lieder & Griffiths, 2019), and one that fits well with the normativist view defended here. Different algorithms that approximate ideal Bayesian inference—that is, heuristics—will fall short of the ideal in different ways, giving rise to different patterns of irrationality—that is, biases. But these biases or irrational powers are not essential to belief as a type, functionally defined at the computational level, precisely because alternative algorithmic realizations will give rise to alternative biases or irrational powers.[[23]](#footnote-23) Spock could have beliefs about the percentage of African countries in the U. N. even if Vulcan cognition evolved differently so as not to use an anchoring and adjustment heuristic.

 This fits well with how Kahneman and Tversky themselves often present their view (Kahneman & Tversky, 1996). Kahneman was asked in an interview whether artificial intelligences should be expected to demonstrate perfect rationality, or the same cognitive biases humans have (like anchoring, the conjunction fallacy, etc.), or a kind of “emergent weirdness” in the form of idiosyncratic cognitive biases of their own, reflecting their own distinctive underlying architecture. Kahneman’s response:

Emergent weirdness is a good bet. Only deduction is certain. Whenever an inductive short-cut is applied, you can search for cases in which it will fail… By their very nature, heuristic shortcuts will produce biases, and that is true for both humans and artificial intelligences, but the heuristics of AI are not necessarily the human ones (Freakonomics, 2011).

 There is often said to be a “Great Rationality Debate” in cognitive science (Stein, 1996), a largescale disagreement over just how irrational human beings are, with Bayesians counted among the “Panglossians” who downplay the supposed failures of rationality (Stanovich, 2012). But the normativist position defended here does not need to be especially Panglossian. What matters to the normativist view is not *how much* human irrationality there is—maybe the average person commits the conjunction fallacy 17 times a day—but whether the irrational causal powers that our beliefs plainly do possess are *essential* to belief as a type. Do they form part of the higher-level real pattern that our computational level theories in cognitive science should be picking up on? No, says the defender of Bayesian computational models. No, says the normativist.

 To adopt a metaphor from Dennett (1981), imagine a vastly superior intelligence—from Mars, let us say, following Dennett—with comprehensive knowledge of the microphysics of the world, but blind to normativity. Such a being might be able to predict with complete accuracy how the physical world will unfold throughout the rest of time, but it will miss various higher-level real patterns if normativism is right. Given the causal profile of {1R, 2R, 3I, 4I, 5A, 6A}, the Martian will be unable to recognize what unites the subset {1R, 2R}, or what distinguishes it from other subsets like {1R, 4I, 6A} or {3I, 5A,}. By missing out on rational normativity, the Martian misses what is shared across different possible algorithmic implementations of the Rational Model of Categorization, or for that matter what is shared across Bayesian accounts of categorization, memory, causal learning, language processing, and so on for various cognitive processes. It is often said to be a virtue of the Bayesian approach how it *unifies* cognitive phenomena (Griffiths, et al. 2010; Tenenbaum, et al. 2011; Hartmann & Colombo, 2017). The normativist can add that it is precisely rational normativity that provides the unification, as illustrated by the Martian metaphor.

 I have been arguing that my Bayesian-inspired version of normativism is consistent with findings of irrationality familiar from the heuristics and biases program. It is not, however, consistent with all views come what may—it is not unfalsifiable. Mandelbaum (2019) criticizes Bayesian approaches partly by pointing to the belief disconfirmation effect, a form of belief polarization in which subjects *increase* their belief in a given proposition in response to evidence against it. Mandelbaum proposes that the effect should be understood not as a kind of cognitive malfunction, or as the result an algorithm that approximates Bayesian inference but falls a bit short, but rather in terms of the proper functioning of a kind of *psychological immune system* (Gilbert, 2006). On this view, people identify with certain beliefs that they hold especially deeply, including for instance religious or political beliefs. This causes a kind of psychological discomfort upon receiving disconfirming evidence. To fend off the threat to one’s sense of self, the psychological immune system responds to such disconfirming evidence by reaffirming the belief with greater strength than before, adjusting credences in the opposite direction of what Bayesian norms of rationality require.[[24]](#footnote-24)

 For my purposes here, I will not try to argue against Mandelbaum’s view—although see for instance Jern, et al. (2014) for a rational analysis of belief polarization. Rather, I mention the position to help clarify how the debate over normativism might proceed going forward. There are alternatives to normativism available, alternatives where what makes up the higher-level real pattern, what is essential to belief, are not rational powers per se, but causal powers involved in protecting one’s self-conception regardless of whether they are rational, irrational, or arational. Insofar as the belief disconfirmation effect gives us reason to accept Mandelbaum’s psychological immune system, it gives us reason to reject normativism as I have been defending it. Going forward, it is challenges like this that normativists will need to fend off.

 I turn now to a second problem facing Bayesians, one pressed especially by Danks (2008, 2018). Rational Analysis purports to explain *why* human cognition works as it does: it is because the way the mind operates is optimal. However, in order for such an explanation to be correct, optimality or rationality need to play a genuine causal role. It needs to be the case that evolution or learning exerted a kind of pressure on human cognition, pushing us to use Bayesian updating in categorization, for example, and doing so *because* it is rationally optimal. In the absence of such a causal claim, Rational Analysis explanations would be false or at least deficient. And, Danks contends, Rational Analysis proponents typically provide no sustained argument for this claim that they need.

 In response, I say that the metaphysical framework set out in this paper can be used to develop the sort of argument that Rational Analysis proponents need. To illustrate how it would go, suppose that human categorization in fact does use a particle filter. There is then a question of *why* a particle filter—why not some other type of algorithm, why not a non-Bayesian one? Adopting the metaphysical framework I have defended here, Rational Analysis proponents should say that it is because of the rational causal powers the particle filters possess—the filter was selected by evolution or learning because of these rational powers. And in connection, they should add, it is *not* because of whatever irrational or arational powers the particle filter possesses. Human categorization uses a particle filter *despite* the irrational order effect it gives rise to, not *because* of that order effect.

 Now, if categorization were the only domain that researchers could draw on, that would pose a challenge for establishing this causal claim. What would be the basis for privileging rational powers over the irrational and arational ones, for regarding just the rational powers as the explanation for why this particle filter is used? But suppose instead that human cognition approaches optimality across many different domains, just as Bayesians typically hold: not just in categorization but also in causal learning and reasoning, language processing, intuitive physics, and so on. And suppose in addition that different algorithms are used in these different domains, giving rise to different idiosyncratic patterns of irrationality and arationality. As a result, the specific order effect observed in categorization is not found in these other domains.

 This could justify privileging the rational powers. If in other domains, human cognition uses distinct algorithms that approximate ideal Bayesian inference but without giving rise to the same order effect, that would seem to suggest that it is not because of the order effect that human categorization uses the particle filter that it does, but because of its ability to approximate ideal Bayesian inference. In the scenario imagined, it would be as if Bayesian updating is a kind of magnet that in domain after domain keeps drawing human cognition toward optimality, while relying on different heuristics with different biases to achieve this result in different domains.

In that event, the explanation for why human cognition is so Bayesian should appeal to something that the various algorithms have in common, with their capacity to approach rational optimality being the natural candidate. The explanation should not appeal to something that differentiates the various algorithms, like the idiosyncratic biases the algorithms give rise to. Now, to be sure, Bayesian work on algorithmic realization is still in its early stages, and so I don’t want to overstate the present empirical strength of the argument being envisioned. However, I do think this at least has the makings of a promising response to Danks’ objection,.

To think through the issue further, consider how concerns about Bayesian vacuity might arise here (Bowers & Davis, 2012; Glymour, 2011).[[25]](#footnote-25) According to such concerns, any pattern of behavior will fit some Bayesian model, and so the result that domain after domain can be made to fit with some Bayesian model or another tells us nothing substantive about the world. But in that case, this sort of convergence in which domain after domain can be given a Bayesian treatment would be useless for addressing Danks’ objection—the objection requires something substantive in response.

In response to this sort of concern, however, we can imagine a more substantive convergence across models, so that they agree not just in being broadly Bayesian, but in the more detailed causal structures they describe—the set of causal powers they attribute. So for instance, a Bayesian model of causal learning and reasoning might posit certain rational powers to account for the phenomenon of “backward blocking” (Sobel et al., 2004), which would then impose constraints on the rational powers that can be used to account for other elements of causal learning and reasoning. Or the given Bayesian model of causal learning and reasoning might impose constraints on the causal powers posited by models of categorization, given that categorization depends in part on representations of causal structure (cf. Danks, 2014: 30).

 To stick with this last example, if the set of rational powers posited by a given Bayesian model of causal learning and reasoning were to match up with the set of rational powers posited by the Rational Model of Categorization, that would support a kind of convergent realist argument in favor of regarding those particular rational powers as real elements of the world: they keep showing up in empirically successful models, so they must be real. And if such rational powers are real, they can play a genuine explanatory role and figure in the response to Danks’ objection. Now, it is an empirical question whether the causal powers described by different models will match up in this way. But in that case, it is not a vacuous claim we are entertaining—it is something that could empirically turn out to be false.

 As this discussion highlights, the metaphysical view I have defended in this work is hostage to empirical fortune in various respects. But so it goes in the metaphysics of science, where the aim is not to construct empirically impenetrable fortresses of a priori speculation, but to use the tools of metaphysics to clarify our leading scientific views and work out their implications. Here I have tried to bring normativism and the subset model of realization into contact with Bayesian cognitive science, thereby furthering our understanding.

WORKS CITED

Anderson J. R. (1990). *The Adaptive Character of Thought*. Psychology Press.

Anderson, J. R. (1991). “The Adaptive Nature of Human Categorization,” *Psychological*

*Review*, 98.3: 409-429.

Anderson, J. R. & Milson, R. (1989). “Human Memory: An Adaptive Perspective,”

*Psychological Review*, 96.4: 703-719.

Bahník, S., Englich, B., and Strack, F. (2017). “Anchoring Effect,” in Pohl, R. F. (ed.)

*Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment, and Memory, 2nd*

*Edition*. Routledge: 223-24o.

Block, N. (1980). “Troubles with Functionalism,” in *Readings in the Philosophy of*

*Psychology*. Harvard University Press: 268-305.

Boghossian, P. (2008). *Content and Justification: Philosophical Papers*. Oxford University

Press.

Bowers, J. & Davis, C. (2012). “Bayesian Just-So Stories in Psychology and

Neuroscience,” *Psychological Bulletin*, 138: 389-414.

Brandom, R. (1994). *Making it Explicit*. Harvard University Press.

Broome, J. (2015). “Reason Versus Ought,” *Philosophical Issues*, 25:

Chater, N. & Manning, C. D. (2006). “Probabilistic Models of Language Processing and

Acquisition,” *Trends in Cognitive Sciences*, 10.7: 335-344.

Cherniak, C. (1986). *Minimal Rationality*. MIT Press.

Danks, D. (2008). “Rational Analyses, Instrumentalism, and Implementations,” in

Chater, N. and Oaksford, M. (eds.) *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford University Press: 59-75.

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. The

MIT Press.

Danks, D. (2018). “Probabilistic Models,” in Sprevak, M. & Colombo, M. (eds.) *The*

*Routledge Handbook of the Computational Mind*. Routledge.

Davidson, D. (1970). “Mental Events,” in Foster, L. & Swanson, J. W. (eds.) *Experience and*

*Theory*. Duckworth. Page references to reprint in Davidson (1980).

Davidson D. (1973). “Radical Interpretation,” *Dialectica*, 27: 314-328.

Davidson, D. (1980.) *Essays on Actions and Events*. Clarendon Press.

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Clarendon Press.

Davidson, D. (1997). “Indeterminism and Antirealism,” in Kulp, C. B. (ed.)

*Realism/Antirealism and Epistemology*. Rowman & Littlefield: 109-122.

Denison, S., Bonawitz, E., Gopnik, A. & Griffiths, T. L. (2013). “Rational Variability in

Children’s Causal Inferences: The Sampling Hypothesis,” *Cognition*, 126.2: 285-300.

Dennett, D. C. (1981). “True Believers: The Intentional Strategy and Why It Works,” in

Heath, A. F. (ed.) *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of London*. Clarendon Press: 150-167.

Dennett, D. C. (1987). *The Intentional Stance*. Bradford Book.

Dennett, D. C. (1991). “Real Patterns,” *Journal of Philosophy*, 88.1: 27-51.

Dennett, D. C. (2017). “Darwin and the Overdue Demise of Essentialism,” in Smith, D. L.

(ed.) *How Biology Shapes Philosophy*. Cambridge University Press: 9-22.

Ehring, D. (2011). *Tropes: Properties, Objects, and Mental Causation*. Oxford University

Press.

Elqayam, S. & Evans, J. St. B. S. (2011). “Subtracting ‘Ought’ From ‘Is’: Descriptivism

Versus Normativism in the Study of Human Thinking,” *Behavioral and Brain Sciences*, 34: 233-290.

Elqayam, S. & Over, D. O. (2016). “From Is to Ought: The Place of Normative Models in

the Study of Human Thought,” *Frontiers in Psychology*, 7: Article 628.

Endicott, R. (2012). “Resolving Arguments by Different Conceptual Traditions of

Realization,” *Philosophical Studies*, 159.1: 41-59.

Eriksson, L. & Hájek, A. (2007). “What Are Degrees of Belief?”, *Studia Logica*, 86: 183-213.

Fine, K. (1994). “Essence and Modality,” *Philosophical Perspectives* 8: 1-16.

Fodor, J. (1974). “Special Sciences,” *Synthese*, 28.2: 97-115.

Freakonomics. (2011). “Daniel Kahneman Answers Your Questions,” retrieved from

<https://freakonomics.com/2011/11/28/daniel-kahneman-answers-your-

questions/>

Gerstenberg, T. & Tenenbaum, J. B. (2017). “Intuitive Theories,” in Waldman, M. R. (ed.)

*The Oxford Handbook of Causal Reasoning*. Oxford University Press: 515-547.

Gilbert, D. (2006). *Stumbling on Happiness*. Alfred A. Knopf.

Glymour, C. (1980). *Theory and Evidence*. Princeton University Press.

Glymour, C. (2011). “Osiander’s Psychology,” *Behavioral and Brain Sciences*, 34: 199-200.

Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. (2008). “Bayesian Models of Cognition,” in

 Sun, R. (ed.) *The Cambridge Handbook of Computational Psychology*. Cambridge University Press: 59-100.

Holyoak, K. J. & Cheng, P. W. (2011). “Causal Learning and Inference as a Rational

Process: The New Synthesis,” *Annual Review of Psychology*, 62: 135-163.

Howes, A., Warren, P. A., Farmer, G., El-Deredy, W., & Lewis, R. L. (2016). “Why

Contextual Preference Reversals Maximize Expected Value,” *Psychological Review*, 123.4: 368-391.

Icard, T. (2016). “Subjective Probability as Sampling Propensity,” *Review of Philosophy*

*and Psychology*, 7.4: 863-903.

Icard, T. (2018). “Bayes, Bounds, and Rational Analysis,” *Philosophy of Science*, 85.1: 79-

101.

Jern, A., Chang, K. M., & Kemp, C. (2015). “Belief Polarization is not Always Irrational,”

*Psychological Review*, 121.2: 206-224.

Kahneman, D., Slovic, P., and Tversky, A. eds. (1982). *Judgment Under Uncertainty:*

*Heuristics and Biases*. Cambridge University Press.

Kahneman, D. & Tversky A. (1996). “On the Reality of Cognitive Illusions,” *Psychological*

*Review*, 103.3: 582-591.

Kahneman, D. & Tversky, A. (eds.). (2000). *Choices, Values, and Frames*. Cambridge

University Press.

Khaw, M. W., Li, Z., & Woodford, M. (2017). “Risk Aversion as a Perceptual Bias,” *NBER*

*Working Paper No. 23294*. National Bureau of Economic Research.

Kim, J. (1992). “Multiple Realization and the Metaphysics of Reduction,” *Philosophy and*

*Phenomenological Research*, 52.1: 1-26.

Kolodny, N. (2005). “Why Be Rational?”, *Mind* 114: 509-563.

Knill, D. & Richards, W. (eds.). (1996). *Perception as Bayesian Inference*. Cambridge

University Press.

Ladyman, J. and Ross, D., with Spurrett, D. and Collier, J. (2007). *Everything Must Go:*

*Metaphysics Naturalised*. Oxford University Press.

Lewis, D. K. (1970). “How to Define Theoretical Terms,” *Journal of Philosophy*, 67: 427-446.

Levy, R., Reali, F. & Griffiths, T. L. (2009). “Modeling the Effects of Memory on Human Online Sentence Processing with Particle Filters,” in Koller, D., Schurmans, D., Bengio, Y., & Bottou, L. (eds.), *Advances in Neural Information Processing Systems*, vol. 2. NIPS Foundation: 937-944.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). “Burn-in, Bias, and the Rationality

of Anchoring,” in Bartlett, P., Pereira, F. C. N., Bottou, L., Burgess, C. J. C., & Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems* 26.

Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2017). “The Anchoring Bias

Reflects Rational Use of Cognitive Resources,” *Psychonomic Bulletin & Review*, 25: 322-349.

Lieder, F. & Griffiths, T. L. (2019). “Resource-Rational Analysis: Understanding Human

Cognition as the Optimal Use of Limited Cognitive Resources,” *Behavioral and Brain Sciences*, 4.43: 1-16.

McHugh, C. & Whiting, D. (2014). “The Normativity of Belief,” *Analysis*, 74.4: 698-713.

Maloney, L. T. & Mammasian, P. (2009). “Bayesian Decision Theory as a Model of

Human Visual Perception,” *Visual Neuroscience*, 26.1: 147-155.

Mandelbaum, E. (2019). “Troubles with Bayesianism: An Introduction to the

Psychological Immune System,” *Mind and Language*, 34.2: 141-157.

Mandelbaum, E., Won, I., Gross, S., & Firestone, C. (2019). “Can Resources Save

Rationality? ‘Anti-Bayesian’ Updating in Cognition and Perception,” *Brain and Behavioral Sciences*, 4.43: 31-32.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and*

*Processing of Visual Information*. MIT Press.

Medin, D. L. & Schaffer, M. M. (1978). “Context Theory of Classification Learning.”

*Psychological Review*, 85.3: 207-238.

Medin, D. L. & Bettger, J. G. (1994). “Presentation Order and Recognition of

Categorically Related Examples,” *Psychonomic Bulletin & Review*, 1: 250-254.

Melnyk, A. (2003). *A Physicalist Manifesto: Thoroughly Modern Materialism*. Cambridge:

Cambridge University Press.

Oaksford, M. (2014). “Normativity, Interpretation, and Bayesian Models.” *Frontiers in*

*Psychology,* 5, Article 332.

Oaksford, M. & Chater, N. (1999). “Ten Years of the Rational Analysis of Cognition,”

*Trends in Cognitive Science*, 2: 57-65.

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to*

*Human Reasoning*. Oxford University Press.

Polger, T. W. & Shapiro, L. A. (2016). *The Multiple Realization Book*. Oxford University

Press.

Putnam, H. (1967). “Psychological Predicates,” in Capitan, W. H. & Merrill, D. D. (eds.),

*Art, Mind, and Religion*. University of Pittsburgh Press: 37-48.

Reed, S. K. (1972). “Pattern Recognition and Categorization.” *Cognitive Psychology*, 3: 382-

407.

Rescorla, M. (2013). “Rationality as a Constitutive Ideal,” in Lepore, E. & Ludwig, K.

(eds.), *A Companion to Davidson*. Wiley-Blackwell: 472-488.

Rescorla, M. (2015). “Bayesian Perceptual Psychology,” in Matthen, M. (ed.), *The Oxford*

*Handbook of the Philosophy of Perception*,. Oxford University Press: 694-716.

Rescorla, M. (2019). “A Realist Perspective on Bayesian Cognitive Science,” in Chan, T. &

Nes, A. (eds.), *Inference and Consciousness*. Routledge: 40-74.

Rey, G. (2007). “Resisting Normativism in Psychology,” in McLaughlin, B. P. & Cohen, J.

(eds.) *Contemporary Debates in Philosophy of Mind*. Blackwell: 69-84.

Rips, L. J. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking*. A

Bradford Book.

Rosen, G. (2015). “Real Definition,” *Analytic Philosophy*, 56.3: 189-209.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). “Rational Approximations to

Rational Models: Alternative Algorithms for Category Learning,” *Psychological Review*, 117.4: 1144-1167.

Sanborn, A. N., Mansinghka, V. M., & Griffiths, T. L. (2013). “Reconciling Intuitive

Physics and Newtonian Mechanics for Colliding Objects.” *Psychological Review*, 120.2: 411-437.

Sanborn, A. N. & Chater, N. (2016). “Bayesian Brains without Probabilities,” *Trends in*

*Cognitive Sciences*, 20.12: 883-893.

Shiffrin, R. M. & Steyvers, M. (1997). “A Model for Recognition Memory: REM—

Retrieving Effectively from Memory,” *Psychonomic Bulletin & Review*, 4: 145-166.

Sellars, W. (1956). “Empiricism and the Philosophy of Mind,” in *Science, Perception and*

*Reality*, 1963 and 1991. Ridgeview: 127-196.

Shoemaker, S. (1980). “Causality and Properties,” in Van Inwagen, P. (ed.) *Time and*

*Cause: Essays Presented to Richard Taylor*. Reidel Publishing: 109-135.

Shoemaker, S. (2001). “Realization and Mental Causation,” in Gillett, C. & Loewer, B.

(eds.) *Physicalism and its Discontents*. Cambridge University Press.

Shoemaker, S. (2007). *Physical Realization*. Oxford University Press.

Sider, T. (2020). *The Tools of Metaphysics and the Metaphysics of Science*. Oxford University

Press.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). “Children’s Causal Inferences from

Indirect Evidence: Backwards Blocking and Bayesian Reasoning in Preschoolers,” *Cognitive Science*, 28: 303-333.

Stanovich, K. E. (2012). “On the Distinction Between Rationality and Intelligence:

Implications for Understanding Individual Differences in Reasoning,” in Holyoak, K. J. & Morrison, R. G. (eds.), *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press.

Stanovich, K. E. (2013). “Why Humans Are (Sometimes) Less Rational than Other

Animals: Cognitive Complexity and the Axioms of Rational Choice,” *Thinking and Reasoning*, 19.1: 1-26.

Stein, E. (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive*

*Science*. Clarendon Press.

Sutherland, S. (2013.) *Irrationality: The Enemy Within*. Pinter & Martin.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). “Bayesian Models of

Cognition Revisited: Setting Optimality Aside and Letting Data Drive Psychological Theory.” *Psychological Review*, 124.4: 410-441.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). “How to Grow a

Mind: Statistics, Structure, and Abstraction.” *Science*, 11.331: 1279-1285.

Tversky, A. and Kahneman, D. (1974). “Judgment under Uncertainty: Heuristics and

Biases,” *Science*, 185: 1124-1131.

Van Rooij, I. (2008). “The Tractable Cognition Thesis,” *Cognitive Science*, 32.6: 939-984.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2009.) “One and Done?

Optimal Decisions from Very Few Samples,” in Taatgen, N. & van Rijn, H. (eds.) *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Cognitive Science Society: 148-153.

Wedgwood, R. (2002). “The Aim of Belief,” *Philosophical Perspectives*, 16: 267-297.

Wedgwood, R. (2007a). *The Nature of Normativity*. Oxford University Press.

Wedgwood, R. (2007b). “Normativism Defended,” in McLaughlin, B. P. & Cohen, J.

(eds.) *Contemporary Debates in Philosophy of Mind*. Blackwell: 85-101.

Wedgwood, R. (2017). *The Value of Rationality*. Oxford University Press.

Wilson, J. (1999). “How Superduper Does a Physicalist Supervenience Need to

 Be?” *Philosophical Quarterly*, 49: 33-52.

Wilson, J. (2011). “Non-reductive Realization and the Powers-Based Subset Strategy,”

*The Monist*, 94.1: 121-154.

Xu, F. & Tenenbaum, J. B. (2007). “Word Learning as Bayesian Inference.” *Psychological*

*Review*, 114.2: 245-272.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). “The Bayesian Sampler: Generic

Bayesian Inference Causes Incoherence in Human Probability Judgments,” *Psychological Review*.

1. Psychologists use the term “normativism” for the methodological thesis that normative accounts should guide psychological theorizing, a thesis that has come in for critical scrutiny (Elqayam & Evans, 2011; Elqayam & Over, 2016; Tauber et al., 2017). This is different from how philosophers like myself use the term, for the metaphysical thesis that belief has a normative essence. However, there is potentially a connection if the view defended in this paper is correct: metaphysical normativism might be part of the best explanation for why methodological normativism is empirically successful. [↑](#footnote-ref-1)
2. In addition to predicting such empirical findings, the Rational Model also promises to offer new theoretical insights into categorization. For instance, Sanborn et al. (2010) note that it is possible to reconceive of the dispute between exemplar theory and prototype theory, regarding how categories are represented, as a dispute between alternative versions of the Rational Model, with exemplar theory in effect making use of nonparametric density estimation while prototype theory uses parametric methods. This reconceptualization helps clarify just where the models differ in their assumptions and also helps unify them by highlighting what they have in common. It also points to room in logical space for approaches that fall between the two or make use of both. [↑](#footnote-ref-2)
3. Wedgwood (2002 and 2017) takes the fundamental norm governing belief to be a norm of truth, and then takes the norms of rationality to hold derivatively. This view is compatible with my position. [↑](#footnote-ref-3)
4. There is also a debate, sparked by Kolodny (2005), over whether rationality is genuinely normative. I will assume without argument that it is, but see Wedgwood (2017) for a response to Kolodny that fits comfortably with the normativist view defended here. [↑](#footnote-ref-4)
5. Wedgwood (2007a and 2007b) develops his framework in terms of rational *dispositions* rather than powers, but I take this to be just a verbal difference. I prefer the term “powers” in order to set up the discussion of the causal powers subset account of multiple realization in §3. [↑](#footnote-ref-5)
6. Harman (1995: 184) suggests that the constraints of theoretical and practical rationality can conflict. For instance, it may be practically rational but theoretically irrational to engage in positive thinking. If I were forced to allow for this, I could revise the account in the text and say that rational powers are those whose exercise complies with some norm of rationality without violating any, while irrational powers are those whose exercise violates some norm. [↑](#footnote-ref-6)
7. It is worth noting that cognitive scientists sometimes apply the framework of Bayesian decision theory outside the domain of decisions narrowly conceived. For instance, Bayesian accounts of vision often take the visual system to have a utility function that determines the percept experienced by a subject (Maloney & Mamassian, 2009; Rescorla, 2015). The upshot is that the appeal to Bayesian decision theory will have broader application than might be initially obvious. [↑](#footnote-ref-7)
8. Stanovich (2013) reviews empirical research showing that nonhuman animals like pigeons, rats, and chimps behave in ways that comply with various norms of decision theory better than human beings do. Now, there may be important arguments to be made that such nonhuman animals lack beliefs, but that they are insufficiently irrational to believe anything is not one. [↑](#footnote-ref-8)
9. See n. 14 on how to make this allowance for exceptions more rigorous. [↑](#footnote-ref-9)
10. In their review of recent work, McHugh & Whiting (2014) write that “it is surprisingly rare to find arguments for Normativism,” despite how many proponents of the view there are. None of the arguments they do review are empirical. [↑](#footnote-ref-10)
11. In principle, it would be possible to develop an empirical defense of normativism that appealed to non-Bayesian (but still rational) approaches. However, my argument in this section appeals especially to work by Bayesians on how Bayesian updating might be algorithmically realized in human beings. Non-Bayesian views that have done less to address the issue of realization won’t be able to advance this sort of argument. Thanks to an anonymous referee for pressing me on the point. [↑](#footnote-ref-11)
12. Shoemaker (2007: 18) also uses the Ramsey-Lewis method within his version of the subset account. [↑](#footnote-ref-12)
13. Icard (2016) defends a view of credences that is in the vicinity of my position, focusing on sampling algorithms that approximate Bayesian inference. [↑](#footnote-ref-13)
14. One complication that can be passed over for some purposes but is worth acknowledging here is that we might want to loosen functional definitions a bit, so that what is required for being in *H* is not possessing *all* of the causal powers mentioned in the functional definition but rather *sufficiently many of them*—see Lewis’s (1970: 432) account of *near-realization*, which requires that the disjunction of the conjunction of the clauses of the Ramsey sentence of the causal theory be true. I mention this here because we will want to allow that a system can have a credence state functionally defined by a Bayesian computational model even if the system deviates a bit from the model, even if it is not quite as ideally rational as the model describes. [↑](#footnote-ref-14)
15. As Endicott (2012) observes, different philosophers understand the realization in subtly different ways, including that they conceive of it as having different relata. On my view, the realization relation obtains in the first place between types, but then in a derivative sense we can speak of the realization of tokens, or of theories, or of models. [↑](#footnote-ref-15)
16. Or just in case *L* possesses sufficiently many of such powers, given near-realization; see n. 14. [↑](#footnote-ref-16)
17. This talk of “inheritance” is taken from Kim’s (1992) *causal inheritance principle*. See (omitted). [↑](#footnote-ref-17)
18. This is not the only way to motivate the claim—see (omitted) for further arguments. [↑](#footnote-ref-18)
19. Both Shoemaker (2007) and Wilson (2011) reject the token identity theory in their development of the subset account of realization, and so the discussion here marks a break from their views. Other defenders of the subset account embrace a token identity theory, however, including Ehring (2011) and (omitted). [↑](#footnote-ref-19)
20. Dennett (2017) argues against essences, a view that other naturalistic philosophers share. Part of my response to this position, as suggested in the text, is that the appeal to essences helps us make sense of real patterns, allowing us to distinguish between what comprises the pattern (what is essential) and what does not (what is inessential). In a spirit of concession, perhaps I can allow that this calls for a metaphysically thinner conception of essences than what Dennett is rejecting. [↑](#footnote-ref-20)
21. We can still say that the Rational Model functionally defines credence states about categorization even if it attributes to subjects insensitivity to stimuli order if we appeal here to near-realization; again see n. 16. [↑](#footnote-ref-21)
22. To provide an additional example of the approach: Levy et al. (2009) observe that a Bayesian computational model of sentence-processing captures a wide range of psycholinguistic findings but fails to predict human struggles with “garden path sentences.” In response, they descend from the computational to the algorithmic level and show that a particle filter that approximates ideal rational inference is able to capture those findings correctly predicted by the Bayesian computational model, while in addition predicting human troubles with garden-path sentences. [↑](#footnote-ref-22)
23. Ideal Bayesian computations are often intractable (van Rooij, 2008), and so the best that can be achieved are algorithms that approximate the ideal but fall short in certain ways. Some critics take this intractability to show that the Bayesian models need to be understood in instrumentalist rather than realist terms, but see Rescorla (2019) for a defense of Bayesian realism. [↑](#footnote-ref-23)
24. Mandelbaum (2019) acknowledges that the findings of irrationality familiar from the heuristics and biases literature are consistent with the Bayesian approach, as I have been arguing in this section. He turns to the psychological immune system to try to advance the discussion beyond this point. [↑](#footnote-ref-24)
25. Thanks to an anonymous referee for pressing me on this point. [↑](#footnote-ref-25)