Troubles with Bayesianism: An Introduction to the Psychological Immune System

1. Bayes: Local and Imperial

Bayesianism, in one form or another, has never been more popular than it is now. Its use in normative inquiries (e.g., in formal epistemology) has been prominent for some time. But recently theorists have tried to extend Bayesianism to a series of descriptive endeavors. A decade or so ago only a handful of cognitive scientists attempted to explain mental processing by Bayesian lights. Now anywhere one looks one can see philosophers and cognitive scientists alluding to the Bayes' rule in order to explain some phenomenon or another.

Bayesianism's appeal isn't hard to see: it allows for the possibility of a single mental mechanism—Bayesian updating—to unify mental processes as diverse as word learning (Xu and Tenenbaum 2007), belief updating (Bennett 2015), conditional reasoning (Oaksford and Chater 1994) the development of moral judgments (Nichols et al. forthcoming), domain general reasoning (Vul and Pashler 2008), reward prediction error learning (Hohwy 2013, Clark 2013), compositionality in the Language of Thought (Goodman et al. 2015a), causal reasoning (Gweon and Schulz 2011) and reinforcement theory learning (Vlassis 2012) to name just a few recent domains of interesting work falling under the Bayesian banner. The sheer generality of Bayesianism allows a scope unmatched by most theories, save for discredited ones like Radical Behaviorism (Skinner 1974) and Associationism (Mandelbaum 2016).

Moreover, if one tries to reverse engineer the mind Bayesianism has few competitors (Tenenbaum et al. 2011). Though there are other computational models one can use that aren't necessarily Bayesian, the relative success of Bayesian models in engineering and machine learning should bolster one's confidence in Bayesianism.[1]

But what exactly is descriptive Bayesianism (hereafter referred to simply as 'Bayesianism')? Since there is no simple idea that separates out Bayesians from non-Bayesians, it will take a bit of work to detail the contours of the theory. We can start by separating Bayesians into two camps. Call the first 'Imperial Bayesians.' Imperial Bayesians think that the Bayes' rule is, in some way or another, approximated by all mental processes. For Imperialists it is not an accident that Bayesian analyses arise in a diverse set of findings since they believe that all mental processing—perception and cognition—aims at approximating a Bayesian ideal. In contrast to Imperial Bayesians there are what I'll call 'Local Bayesians.' Local Bayesians differ from Imperial Bayesians merely in the scope of their theories: whereas Imperialists think all mental processing is Bayesian, Localists think that only some mental processes are Bayesian (and may be agnostic on the global question). That is, Localists can still posit a heterogeneous array of mental mechanisms, of which Bayesian processing is just one. Arguing that all forms of Local Bayesianism are false root and branch would thus be too large a task for a single essay: to argue that no mental process is Bayesian, one would have to go through each mental process one by one showing that its processing cannot be interpreted in a Bayesian fashion. For the rest of the essay I will remain neutral on the question of the truth of Local Bayes. Instead, my focus will be on Imperial Bayes. Although no theorist may hold the exact Imperial Bayes position, ones extremely close to it are widely held (e.g., Chater and Oaksford 2009, Tenenbaum et al. 2011, Friston 2012, Hohwy 2013, Clark 2013).[2] Moreover, many of the criticisms of Imperial Bayes will

[1] See, e.g., Schmidhuber (2015) for an overview of Deep Learning models, which needn't be Bayesian.
[2] One may be inclined to separate out the Methodological Imperialists from the Radical (or 'fundamentalist,' Jones and Love 2011) Imperialists, just as one might have for Behaviorism. As far as I can tell, Clark and

apply to specific applications of Local Bayes. Since the bulk of my critique will focus on specific examples, this critique can also be read as a criticism as some specific forms of Local Bayes.

2.The Algorithmic and Computational: Optimality and Bayesianism

Marr (1982) famously outlined three levels of explanatory desiderata for mental processes: the computational, the algorithmic, and the implementational. The computational level describes the problem the system is trying to solve. The algorithmic level describes the actual algorithms the process utilizes to solve the problem specified at the computational level. Finally, the implementational level describes how the algorithms are physically implemented.

Part of the value of the computational level is purported to be that specifying the problem that the system is trying to solve should help constrain the types of solutions the system might use—that is, the computational level goals should constrain our search for algorithmic level models (Oaksford and Chater 2009). There has been confusion as to whether Bayesianism is meant to apply to the computational or algorithmic level (Jones and Love 2011, Oaksford and Chater 2007). Computational Bayesians claim that Bayesian analysis shows how a system would solve a problem, assuming it were to solve it optimally. Computational Bayesians are thus another variety of normative Bayesians, and are technically agnostic as to how actual mental processing unfolds. In contrast, algorithmic Bayesians make claims about how actual mental processing works. Algorithmic Bayesians are committed to the idea that the optimal, rational way for a mental mechanism to solve a given problem is the actual way the mental mechanism solves the problem. It is the algorithmic conception of Bayesianism that is committed to descriptive answers for how we process and, consequently,

Hohwy are more akin to the radical end of the spectrum, whereas Tenenbaum and his collaborators are inclined towards the methodological position.

algorithmic (and not merely computational) Bayes is the position that is the aim of my critique.[3]

Abstractly, my strategy is as follows: find some process (p) for which theorists claim p operates in an optimal Bayesian way when solving a task (t). If it can be shown that *p* does not so operate in an optimal way, then we can conclude that a) Local Bayesianism is false with regard to *p* and b) Imperial Bayesianism is false *tout court*.

Of course, there are hurdles to meet along the way. For one thing, merely showing that *p* sometimes acts in a suboptimal way wouldn't itself be enough to disprove that *p* is in fact optimal for solving *t*. It might be that *p*'s suboptimality in this case isn't due to its core processing, but do to some performance constraint or other. That is, one could still hold that the core competence of *p* is optimal in regards to *t* but also believe that sometimes outside factors conspire so that *p* performs suboptimally.[4] This isn't a mere possibility, but instead a serious problem that grinds most discussions of the data to a standstill. For a concrete example, take the heuristics and biases literature, which is rife with findings about human irrationality. We are forever hearing how people ignore base rates (Kahneman and Tversky 1973), fall for the conjunction fallacy (ibid.), are deceived by the disjunction effect (Tversky and Shafir 1992), insufficiently adjust from irrelevant anchors (Epley and Gilovich 2001), affirm the consequent (Wason 1971), probability match instead of maximize (West and

[3] Danks (2013) argues that Marr's three levels are cross-cut by questions of instrumentalism(/realism) and optimality. Because of this Danks argues one cannot merely equate computational level processing with optimal processing. I agree with the general moral Danks draws, but in the specific case of Bayesian processing theorists are in fact explicitly committed to the optimality condition (see, e.g., Oaksford and Chater 1994; Weiss et al. 2002; Griffiths and Tenenbaum 2006; Norris 2006; Bogacz 2007; Feldman et al. 2009; Girshick et al. 2011). It is the commitment to optimal processing that is the hallmark of Bayesian theories—the more one loosens this commitment, the less clear it is that the theory under scrutiny is Bayesian (as opposed to say, any old theory of utilizing probabilistic updating in some fashion). For critiques of Bayesianism because of its connection to optimality see Jones and Love 2011; Elqayam and Evans 2011; Bowers and Davis 2012).

[4] Or maybe it's not that *p* processes suboptimally per se, but instead that *p* looks to be engaged but is in fact bypassed, or that *p* in fact did process optimally, but its output was overridden by a separate process, or any other way that *ceteris* may not be *paribus*.

Stanovich 2003), and on and on. Against this backdrop it strikes some as absurd that anyone could argue for optimality. But it isn't. As Bayesians (and others) stress, the heuristics and biases project was set against a backdrop of appreciation of human rationality. Part of the genius of the original Kahneman and Tversky research was creating experimental situations that would reliably cause people to act irrationally. For every article showing that "people do not appear to follow the calculus of chance or the statistical theory of prediction" (Kahneman and Tversky 1973, 237), one can find an article showing people excelling at the same task. So what are we to do when we find evidence that college students from elite universities unabashedly ignore base rates (ibid.) while four-year-olds successfully incorporate base rates (Sobel et al. 2004)?

To break this deadlock, we need to do more than just find examples where people appear to be acting irrationally. What would be needed to show that local Bayesianism is false is to find actions that are not just the result of errors in processing. Rather, the irrationality has to result from a system that is set up to properly output the actions we categorize as irrational.

But perhaps even irrational *outputs* won't be enough in themselves to truly worry Bayesians, for paradigmatic outputs—decisions, motor behaviors, and the like—are interaction effects. What would be truly worrisome is if we found a process that *updated* in a decidedly non-Bayesian fashion. We must find suboptimal processing that is, from the standpoint of the processor, its proper functioning. To put it in our earlier terms, what would be maximally worrisome for the Bayesian would be to show that the core competence of process p in solving t is, in fact, nonoptimal from the Bayesian's own sense of optimality.

Because of the vast differences in different tasks, having a sense of what optimality is across the board is difficult. Nevertheless, Bayesians do provide us with one fixed point which can be utilized regardless of domain: Bayes' rule. Bayes' rule is purported to be the core of optimal processing itself—it's where the normative meets the descriptive. So, if we can find some *p* whose core processing itself contravenes the Bayes' rule, then we can assured that Local Bayes is false for *p* and that Imperial Bayes is false.

But what *p* should one choose to investigate? In order to make the case against Bayes as difficult, and thus compelling, as possible it's best to use a domain at which Bayesianism is most at home: belief updating. After all, Bayes' rule is most easily understood and discussed as a way of updating one's (or a process's) beliefs (/credences)[5] about a given hypothesis. As such, almost any account of Bayesian processing appears to be a version of Bayesian belief updating. Thus if we can show that belief updating itself is, at its core, deeply non-optimal in a way that contravenes Bayes' rule, we can cast skepticism on the broader Bayesian enterprise.

3. Problems for the Bayesian

In this section, I'll canvass some of the problems for Bayesianism. Because of space constraints I'll leave out many issues that are either not as dire as the ones I discuss, or that have been discussed elsewhere.[6]

3.1 Psychological Reality

Bayesians are, in some sense, committed to the idea that we update our beliefs via Bayes' rule. But I can find no theorist who actually thinks that humans update by using an explicit representation of the Bayes' rule.[7] For one thing, although updating via Bayes' rule may be

---

[5] The question of whether it is beliefs or credences that are updated is orthogonal to my focus, and I wish to remain neutral on it for the present discussion. For readability, I will refer to 'beliefs' but readers should feel free to substitute 'credences' as they see fit.

[6] For critiques regarding overfitting see Endress (2013); for worries about variability in decision rules and ad hoc model selection see Marcus and Davis (2013); for problems with probability matching see Eberhardt and Danks (2011). For some reasonable responses from prominent Bayesians see Frank (2013); Goodman et al. (2015b).

[7] For discussion of how explicit and implicit representations differ see Quilty-Dunn and Mandelbaum (MS).

possible in some very circumscribed experimental settings, it would be intractable to do so in real life reasoning. One couldn't have a fully delineated and explicit hypothesis space that one updates every time new data is received (which, on some reasonable readings of new data, is each new instant)[8]. A psychologically literal Bayesian model would also force cognizers to search through all of the posterior distribution in real time, which would be seemingly impossible—the combinatorial explosion would be too immense.[9] Thus, there is a search among Bayesians to find algorithms that approximate Bayesian inference (Vul et al. 2014). For example, some have posited that knowledge representations take the form of probabilistic distributions, and that Bayes' rule is approximated in part via sampling from such distributions (ibid.).[10] In fact, it's recently been argued that mere sampling from the posterior is almost as optimal (for decision making) as using the full posterior, even when one just takes a single sample from the posterior (and often it looks like single samples are themselves pragmatically ideal; ibid.).

Although the questions of psychological reality are important, I find them a bit less pressing than others. For one thing, they have been known for some time (see, e.g., Gigerenzer 2008); for another, figuring out the actual psychological implementation (i.e., the algorithmic-level explanation) for Bayesian reasoning is an active research program, one which many clever theorists are currently engaged in. To bemoan the project because it's *in medias res* seems shortsighted. Nevertheless, how one thinks this program will turn out will inform how optimistic one is about the long-term prospects of a Bayesian cognitive science.

---

[8] For a concrete example, see Endress (2013), which calculates that the Franks and Tenenbaum's (2011) model would demand that infants process 900 counterfactual syllable triplets (e.g., *di di je*) per second.

[9] It is, I think, worth noting that Bayesians aren't the only ones in this type of predicament. Chomsky's Minimalist Program (Chomsky 1995) appears to have similar consequences (e.g., see the extreme amount of possible sentences that are partially derived but crash before Spell-Out).

[10] 'In part' is there because there is much more to Bayesian (or any) decision making than merely sampling from a posterior—one must also use the posterior (or samples of) to make a decision of what one should do. Sampling from the posterior doesn't in and of itself dictate one's decision (or response), though it's often useful to speak as if it did.

For what it's worth, although I think Bayesianism is probably not true of how we update beliefs, I don't think its falsity is due to the impossibility of having evolved an approximate Bayesian processor.

But the problem of psychological reality puts an earlier worry into sharper focus: if we cannot rely on Bayes' rule being explicitly represented and followed, then how can we import a sense of optimality across tasks, even tasks about belief updating? If we are just approximating an optimal updater, then would deviations from the optimal really be counterexamples?

In order to get around this worry one would need to show clean evidence that no approximate Bayesian processor, no matter how its instantiated, should ever produce. Moreover, to be maximally convincing such evidence must be caused by a process whose function it is to produce such outputs. Focusing on belief updating, we have three candidates: in the first case, we fail to learn information that we should learn (a type of learning blindness), and in the second, we do not update when we should update (belief perseverance). The third and most pressing case is one of learning perversity—receiving evidence that ~P and yet increasing our belief that P.

3.2 Belief Perseverance and Not Learning What Should Be Learned

It's long been known that an organism doesn't learn everything one's learning theory predicts it should. Associationists and Behaviorists predicted that whatever properties were associated (or reinforced) in one's environment should be thereby associated in one's mind. But of course there are always more combinations of properties instantiated (/reinforced) in one's environment than are ever learned. Consider a rat in a cage that, on some pattern of reinforcement, will be shocked in conjunction with being shown a light. A deep problem for Behaviorists was to explain why, given some pattern of reinforcements, rats would learn that the light leads to the shock but given other patterns, rats would learn that the cage itself leads to the shock, ignoring the role of the light altogether (Mandelbaum 2016). Though Associationists and Behaviorists didn't have the theoretical tools to predict these patterns of learning, the Bayesians do: the rats will learn whatever stimulus is a better predictor of the shock.

The reliance on prediction allows Bayesians to explain lots of instances of failures to learn that Associationists and Behaviorists couldn't (see, e.g., Bayesian explanation of Kamin blocking, Sobel et al. 2004). But Imperial Bayesians also have problems explaining why some information that should be learned isn't. Perceptual examples abound. One can know that the figures in the Ames room or the lines of the Müller-Lyer are the same length and yet one cannot learn to see it so.

Though the failures of perceptual systems to learn, or update, some information is a problem for Imperial Bayesianism, these failures don't strike at the core of Bayesianism. For instance, one can deal with these failures by adding a bit more structure to the overall architecture of the mind. An Imperial Bayesian can posit that perceptual systems are encapsulated from the rest of the mind and perhaps such encapsulation would be enough to explain away the lack of updating in perceptual systems.[11] Moreover, perhaps the Bayesian will have to posit some innate information—such as the innate information that there is only one overhead light source. But doing so needn't affect the core of Bayesianism. After all, (non-Jeffreys) priors have to come from somewhere, and it's empiricism, not Bayesianism per se that is at odds with innate priors.

---

[11] Note that, although it is consistent for an Imperial Bayesian to believe in informational encapsulation of perceptual systems, believing in full-fledged modularity would be more or less impossible. That is because modularity entails that the different modules utilize different domain-specific algorithms (Mandelbaum 2015). It is the idea of a disparate suite of domain specific algorithms that is inconsistent with Imperial Bayes.

But the problems aren't that simple to sidestep. Similar lack of learning can be seen in cognition. Rats are prepared to learn that an audiovisual stimulus signals a shock, and they are prepared to learn that a gustatory stimulus signals nausea. Indeed, they are so 'prepared' to learn this that they need only one instance to make the induction (Garcia and Koelling 1966). But rats are contraprepared to learn that an audiovisual stimulus signals nausea or that a gustatory stimulus signals shock; that is, they cannot learn these contingencies (ibid.). Interestingly enough, humans cannot either (Baeyens et al. 1990).[12]

But again, the enlightened Imperial Bayesian can, by invoking a little architecture and nativism, explain away these presumptive counterexamples. Taste aversion learning is innate if anything is, and one can imagine priors for contingencies here being close to 1 or unmovable because of how they are otherwise stored. Some Bayesians, like Tenenbaum, welcome nativism (though others—like Clark and Hohwy—don't particularly). The more one resists nativism and other architectural constraints, the bigger these problems are. But not all problems of failures to learn involve evolutionarily significant properties (see Danks 2006). And regardless, there are central problems afoot for all Bayesians when it comes to belief perseverance for properties that aren't evolutionarily significant, problems that no amount of nativism or architecture can help solve.

Take a moment to think about the relationship between firefighters and risk preference. Do you think better firefighters are more risk averse or more risk seeking? If you are like most people studied, you a) have no antecedent opinion and b) can easily think up causal stories to explain why either case would be true. In a series of studies Anderson and colleagues examined belief perseverance about firefighting and risk preference (Anderson et

---

[12] For example, imagine becoming nauseated after drinking something that was floridly colored and had a particular aftertaste. People will not infer that it was the coloring that made them sick, only the taste; they will freely drink other substances that have the same color, but none that have the same smell or taste.

al. 1980; Anderson 1983; Anderson and Sechler 1986; Slusher and Anderson 1989). Subjects were induced to form a theory about the connection in a number of different ways, e.g., by reading fictitious case histories, or encountering fictitious data. Other subjects were merely asked to think about one type of relationship. Subjects in all the conditions were then given counterattitudinal evidence—evidence that the relationship was actually the opposite of what they had thought (e.g., that risk-aversion led to better firefighting than risk-seekingness). Whether subjects read anecdotes or perused charts, and whether they came up with their own causal link between the properties or were merely given one by the experimenters, all subjects showed the same tendency to have their beliefs persevere in the case of the counterevidence.[13] This held regardless of how the counterevidence was presented. That is, some subjects merely contemplated a hypothetical relationship between risk-seekingness and firefighting, and then were shown (fictional) mounds of data showing that in fact the relationship went the opposite way from which the subject supposed and yet the subject still wouldn't change their belief, even though the belief just arose via hypothetical contemplation.

Belief intransigence of this sort is deeply problematic for the Imperial Bayesian. After all, learning causal connections between two seemingly disparate properties is exactly the type of scenario for which Bayesian updating is tailor-made. Nevertheless, if one looks in the right way, one can find belief perseverance in many causal learning paradigms.[14] For instance, subjects in Taylor and Ahn (2012) were tasked with learning the causal connections between fictitious diseases. In the first 20 trials subjects were introduced to two fictitious

[13] Although all versions of the experiment showed subjects' tendency to persevere, some manipulations caused stronger perseverance effects than others. For example, having the subjects self-generate the relationship between the variables was more effective than giving the subjects the relationship (see, e.g., Davies 1997), and asking the subject to form some causal relationship, even if just based on a single anecdote, made for a powerful and intransigent belief (Anderson 1989).

[14] Sadly, one also finds it wherever the (in)effectiveness of debriefing is under investigation (e.g., Valins 1972; Ross et al. 1975; Wegner et al. 1985), or in studies of misinformation more generally (e.g., Ecker et al. 2011).

diseases B(urlosis) and C(aprix). Each trial was supposedly another patients chart and the patient could have B, C, both, or neither. Subjects were also told that there may be other conditions not yet listed here that may be introduced later. After 20 trials, subjects easily formed beliefs between the absence and presence of the two properties. Let's take, for instance, a subject that was in the condition where having B predicted having C (i.e., the patient would see that any patient that had C would also have B, and any patient that didn't have C didn't have B). Such a subject reliably formed the belief that B led to C. After the 20 trials, subjects were then introduced to another fictitious disease A(blique) and asked what the relationship between the three diseases were. Just as in Kamin blocking, subjects were 'blocked' here. Even given another 20 trials where A in fact lead to both B and C, subjects would persevere in their original hypothesis. Taylor and Ahn couldn't model the results using any Bayesian models, but the problem here is larger than just this one study: the moral is that paradigms where we should be seeing the most Bayesian successes—causal learning paradigms—in fact lead to failures of belief updating because of belief perseverance. The Bayesian challenge is to explain how such perseverance is consistent with Bayesianism and to predict when such perseverance will arise.

3.3 Belief Polarization

The biggest stumbling block for Bayesian theories of belief updating is a species of belief polarization. Though it's often discussed as a single phenomenon, 'belief polarization' is an umbrella term covering two distinct effects, biased assimilation-based polarization and belief disconfirmation based polarization. I take these in turn.

3.3.1 Polarization via Biased Assimilation

By far the most widely discussed polarization phenomenon is biased assimilation. Biased assimilation is a phenomenon about how people gather and scrutinize evidence. For example, in the most-cited biased assimilation study, subjects were given equivocal evidence about the efficacy of the death penalty (Lord et al. 1979). Specifically, they encountered two pieces of inconsistent evidence: one a summary of a study that claimed that states that had the death penalty subsequently had lower murder rates, and the other a summary asserting the opposite, that states with the death penalty had higher murder rates than states without.

Prima facie, one might think that when one is confronted with equivocal evidence, one's beliefs should become, if anything, more tempered, not more extreme. But that is not what was found. Subjects' beliefs strengthened in the direction of their antecedently held belief; the death penalty proponents ended up being even more pro-death penalty, the death penalty opponents became even more anti-death penalty.[15] Though the Lord study just mentioned is the most discussed result of the literature, it isn't nearly the first. It came after almost two decades of dissonance research into the 'selective exposure' effect.[16] Selective exposure effects work in a similar way to the Lord et al.'s study, with the one important difference being that subjects in a selective exposure paradigm are allowed to choose whether to encounter or avoid certain pieces of information. To use a canonical example, imagine a subject who was deciding between buying a Honda and a Toyota, and recently decided to buy the Honda. This subject might then be given a magazine that contains advertisements for both Toyotas and Hondas and asked to peruse the magazine at her leisure. Experimenters would then surreptitiously track how long she looked at Honda ads and Toyota ads as she thumbed through the magazine. Subjects who just bought a Honda would spend much more time looking at Honda ads than at Toyota ads, and spend very little

[15] What a 'proponent' amounted to is someone who antecedently supported the death penalty, thought it had a deterrent effect, and thought the studies backed them up (*mutatis mutandis* for the death penalty opponent).
[16] One can be excused for not knowing this since, curiously, Lord et al. never once refer to this rather voluminous literature. For a classic selective exposure effect (where subjects who smoke push a button to add static to an anti-smoking message) see Brock and Balloun (1967); for a wider overview of the phenomenon see Zillman and Bryant (2013.)

if any time looking at Toyota ads. Seeing the pro-attitudinal advertisement would then lead the subjects to become more confident in their antecedent attitude (that Hondas are better than Toyotas).

Thus in both biased assimilation and selective exposure experiments we find a type of belief polarization. But the type of polarization found here is in how one handles the evidence before them. In the selective exposure paradigm the workings of dissonance dictate where the subjects will attend. For example, the more strongly the subjects hold their beliefs, the more strongly they'll avoid counterattitudinal evidence and encounter pro-attitudinal evidence (Brannon et al. 2007). Note that the effect here is really an effect of avoidance—just like the patient who avoids the doctor's call to maintain their belief in their health, the subject's antecedent belief keeps them sequestered away from information which might disconfirm what they believe.

Unlike the selective exposure researchers, Lord et al. (1979) didn't control for different mechanisms that could lead to their biased assimilation, though one can still speculate. It's reasonable to suppose that their finding is due to differential scrutiny, where subjects thought much harder about the counterattitudinal studies than the proattitudinal ones. The more effort they put in, the more counter arguments they came up with; when they compared their counterexamples to the lack of counterexamples that arose for the pro-attitudinal information (due to their lack of trying to produce such counterexamples) they not only reaffirm but also strengthen their antecedent beliefs.

This type of differential scrutiny is predicted by a few different theories of persuasion (Festinger and Maccoby 1964; Petty and Cacioppo 1986), and differs from the mechanisms at play in selective exposure. Nevertheless, both of these effects pertain to how one gathers evidence: in the one case we ignore evidence, in the other we choose which evidence to scrutinize and which to leave be. Although at first blush biased assimilation seems quite irrational, when seen as a phenomenon of evidence gathering one can argue that it is actually rational. For instance, Kelly (2008) has argued that a rational person could show these biased assimilation effects. Jern et al. (2014) have even gone further to produce Bayesian models that entail biased assimilation effects.[17] So perhaps polarization isn't a problem.

Thus we reach what appears to be another standstill: even though biased assimilation looks bad at first pass, perhaps Bayesians can handle the phenomenon. But the other type of polarization evidence—what happens when one's belief is disconfirmed—has been roundly ignored by all parties in the debate. And it is this evidence that cannot be handled by Bayesian theories of any stripe. Once this effect is clear, we can turn back to the modeling of biased assimilation and see how poorly Bayesian models actually handle the data.

3.3.2 Polarization via Belief Disconfirmation

In the late 1800s, August Petermann was the world's most famous geographer. This was all the more impressive for his being an armchair geographer—he rarely left his perch in Gotha, Germany. In particular, Petermann was famous for his maps of the Polar Regions, and he was a loud proponent of the 'open polar sea' theory—the idea that the ice pack in the Arctic thinned out as one reached the North Pole. Petermann, hypothesized that in the summer the northern Arctic Ocean would be totally free of ice. Of course, he never saw any such thing—in part because he never made it anywhere near the arctic, and in part because the

[17] That said, I find the Jern et al. models to be quite implausible. It is difficult to believe that people actually have priors similar to the ones built into their models. For example, in order to explain the Lord et al. the model dictates that people assume a) that all studies are infused with research bias (so that researchers just uncover effects that are consistent with their own beliefs) and b) that the majority of people disagree with one's own opinion. No evidence is given for either prior, and b) in particular seems quite hard to swallow. It is instructive to go through the details of their models before believing that they have successfully modeled human behavior (their other models have similarly hard to believe priors). For more problems with these models see the next section.

hypothesis isn't true. Nevertheless, his reputation lent credence to the open polar sea hypothesis and multiple voyages attempted to reach the pole, risking their lives on Petermann's guess. In 1875 the *HMS Discovery* and the *HMS Alert* set off to win the pole only to find that Petermann was wrong—there was no open polar sea, just a solid sea of ice. After battling scurvy, snow-blindness, and other maladies, the ships broke free of the pack ice and returned to the UK with the news that there was no open polar sea. Though such news echoed what was already known from other disasters (such as the 1871 voyage of the *Polaris* which met with a similar fate) when Petermann found out that his theory was disconfirmed he doubled down on the theory, not just having the belief persist, but instead actually increase in strength. Petermann began to openly proselytize to others, lobbying the German government to sponsor an expedition to the pole. When Germany wouldn't finance an expedition, Petermann turned his efforts to America, and convinced the owner of the New York Herald (James Gordon Bennett) and the US Navy to back another expedition to the North Pole through the open polar sea. The result was the catastrophic voyage of the *USS Jeanette*.[18]

What caused Petermann to increase his confidence in the open polar sea hypothesis even after receiving the earlier gruesome disconfirming evidence? And more importantly, was he particularly special in his irrationality? Seemingly not. There is a long history of people increasing their beliefs after receiving disconfirming evidence. The *locus classicus* for such evidence is Festinger et al. (1956), where researchers tracked a millennial cult. The cult predicted that the world would end on December 21st, 1954. Cult members didn't merely make some assertions that the world would end then—they staked their lives and reputations on it, quitting their jobs, emptying their savings, and preparing for their future

[18] Petermann died before news of the *Jeannette's* catastrophe made it back to Europe. For more on the *Jeannette* see Sides (2014).

life post-destruction.[19] When the date came and went the group had their belief in the world's impending destruction emphatically disconfirmed. Yet after the disconfirmation the cult members didn't merely accept the disconfirmation, decrease their belief accordingly, and lower their commitment to the group's prophecies; rather they increased their commitment to the cult and began proselytizing in earnest. Again, there was nothing particularly special about this millennial cult: members of 12 of the 13 cults who had made specific millennial prophecies (i.e., picked a particular date on which the world would end), increased their proselytizing and their beliefs in the cult post-prophecy disconfirmation (Dawson 1999).

There are reasons one might be skeptical of these data. For one thing, the number of cults one can track is small. Accordingly, one might think that there are so few of these millennial cults because very few people are so irrational. These are *cult members* after all. Moreover, we don't exactly know what happens with their particular belief. Sure they believed *in* the cult, but what about their belief that the world would end on a particular day—did they increase their credence in that proposition after disconfirmation?

Such worries make the cult literature more suggestive than deeply problematic. But the theme is replicable experimentally in populations outside of cult members, even when we keep the belief's content fixed. And it is this datum—people increasing their belief that p after receiving evidence that not-p—that Bayesianism cannot handle. Take, for instance, Batson (1975), where subjects were split into two groups—those who antecedently believed that Jesus was the Son of God and those who did not.[20] Subjects were then asked to read an article they were told was "denied publication in the New York Times at the request of the World Council of Churches because of the obvious crushing effect it would have on the

[19] They thought they'd be whisked up aboard an alien spaceship and avoid the world-destroying deluge.
[20] Subjects were individuated based on their responses to a pretest attitude measure which asked questions such as "Jesus actually performed miracles," "The Bible contains many errors," "The Bible is the Word of God," and "Jesus was only human."

entire Christian world" (180). The article explained that "scholars in Jordan have conclusively proved that the major writings in what is today called the New Testament are fraudulent" for archeologists had unearthed letters from the authors of the New Testament which stated that they knew that Jesus wasn't the Son of God (ibid.).[21] The article went on to say that through radiocarbon dating the letters were shown to be real, and thus the head researcher on the project was forced to reluctantly conclude that the letters are authentic.

After reading the article, participants were then asked to do two things: say whether they believed the article or not and then take another test to see how their attitudes about Jesus had changed. The results were instructive. Unsurprisingly, those who didn't believe that Jesus was the Son of God tended to believe the article, and then increased their belief that Jesus wasn't the Son of God after reading the article. Those who did believe that Jesus was the Son of God and did not believe the truth of the article didn't have their belief that Jesus was the Son of God change at all. This is also unsurprising: most of these participants had a high belief to begin with and the easiest thing for them to do is to reject the possible disconfirming evidence. Once such evidence is rejected, their belief is not under threat and needn't be managed at all in either direction.

However, the most interesting results came from the group consisting of antecedent believers (theists) who also believed the article to be true. These participants are ones who believed that p (that Jesus was the Son of God) and agreed that they just received convincing evidence that not-p. Like Petermann and the millennial cultists before them, these subjects *increased* their belief in p. That is, they now believed even more that Jesus was the Son of

[21] In particular the letters supposedly said, "I am sure we were justified in stealing away his body and claiming that he rose from the dead. For, although his death clearly proves he was not the Son of God as we had hoped, if we did not claim that he was, both his great teaching and our lives as his disciples would be wasted!" (ibid. 180).

God after receiving information that they accepted that purported to show that Jesus was not the Son of God.

One might be tempted to think that only the religious are irrational, but such a hypothesis is unfounded. The belief disconfirmation effect—increasing p after receiving information that not-p—isn't bound to religion at all. One can uncover it in a variety of guises, whether one is disconfirming the belief that there was a conspiracy to assassinate JFK (McHoskey 1995), disconfirming one's opinions on affirmative action and gun control (Taber and Lodge 2006), disconfirming one's belief that coffee isn't unhealthy (Liberman and Chaiken 1992), disconfirming one's belief in the safety of nuclear power (Plous 1991), disconfirming stereotypes about homosexuals (Munro and Ditto 1997), or disconfirming the societal utility of birth control (Kiesler 1971).

There are two morals worth highlighting from the belief disconfirmation effect. The first is that it is anathema to any Bayesian model; one can choose whatever priors one would like, but an updater that increases belief that p after receiving and accepting not-p cannot be modeled as a non-Bayesian updater. The belief disconfirmation effect's power to break the Bayesian stalemate lies in its perversity: it dictates that one increases their belief when one accepts that the belief is under legitimate threat. It is this perversity of updating that is inherently anti-Bayesian. The second moral of the belief disconfirmation effect is that it isn't accidental or due to some performance effect; rather, it arises because of the workings of the *Psychological Immune System*.

4. The Psychological Immune System

The last claim to be defended is that belief disconfirmation effect isn't a mere error in a system's processing but rather stems from a system that is properly functioning. I don't intend to prove that this is so, but to elucidate a hypothesis that entails the possibility, and show that for all we know it isn't false.

From decades of dissonance research we know that receiving disconfirming evidence puts one into a negative, phenomenologically distinct, motivational state; in other words, receiving counterattitudinal information actually causes discomfort—it *hurts* (Elliot and Devine 1994). People will then change their attitudes not to adjust them in line with a norm of truth (*pace* Velleman 2000), but to escape psychological discomfort. Returning to the religious believers in Batson (1975), those who believed that Jesus was the Son of God but didn't believe the counterattitudinal article, didn't rationally need to adjust their belief in Jesus: the dissonance they felt from reading the article caused them to reject the veracity of the article. But those who accepted the veracity of the article and antecedently believed in Jesus were put into an extremely dissonant state. They resolved this dissonance by reaffirming their antecedent belief, and increasing their belief in Jesus. Such adjustment is in line with what we know of the laws of belief: people will adjust their beliefs to avoid psychological discomfort. And it is this fact that is the basis of the Psychological Immune System (Gilbert 2006). Among whatever other laws there are about belief change, we know that there is a basic Psychological Immune System (henceforth, P.I.S.) at work, constantly adjusting beliefs to ward off serious threats to one's sense of self.

Such adjustments don't just happen for any old counterattitudinal information. Just as the physical immune system doesn't get set off for just any infection, so too the threat that sets off the P.I.S. must be substantial. That is, the disconfirming information must attack beliefs that are strongly held in a subjectively important way—in other words, beliefs that one self-identifies with. The more the person self-identifies with a certain belief, the more likely the P.I.S. will be activated when that belief is under attack.

To their credit, Bayesians have noticed the connection between perverse updating and strength of belief. For instance, Jern et al. (2014) note "A similar result was reported by McHoskey (1995), who asked supporters of the official account of the JFK assassination and supporters of an alternative conspiracy account to read a summary of arguments for each account. Those with strong prior beliefs diverged and those with weak prior beliefs did not. The Bayes nets presented in this section cannot account for the fact that only the participants in these studies with strong prior beliefs diverged" (213). The P.I.S. hypothesis can explain the results that Bayes nets could not model, for the P.I.S. interprets the counterattitudinal evidence as threats to the self that have to be warded off; the greater the threat, the greater the response. Just like the physical immune system, the P.I.S. works the strongest when the threat is greatest.

With the P.I.S. in hand, we can now break the gridlock that has undergirded so much of the Bayesian debate. For example, Bowers and Davis (2012) point out that if people were updating as Bayesians, then soccer goalies should act differently than they do. On penalty kicks, goalies should wait to jump until the ball is kicked. But doing so is rare: most goalies guess which way the ball will be kicked and jump before the kick. Nevertheless, Bayesians have a response to this apparently suboptimal behavior. They argue that it is actually optimal once you understand what goalies are actually maximizing, which isn't just goals allowed but instead regret (Bar-Eli et al. 2009). Bar-Eli et al. argue that goalies are calculating not only the goals they will allow, but also their reaction to the outcome. Roughly, Bar-Eli et al. reason that the goalies will regret not jumping early more than jumping early. Since they are trying to optimize both the goals stopped and their future regret, they will tend to jump early even if that is an overall worse strategy.

Attempting to rectify this situation is difficult business. But it can be sidestepped—as opposed to deciding which explanations are ad-hoc, which priors are actually derived and used, which ends are at play in which cases, etc., we have found a set of cases where belief updating itself is perverse. Belief disconfirmation based updating—raising one's credence that P after accepting not-P, is the one fixed point that no Bayesian can allow.

4.1 Looking Back and Wrapping Up

Now that the P.I.S. hypothesis is on the table, we can interpret some earlier effects in the light of it. Belief polarization due to biased assimilation also appears to be due to the P.I.S. Since encountering disconfirming evidence hurts (and encountering confirmatory information feels good) selective exposure is the P.I.S. working prophylactically. Likewise, in differential scrutiny cases one is motivated to scrutinize and reject the disconfirmation information (while being motivated to just passively accept the confirming information) in order to keep one's beliefs intact.

Similarly, in cases where one's antecedent strength of belief is middling, the P.I.S. would predict effects that are closer to belief perseverance than belief polarization. In these cases the beliefs in question (e.g., the relation between being a firefighter and one's risk preferences) aren't ones that people deeply self-identify with. Thus, the threat isn't large enough to need to reaffirm and increase the strength of belief, so one sees little increase in credence.[22] The beliefs here persist because it feels easier to do so than to change one's beliefs. Take, for instance, a case in which a subject is asked to figure out probabilities that a chip will be taken from a bag. Once participants form their initial belief it is easier to just persist in this belief than to update based on incoming information, especially when

[22] From a measurement standpoint, this is surprising: the more middling the antecedent beliefs are, the more space they have to move on the scale post-disconfirmation. Hence, the null hypothesis should be that one would expect more measurable polarization from middling than extreme attitudes. This makes the existence of belief disconfirmation effects all the more astounding for most of these subjects are near ceiling in their attitudes to begin with. Since the stronger one holds a belief, the more likely they are to polarize after receiving disconfirming evidence, detecting such effects are less likely because of ceiling effects in attitude operationalization.

participants don't particularly care about the contents of the particular beliefs under disconfirmation.[23]

Which brings us to the core of the P.I.S. The P.I.S. takes part in a tradition running from Freud through Festinger, Aronson, and Gilbert: it understands the workings of cognition through principles of cognitive economy—the beliefs one changes (or keeps) are due to what feels easiest to do while keeping one's self-image intact. Similarly, like Freud's unconscious and dissonance theory, the P.I.S. gives cognition an engine: one can leverage the fact that inconsistencies hurt to explain how the shape of one's web of belief will change. In particular, the P.I.S. adds the notion of the self as core to what sorts of inconsistencies hurt the most: ones that challenge the sense of self. Inconsistencies due to beliefs one self-identifies with are the ones that cause the most drive to ward off psychological threats.

This isn't to say that there aren't other laws of belief unconnected to the P.I.S. that may be lurking in cognition. Just as there are multiple disconnected processes with their own laws and generalizations to be had in perception, the same can be true in cognition. In fact, I'm fairly sure there are laws of belief beyond the P.I.S.—for example, laws of belief acquisition that are orthogonal to the P.I.S. (Mandelbaum 2014). Perhaps some other laws of updating align with truth tracking, or even Bayesian updating. Perhaps for beliefs that are very disconnected from the self one can update in the way Bayesians predict.

[23] If one combines the P.I.S. with a theory in which merely entertaining a hypothesis raises the credence in the hypothesis (Mandelbaum 2014), then one can explain an even broader set of findings that were previously deemed to be the result of performance constraints. For example, Pitz et al. (1967) found that "The change towards certainty following a confirming event was greater than change towards uncertainty following a disconfirming event" and that "many subjects continued to revise their probability estimates upwards, or else left them unchanged, following a single disconfirming event" (391). If one assumes that merely contemplating the hypothesis raises the credence then we don't have to conclude, as Pitz et al. do, that "the probability estimation task is too unfamiliar and complex to be meaningful," but instead that the subjects' beliefs adjust based on what feels easiest (ibid.). That said, the P.I.S. would expect different results for more motivated subjects, or for subjects who did self-identify with the task (if say, they thought the task reflected their intellectual competence).

Regardless of whether there are or not, one can glean more general morals from this story. The first is that Quine was wrong: the center of one's web of belief isn't constituted by beliefs that are necessarily true, but rather is constituted by the beliefs with which people self-identify. Outside of academic philosophy, people don't care that 2+2=4 in all possible worlds, but people absolutely do care that they are seen as moral, smart, and competent (Thibodeau and Aronson 1992). Try to convince the average person that in fact there are worlds in which the laws of arithmetic don't hold by telling them that there are mathematicians who have shown this—the average person will probably shrug their shoulders. But try to convince the average person that highly trained ethicists have discovered that in fact they are extremely immoral and be ready for a quarrel. The P.I.S. inverts Quine's web, putting highly contingent propositions—that we are good, smart, competent people—at the center of the web, while banishing truths that have little to do with the self to the periphery.

The second major moral of the P.I.S. is that Imperial Bayesianism is false, and Local Bayesianism is false at least when it comes to belief updating. And since belief updating is the natural home for Bayesianism, this should gives us pause when considering the massive Bayesian takeover that is prophesied to happen in our cognitive science. Ironically, it's not entirely implausible that we end with a picture of the mind where the faculties of sensation and perception are the home of Bayesian updating (Girschick et al. 2011; Rescorla 2016), while the workings of cognition bend away from Bayes and towards conceptions of the self.

Acknowledgements and Funding

(Redacted for review; Funding received from PSC-CUNY Award 68282-00 46)

References

Anderson, C.A. (1983). Abstract and concrete data in the perseverance of social theories: When weak data lead to unshakeable beliefs. *Journal of Experimental Social Psychology*, 19(2), 93–108. doi: 10.1016/0022-1031(83)90031-8

Anderson, C.A. (1989). Causal reasoning and belief perseverance. In D.W. Schumann (Ed.), *Proceedings of the Society for Consumer Psychology*, Knoxville, TN: University of Tennessee Press, 115–120.

Anderson, C.A., Lepper, M.R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, 39(6), 1037–1049. doi: 10.1037/h0077720

Anderson, C.A., & Sechler, E.S. (1986). Effects of explanation and counterexplanation on the development and use of social theories. *Journal of Personality and Social Psychology*, 50(1), 24–34. doi: 10.1037/0022-3514.50.1.24

Baeyens, F., Eelen, P., Van den Bergh, O., & Crombez, G. (1990). Flavor-flavor and color-flavor conditioning in humans. *Learning and Motivation*, 21(4), 434–455. doi: 10.1016/0023-9690(90)90025-J

Bar-Eli, M., Azar, O.H., & Lurie, Y. (2009). (Ir)rationality in action: Do soccer players and goalkeepers fail to learn how to best perform during a penalty kick? *Progress in Brain Research*, 174, 97–108. doi: 10.1016/S0079-6123(09)01309-0

Batson, C.D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology*, *32*(1), 176–184. doi: 10.1037/h0076771

Bennett, D. (2015). The neural mechanisms of Bayesian belief updating. *The Journal of Neuroscience*, 35(50), 16300–16302. doi: 10.1523/JNEUROSCI.3742-15.2015

Bogacz, R. (2007). Optimal decision-making theories: Linking neurobiology with behaviour. *Trends in Cognitive Sciences*, 11, 118–125.Bowers, J. & Davis, C. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138, 389–141. doi: 10.1016/j.tics.2006.12.006

Brannon, L.A., Tagler, M.J., & Eagly, A.H. (2007). The moderating role of attitude strength in selective exposure to information. *Journal of Experimental Social Psychology*, 43(4), 611–617. doi: 10.1016/j.jesp.2006.05.001

Brock, T.C., & Balloun, J.L. (1967). Behavioral receptivity to dissonant information. *Journal of Personality and Social Psychology*, 6(4), 413–428. doi: 10.1037/h0021225

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204. doi: 10.1017/S0140525X12000477

Danks, D. (2006). (Not) learning a complex (but learnable) category. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, Mahwah, NJ: Lawrence Erlbaum Associates, 1186–1191.

Danks, D. 2013. Moving from levels & reduction to dimensions & constraints. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society,* Austin, TX: Cognitive Science Society, 2124-2129.

Davies, M.F. (1997). Belief persistence after evidential discrediting: The impact of generated versus provided explanations on the likelihood of discredited outcomes. *Journal of Experimental Social Psychology*, 33(6), 561–578. doi: 10.1006/jesp.1997.1336

Dawson, L. (1999). When prophecy fails and faith persists: A theoretical overview. *Nova Religio*, 3(1), 60–82. doi: 10.1525/nr.1999.3.1.60

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines*, 21(3), 389–410. doi: 10.1007/s11023-011-9241-3

Ecker, U.K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, 18(3), 570–578. doi: 10.3758/s13423-011-0065-1

Elliot, A.J., & Devine, P.G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67(3), 382–394. doi: 10.1037/0022-3514.67.3.382

Elqayam, S., & Evans, J.S.B. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, 34(5), 233–248. doi: 10.1017/S0140525X1100001X

Endress, A.D. (2013). Bayesian learning and the psychology of rule induction. *Cognition* 127(2), 159–176. doi: 10.1016/j.cognition.2012.11.014

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, 12(5), 391–396. doi: 10.1111/1467-9280.00372

Feldman, N.H., Griffiths, T.L., & Morgan, J.L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116, 752–782. doi: 10.1037/a0017196

Festinger, L., & Maccoby, N. (1964). On resistance to persuasive communications. *The Journal of Abnormal and Social Psychology*, 68(4), 359–366. doi: 10.1037/h0049073

Festinger, L., Riecken, H.W., & Schachter, S. (1956). *When Prophecy Fails*. Minneapolis: University of Minnesota Press. doi: 10.1037/10030-000

Frank, M.C. (2013). Throwing out the bayesian baby with the optimal bathwater: Response to Endress (2013). *Cognition*, 128(3), 417–423. doi: 10.1016/j.cognition.2013.04.010

Frank, M.C., & Tenenbaum, J.B. (2011). Three ideal observer models of rule learning in simple languages. *Cognition*, 120(3), 360–371. doi: 10.1016/j.cognition.2010.10.005

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, 62(2), 1230–1233. doi: 10.1016/j.neuroimage.2011.10.004

Garcia, J., & Koelling, R.A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, 4(1), 123–124. doi: 10.3758/BF03342209

Gigerenzer, G. (2008). *Rationality for Mortals: How People Cope with Uncertainty*. Oxford: OUP.

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

Girshick, A., Landy. M., & Simoncelli, E. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7), 926–932. doi: 10.1038/nn.2831

Goodman, N.D., Frank, M.C., Griffiths, T.L., Tenenbaum, J.B., Battaglia, P.W., & Hamrick, J.B. (2015a). Relevant and robust. A response to Marcus & Davis (2013). *Psychological Science*, 26, 539–541. doi: 10.1177/0956797614559544

Goodman, N.D., Tenenbaum, J.B., & Gerstenberg, T. (2015b). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts*, Cambridge, MA: MIT Press, 623–655.

Griffiths, T.L., & Tenenbaum, J.B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17, 767–773. doi: 10.1111/j.1467-9280.2006.01780.x

Gweon, H., & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524-1524. doi: 10.1126/science.1204493

Hohwy, J. (2013). *The Predictive Mind*. Oxford: OUP. doi: 10.1093/acprof:oso/9780199682737.001.0001

Jern, A., Chang, K.M.K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, 121(2), 206–224. doi: 10.1037/a0035941

Jones, M., & Love, B. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4), 169–188. doi: 10.1017/S0140525X10003134

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, 80(4), 237–251. doi: 10.1037/h0034747

Kelly, T. (2008). Disagreement, dogmatism, and belief polarization. *The Journal of Philosophy*, 105(10), 611–633. doi: 10.5840/jphil20081051024

Kiesler, C.A. (1971). *The Psychology of Commitment: Experiments Linking Behavior to Belief*. New York: Academic Press.

Liberman, A., & Chaiken, S. (1992). Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin*, 18(6), 669–679. doi: 10.1177/0146167292186002

Lord, C.G., Ross, L., & Lepper, M.R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098–2109. doi: 10.1037/0022-3514.37.11.2098

Mandelbaum, E. (2014). Thinking is believing. *Inquiry* 57(1), 55–96. doi: 10.1080/0020174X.2014.858417

Mandelbaum, E. (2015). The automatic and the ballistic: Modularity beyond perceptual processes. *Philosophical Psychology* 28(8), 1147–1156. doi: 10.1080/09515089.2014.950217

Mandelbaum, E. (2016). Associationist theories of thought. *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition), E.N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2016/entries/associationist-thought/>.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

McHoskey, J.W. (1995). Case closed? On the John F. Kennedy assassination: biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology*, *17*(3), 395–409. doi: 10.1207/s15324834basp1703_7

Munro, G.D., & Ditto, P.H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, 23(6), 636–653. doi: 10.1177/0146167297236007

Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H. (Forthcoming). Rational learners and moral rules. *Mind & Language*.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113, 327–357. doi: 10.1037/0033-295X.113.2.327

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4), 608-631. doi: 10.1037/0033-295X.101.4.608

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: OUP. doi: 10.1093/acprof:oso/9780198524496.001.0001

Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, 32, 69–84. doi:

10.1017/S0140525X09000284

Petty, R.E., & Cacioppo, J.T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19, 124–192. doi: 10.1016/s0065-2601(08)60214-2

Pitz, G.F., Downing, L., & Reinhold, H. (1967). Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 21(5), 381–393. doi: 10.1037/h0082998

Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer?. *Journal of Applied Social Psychology*, 21(13), 1058–1082. doi: 10.1111/j.1559-1816.1991.tb00459.x

Quilty-Dunn, J. & Mandelbaum, E. (MS). Inferential transitions.

Rescorla, M. (2016). Bayesian Sensorimotor Psychology. *Mind & Language*, 31(1): 3-36. doi: 10.1111/mila.12093

Ross, L., Lepper, M.R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, 32(5), 880–892. doi: 10.1037/0022-3514.32.5.880

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks,* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Sides, H. (2014). *In the Kingdom of Ice: The Grand and Terrible Voyage of the USS Jeannette.* New York: Doubleday.

Skinner, B.F. (1974). *About Behaviorism*. Knopf: New York

Slusher, M.P., & Anderson, C.A. (1989). Belief perseverance and self-defeating behavior. In R.C. Curtis (ed.), *Self-Defeating Behaviors: Experimental Research, Clinical Impressions,* *and Practical Implications*, New York: Plenum Press, 11–40. doi: 10.1007/978-1-4613-0783-9_2

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(3), 303–333. doi: 10.1016/s0065-2601(08)60214-2

Taber, C.S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755–769. doi: 10.1111/j.1540-5907.2006.00214.x

Taylor, E.G., & Ahn, W.K. (2012). Causal imprinting in causal structure learning. *Cognitive Psychology*, 65(3), 381–413. doi: 10.1016/j.cogpsych.2012.07.001

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331 (6022),1279-1285. doi: 10.1126/science.1192788

Thibodeau, R., & Aronson, E. (1992). Taking a closer look: Reasserting the role of the self-concept in dissonance theory. *Personality and Social Psychology Bulletin*, 18(5), 591–602. doi: 10.1177/0146167292185010

Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science*, 3(5), 305–309. doi: 10.1111/j.1467-9280.1992.tb00678.x

Valins, S. (1974). Persistent effects of information about internal reactions: Ineffectiveness of debriefing. In R.E. Nisbett & H. London (eds.), *Thought and Feeling: The Cognitive Alteration of Feeling States*, Transaction: Piscataway, NJ, 116–124.

Velleman, D. (2000). On the aim of belief. In his *The Possibility of Practical Reason*, Oxford: OUP, 244–281.

Vlassis, N., Ghavamzadeh, M., Mannor, S., & Poupart, P. (2012). Bayesian reinforcement learning. In M. Wiering & M. van Otterlo (eds.), *Reinforcement Learning*, Berlin: Springer, 359–386. doi: 10.1007/978-3-642-27645-3_11

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, 19(7), 645–647. doi: 10.1111/j.1467-9280.2008.02136.x

Vul, E., Goodman, N., Griffiths, T.L., & Tenenbaum, J.B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4), 599–637. doi: 10.1111/cogs.12101

Wegner, D.M., Coulton, G.F., & Wenzlaff, R. (1985). The transparency of denial: Briefing in the debriefing paradigm. *Journal of Personality and Social Psychology*, *49*(2), 338. doi: 10.1037/0022-3514.49.2.338

Weiss, Y., Simoncelli, E.P., & Adelson, E.H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5, 598–604. doi: 10.1038/nn0602-858

West, R.F., & Stanovich, K.E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, 31(2), 243–251. doi: 10.3758/BF03194383

Xu, F., & Tenenbaum, J.B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245–272. doi: 10.1037/0033-295X.114.2.245

Zillmann, D., & Bryant, J. (2013). *Selective Exposure to Communication*. New York: Routledge.