# Biased Belief in the Bayesian Brain: A Deeper Look at the Evidence

Ben M. Tappin [1], Stephen Gadsby [2]

[1] Royal Holloway, University of London, United Kingdom
[2] Monash University, Australia

This article is published in *Consciousness and Cognition* (doi: https://doi.org/10.1016/j.concog.2019.01.006). This document reflects the final version prior to copy-editing and is posted to enable open-access.

## Author Note

We are grateful to Dan Williams, Jakob Hohwy and Jon Baron for comments on an earlier version of the manuscript; to Brendan Nyhan for email correspondence; and to two anonymous reviewers for their helpful comments and suggestions. This research was supported by an Economic and Social Research Council scholarship (awarded to BMT) and an Australian Government Research Training Program scholarship (awarded to SG). The funding bodies had no role in the writing of this manuscript or the decision to submit for publication.

Correspondence concerning this article should be addressed to Ben M. Tappin, Department of Psychology, Royal Holloway University of London, United Kingdom TW20 0EX. Email: benmtappin@googlemail.com.

Abstract

A recent critique of hierarchical Bayesian models of delusion argues that, contrary to a key assumption of these models, belief formation in the healthy (i.e., neurotypical) mind is manifestly non-Bayesian. Here we provide a deeper examination of the empirical evidence underlying this critique. We argue that this evidence does not *convincingly* refute the assumption that belief formation in the neurotypical mind approximates Bayesian inference. Our argument rests on two key points. First, evidence that purports to reveal the most damning violation of Bayesian updating in human belief formation is counterweighted by substantial evidence that indicates such violations are the rare exception—not a common occurrence. Second, the remaining evidence does not demonstrate *convincing* violations of Bayesian inference in human belief updating; primarily because this evidence derives from study designs that produce results that are not obviously inconsistent with Bayesian principles.

Keywords: The backfire effect, confirmation bias, motivated reasoning, Bayesian inference, hierarchical Bayesian models of delusion

# Biased Belief in the Bayesian Brain: A Deeper Look at the Evidence

## 1. Bayesian Models of Cognition

Recent years have witnessed an explosion in theoretical and empirical interest in hierarchical Bayesian models of cognition (Clark, 2013, 2016; Friston, 2005; Hohwy, 2013). An underlying principle of these models is the *predictive processing* view of cognition, which characterizes the human mind as an anticipatory predictive engine whose primary drive is to minimize the mismatch between its model of the world and incoming sensory input—referred to as *prediction error*. The putative structure of this process of "prediction error minimization" is *hierarchical*—from perceptual content at the lower levels through to abstract beliefs at the highest levels. Distinct levels act as priors for the levels immediately below, and only the unpredicted aspect of a signal (i.e., the prediction error) is propagated up the hierarchy.

This account entails a unique story about how humans update their beliefs. Namely, via the adjustment of higher level priors in an attempt to minimize prediction error fed forward from the lower levels of the hierarchy. Furthermore, this process constitutes a neurobiologically-plausible account of how Bayesian inference is implemented in the human brain. Specifically, because a system which operates according to the principles of prediction error minimization will necessarily approximate Bayesian inference over the long term (Hohwy, 2017)[1].

One of the more promising aspects of hierarchical Bayesian models of cognition is their potential to explain *delusional beliefs* as a function of aberrant processing in the aforementioned hierarchy (e.g., Adams et al., 2013; Fletcher & Frith, 2009; Corlett et al., 2010). At the heart of this approach is the notion of *excessive* precision—whether this pertains to overly-precise priors or overly-precise prediction error (Williams, 2018, p. 133). Over-precision is suggested to produce two kinds of delusion-relevant effects. The first is that prediction error signal which ought to be disregarded propagates up the hierarchy, eventually having to be *explained away*—with the agent revising her model of the world and forming a delusional belief in the process. The second is that overly-precise (i.e., delusional) priors exert *top down* influence on incoming sensory evidence—shaping the way the agent sees the world.

In a recent article (*Consciousness and Cognition, 61, 129-147, 2018*), Williams draws on several arguments to advance a critique of these hierarchical Bayesian models of delusion (HBMD). One of his key arguments is that belief formation in the healthy (i.e., neurotypical) mind is manifestly *non-Bayesian*.[2] Thus, he argues, this undermines the case for characterizing delusionality as arising from deficits in otherwise Bayes optimal processing. In support of his argument that belief formation in the neurotypical mind is manifestly non-Bayesian, Williams (2018) cites three phenomena:

1. The Backfire Effect

[1] The notion of Bayesian inference *over the long-term* is key here. Isolated deviations from Bayes optimality do not constitute credible counter examples to the predictive processing account, unless it can be shown that they are part of a *consistent* pattern of non-Bayesian belief formation.

[2] More accurately, the argument rests on whether belief formation *approximates*, rather than exactly implements Bayesian inference (Williams, 2018, p. 140). When using terms such as "Bayesian" or "non-Bayesian", we refer here to processes which approximate Bayesian inference in some principled manner.

2. Confirmation Bias
3. Motivated Reasoning

Here we expand on Williams' (2018) treatment of these phenomena. Specifically, we consider more deeply the extent to which empirical evidence for each of these phenomena demonstrate that belief formation in the neurotypical mind violates Bayesian inference. Ultimately, we conclude that this evidence does not *convincingly* demonstrate a violation of Bayesian inference. Our conclusion rests on two key points. First, while the backfire effect has been documented in a number of studies—and indeed seems problematic for many Bayesian models—a substantial number of recent studies find no evidence of backfire, even under theoretically favorable conditions. This evidence suggests the phenomenon is considerably less prevalent than assumed, and, by Williams' (2018) own criterion, does not therefore undermine Bayesian models of cognition (at least, not convincingly). Second, paradigmatic studies of confirmation bias and motivated reasoning fail to demonstrate convincing violations of Bayesian inference, for three reasons. First, the key outcome variable in many of the relevant studies is not belief updating, but *evidence evaluation*. Second, the designs of these studies often preclude causal inferences on the role of motivation in evidence evaluation. Third, where patterns of evidence evaluation in such studies are subjected to formal Bayesian analysis, they appear consistent with Bayesian principles. Taken together, this makes it difficult to interpret evidence for these phenomena as a convincing refutation of Bayesian inference.

## 2. Evidence of Non-Bayesian Belief Formation

In this section, we offer a more thorough discussion of the empirical evidence underlying the three phenomena cited by Williams: the backfire effect, confirmation bias and motivated reasoning. The empirical literature on each of these topics is vast and heterogeneous, and discussing all the evidence is well beyond the scope of this article. We thus focus primarily on the evidence referred (or alluded) to by Williams (2018).

### 2.1. The Backfire Effect

The backfire effect refers to the phenomenon whereby people become more confident in their initial belief after receiving evidence that contradicts that belief. Williams (2018) states that backfire is "most damning" (p. 142) for the notion that neurotypical belief formation is Bayesian, a position recently—and forcefully—staked out by others (Mandelbaum, 2018). The notion of backfire is indeed problematic for HBMD, primarily because the phenomenon of increasing confidence in $P$ when presented evidence that $\sim P$ appears to directly contradict key principles of Bayesian inference. However, as we will highlight below, the backfire citations in Williams (2018; and in Mandelbaum, 2018) are highly selective, and, as a result, leave an impression of the prevalence of backfire that is at odds with a large body of recent evidence.

Arguably the most famous evidence of backfire—cited by Williams—is reported in Nyhan and Reifler (2010). Across a series of studies, these researchers found that political conservatives in the US became more confident that (i) Weapons of Mass Destruction (WMD) had been found in Iraq, and (ii) tax cuts increased government revenues, after receiving evidence *contradicting* these claims. Less often acknowledged is that Nyhan and Reifler (2010) did not replicate the WMD backfire result in a second experiment, and they also observed no backfire on the issue of stem cell research; suggesting that backfire is

subject to considerable contextual constraints (a point not lost on the authors at the time). Nevertheless, more recent evidence of similar backfire effects has been reported. For example, in 2013, Nyhan, Reifler and Ubel reported that, after receiving a correction to the myth that the Affordable Care Act would create "death panels", US respondents who were both (i) highly knowledgeable about politics and (ii) supporters of Sarah Palin *increased* their belief in the myth. Schaffner and Roche (2017) also found that US Republicans' beliefs about employment rates appeared to update in the opposite direction to that implied by an employment report released by the Obama administration—suggestive of backfire.

The interpretation of other evidence often cited in support of the backfire effect (e.g., Hart & Nisbet, 2012; Nyhan & Reifler, 2015; Nyhan et al., 2014; Zhou, 2016) is complicated by the fact that the "backfiring" in such experiments occurs not on measures of peoples' descriptive beliefs about the world, but, rather, on their preferences and behavioural intentions. In fact, where researchers have measured both descriptive beliefs *and* preferences/behavioural intentions, "backfire" in the latter is often accompanied by *appropriate* descriptive belief change towards the evidence (e.g., Barrera et al., 2018; Nyhan et al., 2014). In these and similar cases, one is hard pressed to conclude that people backfired in their beliefs—without imposing additional assumptions about the relationship that exists between peoples' descriptive beliefs on the one hand, and their preferences and behavioural intentions on the other.

In contrast to the evidence of backfire outlined above, there are numerous recent studies that find no evidence of backfire of the kind reported in Nyhan and Reifler (2010). We highlight a number of the most rigorous and compelling cases below.

In a series of five experiments—comprising more than 10,000 subjects and 52 different political topics—Wood and Porter (2018) report that they failed to observe a single instance of backfire. This was despite inclusion of numerous "hot button" (US) issues in their experiment, such as gun violence, immigration, crime, abortion, and race—providing theoretically-favorable conditions for backfire to emerge. Guess and Coppock (2018) report a similar result. These authors fielded an experiment on the topic of gun control from 22-28 June 2016; approximately 10 days after the mass shooting in Orlando, Florida. They presented subjects with evidence that gun control policy either (i) decreased or (ii) increased gun violence. Despite the context of the experiment—in the aftermath of a highly-publicized mass shooting—and a large, nationally representative sample of US adults (N=2,122), the authors observed scant evidence of backfire. They also report no evidence of backfire in a further two experiments conducted with large convenience samples (*ibid*.). Coppock (2016, Chapter 3) similarly failed to induce backfire in an earlier experiment that included threatening and insulting language alongside the presentation of evidence—again providing theoretically-favorable conditions for backfire. This result was recently replicated by Kim (2018, Study 4).

Continuing on, both Nyhan and colleagues (2017) and Swire and colleagues (2017) recently conducted a series of experiments in which they presented subjects with corrections of inaccurate statements made by US President Donald Trump. Both sets of researchers found that subjects—in particular, those who were supporters of Donald Trump—consistently incorporated these corrections into their beliefs; that is, they did not backfire. Pennycook and Rand (2017) report that tagging politically-favorable news stories as "disputed by fact checkers" did not cause people to backfire—that is, to rate the news story as *more* accurate. On the contrary, as intended, tags caused people to rate politically-concordant

news stories as less accurate. Barrera and colleagues (2018) found that French voters who supported Marine Le Pen updated their beliefs about immigration—if not their voting intentions—abiding the pro-immigration evidence that they received. Hill (2017) likewise observed that both US Republicans and Democrats updated their beliefs after receiving evidence about the truth (or falsity) of various partisan political facts—even if the evidence was politically uncongenial. Finally, Haglin (2017) recently observed no evidence of backfire in the domain of vaccination safety, in an experiment closely modelled after Nyhan and Reifler (2015).

The above studies were conducted over a range of different topics, sampling populations and experimental designs. Of course, they do not reflect a systematic or exhaustive review of the relevant evidence, but, taken together, provide a strong challenge to the notion that backfire is prevalent—or even common—in human belief updating. On this basis, the backfire effect currently provides rather *un*convincing evidence in support of Williams's (2018; and Mandelbaum's, 2018) argument that belief updating in the neurotypical population violates Bayesian inference[3]. As Williams (2018, p. 140) himself states, "the fact that one's cousin Barry once violated Bayes optimality evidently does not undermine Bayesian models of cognition". Likewise, the notion that *some* people *sometimes* backfire does not convincingly undermine the assumption that belief updating in the neurotypical mind approximates Bayesian inference[4].

## 2.2. Confirmation Bias

Broadly speaking, confirmation bias refers to the phenomenon whereby new information is sought out or interpreted in patterns partial to existing beliefs (e.g., Nickerson, 1998)[5]. In what follows, we address both types of confirmation bias independently (i.e., information search and information interpretation). First, consider confirmation bias in the way information is sought out—that is, information *sampling*. We note that non-Bayesian information sampling could feasibly exist *in conjunction with* Bayesian belief updating of the kind assumed by HBMD. In other words, in principle it is possible that the neurotypical mind combines new information with prior beliefs in a manner that approximates Bayesian inference, but samples—seeks out or otherwise obtains—that information in a distinctly non-Bayesian manner; and does not, or *cannot*, correct for this bias in sampling (Fiedler, 2000)[6].

Such a possibility highlights a "moderately Bayesian" perspective on human belief formation, where belief *updating* may approximate Bayesian inference, but information sampling is underpinned by a process orthogonal to Bayesian inference. While some predictive coding accounts of cognition assume that perception, belief *and* action—including information sampling—are all accountable for within a single Bayesian framework (Friston, 2012; Hohwy, 2013), this stronger position is by no means obligatory. In order for their

[3] Added to this, recent work has pointed out that even the observation of backfire in human belief updating does not imply a cast-iron violation of Bayesian inference. In particular, because the phenomenon of backfire can be observed among unbiased Bayesian updaters (Bullock, 2009; Druckman & McGrath, 2018).

[4] This point is arguably *more* problematic for Mandelbaum (2018), whose *central* claim is that the backfire effect (what he terms "belief disconfirmation based polarization") renders a Bayesian account of human belief updating untenable.

[5] The term has also been applied to the phenomenon of "positive test strategy" (e.g., Klayman & Ha, 1987). Since this also concerns information *search*—not belief updating *per se*—we do not discuss it further here.

[6] Interestingly, recent evidence from the domain of US politics suggests that such biased or "selective" sampling of information (on the internet, at least) may be less prevalent than assumed (e.g., Guess, 2018; Guess, Nyhan, Lyons, & Reifler, 2018; Flaxman et al., 2016; Gentzkow & Shapiro, 2011).

claims about aberrant precision estimation to hold, proponents of HBMD need only accept a picture of cognition whereby perception and belief formation—but not action—are underpinned by a hierarchical generative model, operating according to the principles of precision-weighted prediction error minimization. Given that *most* HBMD appear committed only to this weaker assumption, evidence of (confirmation) biased sampling in the neurotypical mind does not bear very strongly on their validity.

What about confirmation bias in the *interpretation* of information? That is, the notion that "people are more receptive to evidence that confirms their prior beliefs" (Williams, 2018, p. 142; as cited in Mercier & Sperber, 2017, p. 218). The most straightforward empirical evidence for this notion is that individuals are prone to rate information as stronger or more convincing if it confirms vs. contradicts their prior beliefs (e.g., Koehler, 1993; Lord et al., 1979; Taber & Lodge 2006; Tappin, Pennycook, & Rand, 2018). This effect is extremely robust; indeed, people are "often *unable* to escape the pull of their prior beliefs, which guide the processing of new information in predictable ways" (Taber & Lodge, 2006, p. 767, our emphasis). But the critical question is whether this evidence demonstrates a *convincing violation* of Bayesian inference.

Gerber and Green (1999) think not (e.g., see p. 199). The view of these scholars is that peoples' tendency to rate information as more convincing if it aligns with (vs. contradicts) prior beliefs does not reveal a credible violation of Bayesian inference. Indeed, they present a simple yet principled Bayesian model of belief formation that *entails* such a tendency. In doing so, they additionally point out that scholars rarely evaluate the aforementioned tendency with respect to a formal Bayesian model (a point also taken up at length in Hahn & Harris, 2014). Importantly, where such evaluation *has* occurred—like in their model—the interpretation of new information in light of prior beliefs is found to be consistent with Bayesian conditionalization (e.g., see Koehler, 1993). Obviously, this does not imply that such a tendency is desirable or normative by any or all judgmental standards (Cao, Kleiman-Weiner, & Banaji, 2018; Ditto et al., 2018a; Koehler, 1993). Nor does it imply that peoples' evaluations of new information approximate Bayesian inference (see below). However, it *does* imply that one is hard pressed to take observation of such a tendency as convincing evidence *against* the notion that belief formation in the neurotypical mind approximates Bayesian inference (more on this in section 2.3).

It must be noted that Bayesian models can accommodate a wide range of belief phenomena (Bullock, 2009). Indeed, unconstrained flexibility in specification of the prior and likelihood function allows almost any pattern of data to be explained. While the flexibility of Bayesian models has drawn criticism—not least from Williams (2018; see also Bowers & Davis, 2012)—it is only indirectly relevant to the point we make above. Which is that classic evidence of confirmation bias in the interpretation of new information is not a *convincing* refutation of Bayesian inference, given that this pattern of evidence has been shown to be consistent with Bayesian principles. Of course, given this flexibility, the reverse is also true: showing that peoples' patterns of information evaluation are consistent with a Bayesian model does not by itself provide convincing *support* for the notion that neurotypical belief formation approximates Bayes, either. In other words, much of this evidence seems to be symmetrically undiagnostic: it neither convincingly undermines nor supports the notion that belief formation in the neurotypical mind approximates Bayesian inference. One way to increase the diagnosticity of such evidence may be to measure not only the "location" of peoples' prior beliefs and evidence evaluations—as is common in much of the relevant research—but, in addition, to map their precision (i.e., confidence) and distributional shape

(Bullock, 2009). This extra information would serve to constrain the Bayesian expectation, and thus provide a clearer picture of whether, how and to what extent the relevant human data violate Bayesian inference (Bullock, 2009; Gershman, 2018).

### 2.3. Motivated Reasoning

As described in Williams (2018, p. 142; as cited in Kunda, 1990, p. 480), motivated reasoning broadly refers to the phenomenon where people "arrive at conclusions that they want to arrive at when accessing, constructing, and evaluating beliefs." The empirical literature on motivated reasoning is vast and heterogeneous, but Williams alludes to two factors that are posited to affect beliefs in a way that undermines Bayesian inference. They are *preferences*—what people desire to be true—and *identity*—what defines people and their important groups (e.g., political parties). Consequently, where we refer to "motivated reasoning", we mean reasoning motivated by these factors (and not others).

First, we note that we concur with Williams (2018) that motivated reasoning constitutes a clear challenge—in principle—to the assumption that human belief updating approximates Bayesian inference. The notion that beliefs are updated conditional on preferences and identities would seem to suggest that belief updating mechanisms are at best *orthogonal* to Bayesian inference, and, at worst, work directly *against* such principles[7]. However, as we will argue, paradigmatic evidence of motivated reasoning does not convincingly demonstrate such an idealized notion of the phenomenon.

Indeed, turning to this evidence we find that—similar to confirmation bias in the interpretation of new information—the outcome variable in what is arguably the paradigmatic motivated reasoning study design is peoples' evaluations of the *reliability* of new information. For recent and authoritative reviews, we refer to Ditto (2009), Ditto et al. (2018b) and Kahan (2016). It will help to describe the typical design here. The design involves randomly assigning people to one of two (or three, if a control is included) treatments; then, in each treatment, people receive some information. Across treatments, almost all characteristics of the information are held constant, save for the *upshot* of the information—which is manipulated to be consistent with either one type of outcome or another (e.g., that gun control laws *reduce* crime or do *not* reduce crime). Researchers measure peoples' evaluations of the reliability of the information on self-report scales, and, typically, covariates (e.g., political identity) that are to be used in analysis. The critical inferential test is then conducted on the *interaction* between treatment (i.e., information) and covariate (e.g., political identity or some other preference for one outcome vs. another). If peoples' evaluations of information reliability are observed to be *conditional on* their preferences or identities—that is, a statistically significant interaction term—motivated reasoning is typically inferred. As before, one can ask the following question: does such conditional evaluation of the reliability of information provide *convincing* evidence of a violation of Bayesian inference?

---

[7] An exception to this may be Bayesian models in which beliefs are assumed to provide utility *per se*, which can explain patterns of belief updating seemingly at odds with simple Bayesian updating (e.g., see Sharot & Garrett, 2016). As noted in Williams (2018, footnote 15, p. 142), such models "transform the issue from simple inference to the kinds of phenomena modelled in Bayesian decision theory". The possibility of such models suggests that even observation of human belief updating that is conditional on preferences/identities cannot be taken as a cast-iron violation of Bayesian inference. Of course, the mere existence of such models cannot be taken as convincing evidence that "motivated" human belief updating approximates Bayesian inference, either (a point also made by Williams, 2018, footnote 15, p. 142; see also section 2.2 in the current paper).

We suggest that it does not. The primary reason is that, in the foregoing designs, motivation is not randomly assigned. Consequently, the critical inferential test is a *treatment* by *covariate* interaction, where only the former is randomized. This precludes the inference that motivation *causes* the observed patterns of information evaluation (e.g., Gerber & Green, 2012; see also Druckman & McGrath, 2018; Kim, 2018). In other words, it prevents the key inference of motivated reasoning. A direct corollary is that the results from these designs are susceptible to (confounding) explanations based on prior beliefs—because the random assignment of information not only varies the consistency of said information with peoples' preferences, political identities, and so on, but also with their prior beliefs (Tappin, Pennycook, & Rand, 2018); an "empirical catch-22" in motivated reasoning research (Ditto et al., 2018b, p. 13). In this light, then, the observed patterns of information evaluation may reduce to "people are more receptive to evidence that confirms their prior beliefs" (Williams, 2018, p. 142)—and, as pointed out above, confirmation bias in the interpretation of new information does not provide particularly convincing evidence of a violation of Bayesian inference.

Indeed, the tendency to judge the reliability of information sources based on how closely the information matches one's existing beliefs is (i) arguably a sensible way of setting prior beliefs about reliability—particularly when other, more diagnostic information is scarce—and (ii) not obviously inconsistent with Bayesian principles (Baron & Jost, 2018; Gerber & Green, 1999; Koehler, 1993). As Hahn and Harris (2014, p. 90) point out:

> From a Bayesian, epistemological perspective, source and evidence characteristics combine to determine the overall diagnostic value of the evidence. Furthermore, the content of the testimony itself may provide one indicator (and in many contexts our only indicator) of the source's reliability. Recent work in epistemology has thus endorsed the position that message content should impact our beliefs about the source.

This point may be concretely appreciated by considering a fairly mundane example, offered in Gerber and Green (1999, pp. 197-198), which we reproduce here:

> Suppose that you are supervising an employee, and you have questions about the employee's competence. After reviewing the employee's work over the past year and speaking to a dozen of his co-workers, you conclude that the employee is not doing a good job. Just as you are about to call him into your office, you hear back from a final co-worker who says that, in his opinion, the employee is very capable. Although there is no reason a priori to consider this new report any less reliable than the dozen reports already given, it is hardly convincing evidence that the employee is in fact a good worker. It is far more likely that the new report is wrong and that the final co-worker either has poor evaluation skills (this co-worker's "study" has a methodological flaw) or has observed an uncharacteristic performance (the co-worker's "study" presents misleading findings due to "random error").

In sum, if one assumes that people consider the information they receive in typical motivated reasoning studies as not *perfectly* reliable—an extremely weak assumption, to be sure—the observation that people condition their evaluation of the reliability of that information on their prior beliefs seems both defensible, and, more importantly, not obviously inconsistent with Bayesian principles (Gerber & Green, 1999; Hahn & Harris, 2014; Koehler, 1993).

Of course, this argument applies only to those types of motivated reasoning study designs described above—that is, where motivation itself is not randomly assigned—and thus where prior beliefs are liable to confound evaluations of information reliability. However, there are different study designs that lay greater claim to ruling out the confounding influence of prior beliefs and/or licensing causal inference on the role of motivation. For example, some designs attempt to equalize prior beliefs across subjects (for a review of such attempts, see Ditto, 2009); while others randomly assign political party *cues*, rather than measuring political identity as a covariate (e.g., Cohen, 2003), or they randomly assign *threat to*—or *affirmation of*—identity (e.g., Cohen, Aronson, & Steele, 2000; Coppock, 2016; Kim, 2018; Lyons, 2016; Nyhan & Reifler, 2018). We briefly consider each of these three designs in turn.

In our estimation, the first type of design is difficult to rigorously implement in practice—given that the relevant prior beliefs are often likely to be numerous and embedded in a network of possibly interdependent beliefs, all of which may be brought to bear on reasoning (Gershman, 2018). Furthermore, recalling our earlier point, the "location" of subjects' priors is typically all that is measured in such cases—not their precision or distributional shape (cf. Bullock, 2009). In the absence of the latter, it seems hard to confidently rule out their influence. Nevertheless, even if one assumes that prior beliefs are ruled out in paradigmatic motivated reasoning studies (e.g., Ditto & Lopez, 1992; Ditto et al., 1998; Ditto, Munro, et al., 2003), it is still not clear that the results from these studies are a convincing refutation of Bayesian inference. Specifically, because the results are typically interpreted as evidence for the "quantity-of-processing" (QOP) model of motivated reasoning (Ditto, 2009)—which departs in a crucial way from the classic model of motivated reasoning outlined by Kunda (1990) and cited by Williams (2018).

In brief, the classic model assumes that people recruit cognitive processes to reach a particular, desired conclusion—explaining why they evaluate desirable information as more reliable than otherwise-identical undesirable information. The QOP model, on the other hand, explains this pattern of information evaluations by simply assuming that negatively-valenced stimuli trigger more detailed cognitive processing than (otherwise-identical) positively-valenced stimuli; an asymmetry putatively underpinned by the adaptive advantage of allocating more cognitive resources to analyze threats (Ditto, 2009). The upshot of this asymmetry in the quantity of cognitive processing is two-fold. First, "it is almost inevitable that people will be more likely to consider multiple explanations for unwanted outcomes than wanted ones" (Ditto, 2009, p. 34). Second, as a direct result, "people will be more *uncertain* about the validity of preference-inconsistent than preference-consistent information" (p. 34, emphasis in the original). It is this difference in uncertainty, according to the QOP model, that causes individuals to evaluate undesirable information less favorably than otherwise-identical desirable information. Far from refuting Bayesian inference, then, a key assumption of the QOP model—that individuals condition their evaluation of information reliability on its (perceived) uncertainty—seems quite *consistent with* Bayesian principles. Whether a tendency to devote greater cognitive processing to threatening stimuli *per se* constitutes a violation of Bayesian inference is unclear. However, it does not seem unreasonable to consider that people may have prior experience of the benefits of allocating their cognitive resources this way (to our knowledge, this is an open question).

Regarding the second design type—party cues—as noted by Ditto and colleagues (2018a, p. 8) it appears quite reasonable for people to incorporate whether their political party

endorse (or oppose) the information at hand. Indeed, such reliance may *reflect* the role of prior beliefs—for example, beliefs about the relative trustworthiness of one's in-party elites—rather than ruling them out (Druckman & McGrath, 2018). The final design type—threatening or affirming identity—seems a promising design in principle to isolate causal effects of identity-motivation on belief formation. In practice, however, recent attempts at this have met with mixed results (e.g., Coppock, 2016, Chapter 3; Kim, 2018, Study 4; Lyons, 2016; Nyhan & Reifler, 2018). Overall, the evidence for motivated reasoning from these alternative designs either (i) does not convincingly rule out the influence of prior beliefs or (ii) where it does, the evidence appears mixed—and certainly not decisive one way or the other—or is interpreted as support for a model whose key insight appears quite consistent with Bayesian principles. As before, then, we defer to Williams' (2018, p. 140) own criterion: "the fact that one's cousin Barry once violated Bayes optimality evidently does not undermine Bayesian models of cognition". Abiding this criterion, we are reluctant to consider such mixed evidence of motivated reasoning a convincing refutation of Bayesian inference.

## 3. Conclusion

We have argued that the three phenomena cited by Williams (2018)—the backfire effect, confirmation bias and motivated reasoning—do not convincingly refute the notion that belief formation in the neurotypical mind approximates Bayesian inference. Our argument hinged on our deeper examination of the empirical evidence underlying these phenomena. Indeed, our aim in this paper was not to advance the case that human belief formation *does* approximate Bayesian inference; but, rather, to argue that classic evidence of backfire in belief updating, confirmation bias and motivated reasoning does not convincingly *undermine* such a case.

We drew attention to substantial recent evidence that indicates the backfire effect is not as widespread as perhaps assumed; on the contrary, it appears to be the rather rare exception. Furthermore, we highlighted that paradigmatic evidence of confirmation bias and motivated reasoning derives from study designs that (i) measure how people evaluate the reliability of new information, and, in the latter case, (ii) often do not permit causal inferences on the role of motivation. Insofar as these study designs reveal that people condition their evaluation of new information on their prior beliefs, the results are not obviously inconsistent with Bayesian inference. On the contrary, such results appear consistent with Bayesian principles (Gerber & Green, 1999; Hahn & Harris, 2014; Koehler, 1993). Likewise, we pointed out that studies that lay claim to ruling out the role of prior beliefs are taken as evidence for a model of motivated reasoning whose key assumptions appear quite *consistent with* Bayesian principles.

Therefore, in our view one is hard pressed to conclude that evidence of the backfire effect, confirmation bias and motivated reasoning demonstrates that human belief formation is manifestly non-Bayesian. As such, contra Williams (2018), we conclude that these phenomena do not convincingly undermine hierarchical Bayesian models of delusion.

References

Adams, R., Stephan, K., Brown, H., Frith, C., & Friston, K. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*, 47.

Baron, J., & Jost, J. T. (2018). False equivalence: Are liberals and conservatives in the US equally "biased?" *Perspectives on Psychological Science*. Retrieved from https://www.sas.upenn.edu/~baron/papers/dittoresp.pdf

Barrera, O., Guriev, S. M., Henry, E., & Zhuravskaya, E. (2018). Facts, alternative facts, and fact checking in times of post-truth politics. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3004631

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*, 389-414.

Bullock, J. G. (2009). Partisan bias and the Bayesian ideal in the study of public opinion. *The Journal of Politics*, *71*, 1109-1124.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2018). People make the same Bayesian judgment they criticize in others. *Psychological Science*, 1-12.

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

Clark, A. (2016). *Surfing uncertainty*. Oxford: Oxford University Press.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, *85*, 808-822.

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin*, *26*, 1151-1164.

Coppock, A. (2016). *Positive, small, homogeneous, and durable: Political persuasion in response to information*. PhD Dissertation: Columbia University. https://academiccommons.columbia.edu/doi/10.7916/D8J966CS

Corlett, P., Taylor, J., Wang, X., Fletcher, P., & Krystal, J. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology*, *92*, 345–369.

Ditto, P. H. (2009). Passion, reason, and necessity: A quantity-of-processing view of motivated reasoning. *In Delusion and self-deception: Affective and motivational influences on belief formation*, 23-53.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*, 568-584.

Ditto, P. H., Clark, C. J., Liu, B. S., Wojcik, S. P., Chen, E. E., Grady, R. H., ... & Zinger, J. F. (2018a). Partisan bias and its discontents. *Perspectives on Psychological Science*. Retrieved from https://www.researchgate.net/publication/328916374_Partisan_Bias_and_Its_Discontents

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., ... & Zinger, J. F. (2018b). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, 1-19.

Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A., & Lockhart, L. K. (2003). Spontaneous skepticism: The interplay of motivation and expectation in responses to favorable and unfavorable medical diagnoses. *Personality and Social Psychology Bulletin*, *29*, 1120-1132.

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology*, *75*, 53-69.

Druckman, J., & McGrath, M. C. (2018). The evidence for motivated reasoning in climate change preference formation. *Working Paper*. Retrieved from https://www.ipr.northwestern.edu/publications/docs/workingpapers/2018/wp-18-22.pdf

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*, 659-676.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly, 80*, 298-320.

Fletcher, P., & Frith, C. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience, 10*, 48–58.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*, 815-836.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage, 62*, 1230–1233.

Gentzkow, M., & Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics, 126*, 1799-1839.

Gerber, A. S., & Green, D. P. (1999). Misperceptions about perceptual bias. *Annual Review of Political Science, 2*, 189-210.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gershman, S. J. (2018). How to never be wrong. *Psychonomic Bulletin & Review*, 1-16.

Guess, A. (2018). (Almost) everything in moderation: New evidence on Americans' online media diets. *Working Paper*. Retrieved from https://webspace.princeton.edu/users/aguess/Guess_OnlineMediaDiets.pdf

Guess, A., & Coppock, A. (2018). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, 1-19.

Guess, A., Nyhan, B., Lyons, B, & Reifler, J. (2018). Avoiding the echo chamber about echo chambers: Why selective exposure to like-minded political news is less prevalent than you think. *Knight Foundation White Paper*. Retrieved from https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf

Haglin, K. (2017). The limitations of the backfire effect. *Research & Politics, 4*, 2053168017716547.

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased? Motivated reasoning and rationality. In *Psychology of Learning and Motivation* (Vol. 61, pp. 41-102). Academic Press.

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research, 39*, 701-723.

Hill, S. J. (2017). Learning together slowly: Bayesian learning about political facts. *The Journal of Politics, 79*, 1403-1418.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition, 47*, 75–85.

Kahan, D. M. (2016). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. *In Emerging trends in the social and behavioral sciences*, 1-16.

Kim, J. W. (2018). Evidence Can Change Partisan Minds: Rethinking the Bounds of Motivated Reasoning. *Working Paper*. Retrieved from https://jinwookimqssdotcom.files.wordpress.com/2018/09/jmp-sep-30.pdf

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211-228.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*, 28-55.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480-498.

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*, 2098-2109.

Lyons, B. A. (2016). *Unbiasing information search and processing through personal and social identity mechanisms*. PhD Dissertation: Southern Illinois University. https://opensiuc.lib.siu.edu/dissertations/1248/

Mandelbaum, E. (2018). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language*, 1-17.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175-220.

Nyhan, B., Porter, E., Reifler, J., & Wood, T. (2017). Taking corrections literally but not seriously? The effects of information on factual beliefs and candidate favorability. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2995128

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303-330.

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, *33*, 459-464.

Nyhan, B., & Reifler, J. (2018). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, 1-23.

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: a randomized trial. *Pediatrics*, *133*, e835.

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical Care*, *51*, 127-132.

Pennycook, G., & Rand, D. G. (2017). The implied truth effect: attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384

Schaffner, B. F., & Roche, C. (2017). Misinformation and motivated reasoning: Responses to economic news in a politicized environment. *Public Opinion Quarterly*, *81*, 86-110.

Sharot, T., & Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in Cognitive Sciences*, *20*, 25-33.

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. (2017). Processing political misinformation: comprehending the Trump phenomenon. *Royal Society Open Science*, *4*, 160802.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*, 755-769.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2018). Rethinking the link between cognitive sophistication and identity-protective cognition in political belief formation. Available at PsyArXiv https://psyarxiv.com/yuzfj/

Williams, D. (2018). Hierarchical Bayesian models of delusion. *Consciousness and Cognition*, *61*, 129-147.

Wood, T., & Porter, E. (2018). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 1-29.

Zhou, J. (2016). Boomerangs versus javelins: how polarization constrains communication on climate change. *Environmental Politics*, *25*, 788-811.