Routledge
Taylor & Francis Group

Philosophical Psychology

ISSN: 0951-5089 (Print) 1465-394X (Online) Journal homepage: www.tandfonline.com/journals/cphp20

Bayes or Pascal? The computations underlying motivated reasoning

Francesco Rigoli

To cite this article: Francesco Rigoli (19 Jan 2026): Bayes or Pascal? The computations underlying motivated reasoning, Philosophical Psychology, DOI: 10.1080/09515089.2026.2615690

To link to this article: https://doi.org/10.1080/09515089.2026.2615690

Published online: 19 Jan 2026.

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2026.2615690

Routledge
Taylor & Francis Group

Bayes or Pascal? The computations underlying motivated reasoning

Francesco Rigoli

Department of Psychology and Neuroscience, City St George's, University of London, London, UK

ABSTRACT

The construct of motivated reasoning has inspired an influential body of research. However, most theories of this construct are expressed in a verbal form. This is somewhat limited in light of contemporary research in cognitive science that emphasizes the insight afforded by employing computational modeling. To address this, the paper introduces a computational model of motivated reasoning. The model builds on previous accounts of belief formation based on Bayesian inference by adding computations concerning value or utility. The result is an interpretation of motivated reasoning as being akin to a process reflecting an unconscious Bayesian decision, in a way that is reminiscent of the famous Pascal's wager. This framework is broadly consistent with empirical evidence, especially about the effect of loss function asymmetries on probability judgments, about the confirmation bias, and about the backfire effect. Moreover, it is compatible with evolutionary explanations of motivated reasoning that interpret this phenomenon as ensuing from self-deception. The model helps understanding the computational principles behind the concept of motivated reasoning. Moreover, it facilitates the comparison between perspectives that downplay motivated reasoning and theories that emphasize its role. This may inform empirical research aimed at establishing the real contribution of motivated reasoning during belief formation.

ARTICLE HISTORY
Received 12 June 2025
Accepted 5 January 2026

KEYWORDS
Bayesian decision; motivated reasoning; probability judgment; confirmation bias; backfire effect; self-deception

Introduction

"I believe that it is raining outside." Beliefs like this are part and parcel of everyday life. How are such beliefs formed? This is one of the central questions within social and cognitive science research: addressing it can elucidate phenomena that are important in contemporary society such as, among others, propaganda, polarization, and resistance to scientific communication. It is widely assumed that, if a person entertains a given belief,

CONTACT Francesco Rigoli francesco.rigoli@citystgeorges.ac.uk Department of Psychology and Neuroscience, City St George's, University of London, Northampton Square, London EC1V 0HB, UK

she must have observed some supporting evidence. For example, the person's claim that it is raining may arise because she heard a thunder, or because she spotted a lightning. Moreover, it is evident that not only beliefs are grounded upon evidence, but also upon prior expectations. For example, the person may presuppose that rain is virtually impossible during the dry season. The thunder and the lightning notwithstanding, this prior knowledge may lead the person to conclude that, after all, the rain remains very unlikely.

The notion that novel evidence and prior expectations are the key ingredients of belief formation informs a venerable tradition in philosophy and in the social sciences. Authors like Plato, Aristotle, Locke, Hume, and Kant have all advanced versions of this idea (Niiniluoto et al., 2004). This view is also central to rational inference theory, perhaps the major framework in contemporary cognitive science research (e.g., Austerweil & Griffiths, 2011; Dasgupta et al., 2020; Griffiths & Tenenbaum, 2006; Hahn & Oaksford, 2007; Jern et al., 2014; Oaksford & Chater, 2007; Tenenbaum et al., 2006; Vul et al., 2014; Zmigrod, 2022; Zmigrod et al., 2023), as well as to proposals emphasizing the role of heuristics during belief formation (e.g., Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier, 2011; Gilovich et al., 2002; Kahneman et al., 1982). While rational inference theories entail that prior expectations and novel evidence are combined according to optimal principles (e.g., based on Bayesian statistics), theories based on heuristics assert that they are integrated according to rules of thumb. Yet, the two groups of theories agree that prior beliefs and novel evidence are the unique ingredients of belief formation. Even accounts positing that beliefs are distorted by systematic cognitive biases (e.g., Evans, 1989; Kahneman et al., 1982) ultimately share this view: according to them, representations of prior expectations and novel evidence are deformed by reasoning fallacies, but nonetheless remain the sole factors involved.

Not all authors, however, have been persuaded by the intuition that prior knowledge and novel evidence suffice to explain belief formation. An alternative idea is that, together with these two factors, people's motivation plays a key role too. For example, consider a person who craves a walk outdoor, thus hoping that it is not raining outside. This motivation, according to this perspective, will bias the person's judgment about whether it is raining or not. Such motivated reasoning outlook can already be identified at the dawn of Western philosophy as the position advocated by the sophists in Plato's dialogs (Barrett, 1987). Marginalized thereafter, it resurfaced at the end of the Nineteenth century in the work of authors referred by the philosopher Ricoeur as *masters of suspicion*, that is, in the work of Marx, Nietzsche, and Freud (Ricoeur, 1974). In their footsteps, several social scientists are nowadays persuaded by the idea that motivation is pivotal during belief formation. Indeed, this idea is the backbone of some of the most prominent

frameworks in social psychology, including cognitive dissonance theory (Festinger, 1957; Harmon-Jones & Harmon-Jones, 2007), attribution theory (Harvey & Weary, 1984; Heider, 1958), and social identity theory (Hogg, 2016; Tajfel & Turner, 1986).

The debate about whether motivation contributes to belief formation is far from settled, and many remain skeptical (e.g., Bayes & Druckman, 2021; Erdelyi, 1974; Hahn & Harris, 2014; Mercier & Sperber, 2017). Yet, evidence suggesting that this contribution is real is compelling (e.g., Ditto & Lopez, 1992; Ditto et al., 1998, 2003; Druckman, 2015; Drummond & Fischhoff, 2017; Kunda, 1990; Kahan, 2013, 2016; Klaczynski, 2000; Lodge & Taber, 2013; Lord et al., 1979; Taber & Lodge, 2006; Tappin et al., 2017). A recent example concerns judgments about climate change. If prior beliefs and novel evidence are the whole story, a person should estimate the likelihood of climate change based on aspects such as prior knowledge about climate, recent weather events, and analyses provided by experts on the media. By contrast, people's judgments on climate change appear, according to empirical research, to be shaped primarily by values and group affiliation, aspects possibly under the influence of motivational factors (Feldman et al., 2014; Kahan, 2015; Kobayashi, 2018; Ma et al., 2019; Linden et al., 2018; Wolsko et al., 2016; but see; Bayes & Druckman, 2021).

The contemporary scientific literature presents several theoretical treatments of motivated reasoning (e.g., Ditto, 2009; Ditto et al., 2009; Festinger, 1957; Harmon-Jones & Harmon-Jones, 2007; Harvey & Weary, 1984; Heider, 1958; Hogg, 2016; Jost et al., 2004, 2022; Kahan, 2016; Kunda, 1990; Lodge & Taber, 2013; Tajfel & Turner, 1986). These have offered great insight, yet a potential shortcoming is that, with rare exceptions (e.g., Kim et al., 2010), they are articulated in a verbal form. By contrast, various theories of belief formation that ignore motivated reasoning are grounded on computational modeling, for example in the form of Bayesian inference (Austerweil & Griffiths, 2011; Dasgupta et al., 2020; Griffiths & Tenenbaum, 2006; Hahn & Oaksford, 2007; Jern et al., 2014; Oaksford & Chater, 2007; Tenenbaum et al., 2006; Vul et al., 2014). Compared to computational models, verbal theories are not as precise and, thus, provide somewhat hazier descriptions and more ambiguous predictions. At present, a computational model of motivated reasoning is lacking, and therefore it remains to be established whether this construct can be spelled out in computational terms. What are the computational principles that underly motivated reasoning? The purpose of the present paper is to address this question by developing a computational framework of this construct and by assessing any novel insight that can be gained.

As noted above, at the empirical level there is ongoing debate on whether, and to what extent, motivation is important during belief formation. The paper aims at contributing to this debate by helping to clarify the construct

of motivated reasoning. Put another way, the aim of the paper is theoretical, and not empirical; it is not to establish whether the empirical literature warrants the conclusion that motivation is important during belief formation, but to help elucidating the theoretical principles behind motivated reasoning. Thus, the paper will not enter the empirical debate in detail. Still, it is important to consult the data to assess whether the proposed model is at least plausible, and the paper will do that when appropriate. Given that, as we shall see, the model relies on the formalism of Bayesian decision (Bishop, 2006), it is referred as *Bayesian decision model of motivated reasoning* (BDMR).[1] The next section will offer a systematic description of this framework.

The model

Before presenting the BDMR, it is paramount to provide the definition of two fundamental concepts: the concept of *belief* and the concept of *reasoning*. The article defines the former as a *conscious* mental state associated with acceptance of a specific hypothesis (e.g., "it is raining outside") with some degree of *confidence*. This definition assumes that a belief is a conscious mental representation. This excludes looser formulations of the term that encompass unconscious representations and that are found sometimes in the literature. For example, according to a looser definition, a giraffe can be described as believing that the food is on the tree. Though this way of describing things is useful in certain contexts, here the term "belief" is restricted to conscious mental states which, for example, do not apply to non-human animals. Note also that the definition just proposed introduces the concept of confidence. This implies that, even if two people agree that it is raining outside, one may be staunchly convinced about this while the other may express some degree of uncertainty.

Regarding the concept of *reasoning*, the contemporary literature distinguishes between *intuitive* and *deliberative* processes, the former being largely unconscious and the latter being predominantly conscious (Evans, 2003; Evans, 2008; Kahneman, 2011). Though both contribute to shape beliefs, there is growing evidence that intuitive reasoning is preponderant and is employed by default, while deliberative reasoning comes into play only occasionally and at a later stage (Evans, 2003, 2008; Kahneman, 2011). Based on this distinction, the focus of the BDMR is on explaining how intuitive mechanisms shape beliefs. Thus, all in all, the scenario explored by the model is one where a set of intuitive processes are at work at an unconscious level and eventually give raise to subjective beliefs (each embraced with a given level of confidence) expressed at the conscious level.

Now that the basic concepts have been spelled out, let us introduce the BDMR. The model incorporates previous proposals that rely on Bayesian

inference to explain how, during belief formation, prior representations are integrated with novel information (Austerweil & Griffiths, 2011; Dasgupta et al., 2020; Griffiths & Tenenbaum, 2006; Hahn & Oaksford, 2007; Jern et al., 2014; Oaksford & Chater, 2007; Tenenbaum et al., 2006; Vul et al., 2014). However, since these previous proposals ignore motivational influences, the model views them as limited. To account for the role of motivation, the BDMR integrates Bayesian inference with computations concerning value (or utility).[2] The result is a proposal which is based on a Bayesian decision framework (Bishop, 2006) where beliefs are formed over two stages, the first implementing Bayesian inference and the second capturing the influence of value (Figure 1). To illustrate how this works, consider a person who is pondering two alternative hypotheses about climate change. The first (a climate change hypothesis, H = CHA) claims that the climate has been changing dramatically. The second hypothesis (a climate hoax hypothesis, H = HOA) denies this claim. The BDMR analyses the unconscious reasoning mechanisms that lead a person to believe that either hypothesis is true and describes these mechanisms as being akin to a Bayesian decision process. This can be illustrated as unfolding over two stages. During stage one, the posterior probability of each hypothesis is estimated based on Bayesian inference. If P(H) corresponds to the prior probability of the two hypotheses (with P(CHA) = x, P(HOA) = 1-x, and $0 < x < 1$) and O indicates a new observation made, then, applying the Bayesian Theorem, the posterior probability can be written as:

$$P(H|O) = \frac{P(H)P(O|H)}{P(O)}$$

The prior probability P (H) describes the person's opinion about climate change before any new information is provided (Figure 2).[3] This may derive from past experience (e.g., weather events occurred in the past) as well as from general views about the world (e.g., the view that nature is unchanging). The quantity P(O|H) captures the influence of recording a new observation, for example in the form of experiencing a new weather event

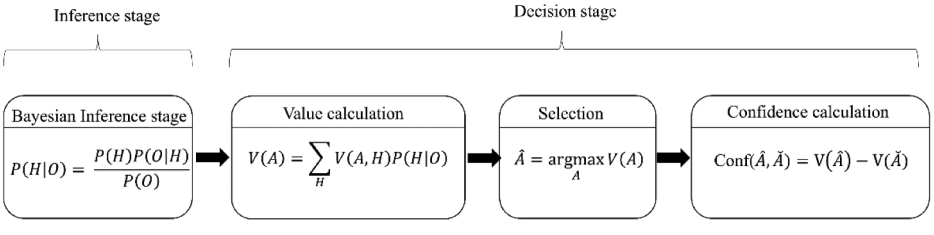

Figure 1. The Bayesian Decision model of Motivated Reasoning (BDMR).

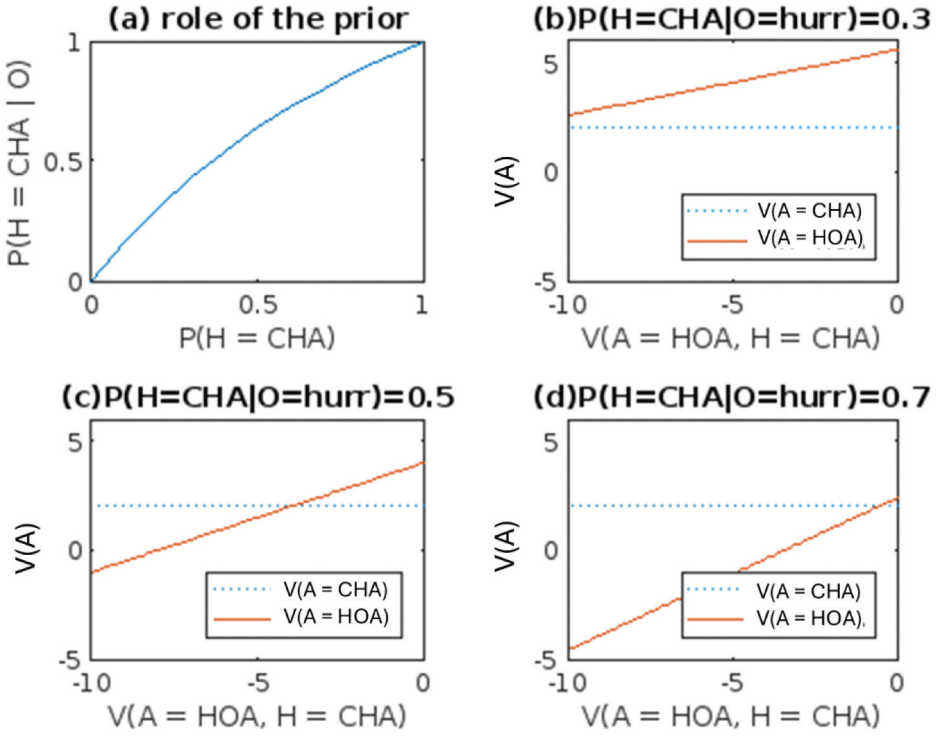

Figure 2. Role of the prior probability and of the value function within the BDMR. Note. The simulated scenario is one where an actor arbitrates between the climate change hypothesis (CHA) and the climate hoax hypothesis (HOA) based on an observation concerning whether a major hurricane has occurred in the country during the last year (O = hurr) or not (O = Nohurr). **(a)** This panel indicates that, as the prior probability P(H = CHA) grows, the posterior P (H = CHA | O) also grows. Here the parameters are as follows: O = hurr; P(O = hurr | H = CHA) = 0.7; P(O = hurr | H = HOA) = 0.4. **(b)** This panel indicates that, as the value of accepting the hoax hypothesis when it is false (V(A = HOA, H = CHA)) grows, the overall value of accepting the hoax hypothesis (V(A = HOA) grows as well, while the value of accepting the climate change hypothesis (V(A = CHA)) remains constant. Here the parameters are as follows: V(A = CHA, H = CHA) = 2; V(A = HOA, H = HOA) = 8; V(A = CHA, H = HOA) = 2; P(H = CHA | O = hurr) = 0.3. **(c)** Same as b, but with P(H = CHA | O = hurr) = 0.5. **(d)** Same as b, but with P(H = CHA | O = hurr) = 0.7.

or of reading a book on the issue. It reflects the probability of experiencing the new observation conditional on either hypothesis being true. For example, P(O = hurricane | H = CHA) indicates the probability of experiencing a hurricane assuming that the climate change hypothesis is true.

So far, the BDMR is equivalent to classic models of beliefs grounded on Bayesian inference (Austerweil & Griffiths, 2011; Dasgupta et al., 2020; Griffiths & Tenenbaum, 2006; Hahn & Oaksford, 2007; Jern et al., 2014; Oaksford & Chater, 2007; Tenenbaum et al., 2006; Vul et al., 2014). However, the BDMR denies that this is the end of the story. According to it, Bayesian inference is

only part of a more complex chain of unconscious mechanisms that shape people's beliefs. Indeed, the BDMR proposes that, once the posterior P (H|O) is computed, stage two ensues, which implements the decision procedure. This is realized based on a value function V (A, H) that depends on which hypothesis is assessed (A) and on which hypothesis is actually true (H) (Figure 2). The value function summarizes all incentives a person predicts to experience by accepting any hypothesis if that hypothesis is true or false. In our example, the value function encompasses four cases:

- V(A = CHA, H = CHA), describing the value of accepting the climate change hypothesis when it is correct;
- V(A = CHA, H = HOA), describing the value of accepting the climate change hypothesis when it is wrong;
- V(A = HOA, H = HOA), describing the value of accepting the climate hoax hypothesis when it is correct;
- V(A = HOA, H = CHA), describing the value of accepting the climate hoax hypothesis when it is wrong.

To appreciate the role of the value function, compare different people while they are evaluating the climate change hypothesis. For a CEO of an oil multinational, the cost of accepting the climate change hypothesis, and of acting accordingly by forfeiting lucrative business opportunities, is very large if this hypothesis turns out to be false. For someone without as much at stake, by contrast, the cost of accepting the climate change hypothesis if this turns out to be false is not as large. This example highlights how the value function can vary greatly among people.

Based on the value function, the BDMR proposes that the expected value associated with accepting either hypothesis is derived. This corresponds to the value expected by choosing that hypothesis, independent of whether the latter turns out to be true or not. Formally, the expected value is calculated by multiplying, for each hypothesis, every value function and the associated posterior probability, and by summing across these products:

$$V(A) = \sum_{H} V(A,H)P(H|O)$$

To see how this works, consider a case where the value functions associated with selecting the climate change hypothesis are V(A = CHA, H = CHA) = 10 and V(A = CHA, H = HOA) = 2, and where the posterior probabilities are $P(H = CHA|O) = 0.2$ and $P(H = HOA|O) = 0.8$. Applying equation 2, the expected value associated with accepting the climate change hypothesis is V(A = CHA) = 10 ×0.2 + 2 ×0.8 = 3.6.

Once the expected values associated with the hypotheses are calculated, one of the hypotheses is finally selected. This, labeled as $\hat{A}$, is the one associated with the largest expected value, formally equal to:

$$\hat{A} = \underset{A}{\operatorname{argmax}}\, V(A)$$

For example, if the expected value associated with the climate change hypothesis is V(A = CHA) = 3.6 and the one associated with the climate hoax hypothesis is V(A = HOA) = 2, then the climate change hypothesis is the one that is chosen ($\hat{A} = CHA$). Crucially, the model posits that the chosen hypothesis ($\hat{A}$) corresponds to the conscious belief embraced by the person. Therefore, insofar as values play a pivotal role in determining which hypothesis is believed to be true, the BDMR interprets beliefs as the result of decisions, and not of inferences. Consider again the oil multinational CEO who attributes a large cost to accepting the climate change hypothesis if this turns out to be false. According to the model, this person is predicted to be more prone to believe in the climate hoax hypothesis compared to people who do not have as much at stake. In this way, the BDMR explains motivated reasoning.

One last element modeled by the BDMR is the level of confidence associated with a belief. The confidence Conf ($\hat{A}, \check{A}$) is equal to the difference in expected value between the chosen hypothesis ($\hat{A}$) and the hypothesis associated with the second-largest expected value ($\check{A}$):

$$Conf\,(\hat{A},\, \check{A}) = V(\hat{A}) - V(\check{A})$$

In the example above where V(A = CHA) = 3.6 and V(A = HOA) = 2, and thus where $\hat{A} = CHA$ and $\check{A} = HOA$, the confidence is equal to Conf ($\hat{A}, \check{A}$) = 3.6 – 2 = 1.6. It is evident that the level of confidence is boosted when the expected value of the selected hypothesis increases compared to the expected value of alternative hypotheses.

In short, the BDMR asserts that beliefs do not stem from inferences, but from unconscious decisions, and proposes a Bayesian framework to characterize the underlying processes. The inferential component remains important, as this is a necessary element of any informed decision. Yet, inference is only part of the story, as this is integrated with value representations to produce the final belief.

To further elucidate the functioning of the model, let us apply it to interpret the unconscious reasoning processes engaged by a person named Sally who has just watched a short interview of a climatologist. Since she lives in a region where the climate has been very regular for years, a priori Sally does not give much credit to the climate change hypothesis (this is

reflected in the prior probability P(H)). Yet, the interviewed climatologist, who is well-respected and well-known on the media, reports various statistics compatible with climate change. To some extent, this new information changes Sally's view, as reflected in a posterior probability which, compared to the prior probability, is more favorable to the climate change hypothesis (formally, P(H = CHA | O = interview) > P(H = CHA)). The inference stage is now completed, and the decision stage ensues. Unconsciously, Sally asks the following questions: What are the implications of accepting the climate change hypothesis if this is correct? And if it is wrong? And what are the implications of rejecting it if the hypothesis is correct? And if it is wrong? Sally's friends ridicule people who are concerned about climate change because they believe that the whole issue is in truth an excuse for allowing the state to control people's life. Sally deeply desires to be admired by her friends, and hence she attaches a very negative value to accepting the climate change hypothesis, even if this may turn out to be true. On this basis, at the end Sally consciously believes that the climate change hypothesis is false. Now, imagine an alternative scenario where Sally is extremely scared by the gloomy forecast about the consequences of climate change. On this basis, the cost of rejecting the climate change hypothesis if this turns out to be true now far outweighs the cost of accepting this hypothesis if this turns out to be false. In this alternative scenario, Sally ends up consciously believing that the climate change hypothesis is true.

An important aspect of the BDMR concerns the relative weight of the inferential component (captured by the Bayesian inference part) vis-à-vis the value component (captured by the decision-making part). The BDMR treats such weight flexibly by adapting the value function to the specific circumstances. For instance, consider a person who is highly motivated to be accurate. Such elevated *accuracy motive* can be modeled by boosting the value of accepting the hypotheses when the hypotheses are correct – in the example above, this corresponds to boosting both V(A = CHA, H = CHA) and V(A = HOA, H = HOA). This minimizes the influence of any bias in favor of either hypothesis, as reflected in a strong motivation to be accurate. Related to this point is the question of what happens when an extremely large payoff is linked with a clearly unrealistic option. Imagine a person who expects to receive one billion dollars if she believes that Putin is the president of the U.S.A.. Does the BDMR predict that, in scenarios like this, people will accept such unrealistic hypotheses? The BDMR does not make this prediction if one assumes that, following standard decision-making models (e.g., Samuelson, 1950; Stewart et al., 2006), values do not correspond to raw payoffs, but to payoffs transformed according to a concave function with an upper bound. If a bounded concave function is employed, then the values attributed to the different hypotheses will not go above a certain threshold. It follows that, whatever the payoffs associated

with the different hypotheses, their impact will be limited, implying that the inferential component will remain important and hence that hypotheses that are clearly unrealistic will be unlikely to be accepted.

All in all, one way to interpret the BDMR is to frame it as a modification of the famous Pascal's wager (Rota, 2017). The philosopher and mathematician Blaise Pascal, who was the father of probability theory, offered a clever argument to justify his Christian faith. He argued that the rational procedure to establish whether the Christian God exists is not limited to estimating the probability of God's existence, but it also requires assessing the utility of accepting the hypothesis that God exists. In other words, it requires answering the following questions: what is the utility of accepting the Christian faith if this turns out to be true? And if it turns out to be false? And what is the utility of rejecting the Christian faith if this turns out to be true? And if it turns out to be false? Pascal, who was a fervent Catholic, reasoned that accepting Christianity if this turns out to be true is linked with an infinite utility that corresponds to the beatitude experienced in heaven after death. On this basis, he concluded that the most rational decision for anyone is to believe in Christ. Leaving aside the religious theme, there are obvious analogies between the Pascal's wager just described and the BDMR.[4] The difference, though, is that, while Pascal conceptualized his wager as a conscious set of rules one should follow to establish what to believe, the BDMR claims that, by and large, these rules are followed already by people, though unconsciously. Put another way, Pascal intended his wager as a *prescription* about what people should believe. The BDMR raises the possibility that it may be more appropriate to view the wager as a *description* of what people already believe.[5]

Now that the BDMR has been illustrated, let us evaluate it in the context of the previous scientific literature. The reminder of the article will discuss various research strands that are relevant to assess the model. As I shall argue, this analysis highlights several aspects, both theoretical and empirical, that are broadly consistent with the logic proposed by the model.

Evolution

When assessing a new psychological model, one of the most important aspects to ascertain is its consistency with the theory of evolution. Let us address this with regard to the BDMR. One way to proceed is to look at contemporary evolutionary theories of motivated reasoning and to ask whether the picture they offer is consistent with the BDMR. One of the most influential of such theories is grounded upon noticing the general importance of deceptive strategies in the context of evolution (Butterworth et al., 2022; Smith et al., 2017; Trivers, 2011; Von Hippel & Trivers, 2011). Instances of organisms that employ deceptive signals

for survival or reproduction, such as the camouflage performed by stick insects to hide from predators, are countless in nature. The human species, according to this argument, is not immune to employing analogous strategies, especially when people communicate with each other. In evolutionary terms, deceiving other members of the social group may sometimes be adaptive. However, it can also be seriously risky if one is discovered by the other group members. To minimize this risk, the theory proposes that humans have evolved the tendency to deceive not only other people, but also themselves. While a conscious lie conveys a series of signals that enhance the risk of detection, these signals are suppressed if one lies also to oneself, thus minimizing the risk of being caught lying (Smith et al., 2017). Moreover, self-deception implies diminished accountability, and thus milder punishment in the case that the truth is discovered (Butterworth et al., 2022). If this evolutionary theory is correct, then motivated reasoning can be interpreted as ensuing from the human propensity for self-deception. According to this view, at an early stage of reasoning people look at the world in a relatively disinterested way, but, at some point, motivational factors come into play – all these processes would unfold unconsciously. This would result in biased beliefs that are in turn communicated to other people.

When the evolutionary theory just described and the BDMR are examined together, the conclusion is that the two are widely compatible. The BDMR can be viewed as providing a computational description of motivated reasoning, interpreted, in evolutionary terms, as a form of self-deception. The model argues that Bayesian inference is the mechanism whereby, as the evolutionary theory suggests, reality is processed in a disinterested way at the unconscious level. Thanks to the decision component, the model also offers a computational description of how the unconscious representations are influenced by value in a way that ultimately produces motivated reasoning and self-deception. Put it simply, the BDMR argues that, unconsciously, people select the hypothesis which is more convenient for them to hold (based on integrating inference and value); next, in a form of self-deception, they experience that belief as the one which is true; finally, in a form of deception toward others, they communicate the belief to other people and thus nudge them to perform certain favorable actions. Consider again the example of the CEO of a multinational oil company pondering the climate change hypothesis. This person may unconsciously acknowledge the plausibility of this hypothesis (via the Bayesian inference stage). Yet, because of the influence exerted by values (via the decision stage), she may consciously believe that the climate change hypothesis is unfounded, she may communicate this belief to other people,

and the latter may in turn behave accordingly (e.g., voting for parties who downplay the climate change issue).

Judgment about probability

Now that the BDMR has been discussed regarding the theory of evolution, let us assess the model in the context of empirical literature. With this regard, one of the most relevant bodies of work is research exploring the impact of motivation upon judgments about probability (Bar-Hillel & Budescu, 1995; Harris et al., 2009; Krizan & Windschitl, 2007; Weber, 1994). This research has investigated whether people's judgments about the probability of a hypothesis vary when the value associated with that hypothesis is manipulated. For example, does the estimated probability of contracting an illness depend on how much a person dreads the illness?

The research on this topic is particularly relevant here because different theories of motivated reasoning entail divergent predictions. Some predict an optimism bias (or wishful thinking) effect, namely, they predict that hypotheses which are more desirable are attributed higher likelihood (Armor & Taylor, 1998; Krizan & Windschitl, 2007; Sharot, 2012). Other theories imply the opposite, namely, a pessimism bias whereby less desirable hypotheses are judged as more likely (Baumeister et al., 2001; Krizan & Windschitl, 2007; Norem, 2001; Shepperd et al., 1996). One last theory is grounded on the notion of loss function asymmetry (Harris et al., 2009; Weber, 1994). The latter occurs when the cost of misjudgment across competing hypotheses is not equal. For example, the cost of dismissing the hypothesis that one has cancer is very high if the person indeed has cancer. At the same time, the cost associated with accepting the hypothesis that one has cancer is relatively small if one actually has no cancer. Because of this loss function asymmetry, the theory predicts that people will typically overestimate the risk of contracting cancer. Among the three theories just described, it is evident that the latter is the closest to the BDMR: both presuppose that loss function asymmetries have an impact upon judgments about probability - Figure 3 illustrates why the BDMR is consistent with loss function asymmetries. By contrast, as demonstrated by Figure 3, the BDMR does not envisage any optimism nor pessimism bias.

Does empirical research support theories of optimism bias, pessimism bias, or of loss function asymmetry? Recent reviews and metanalyses have concluded that evidence for optimism and pessimism bias is mixed: some studies have reported effects compatible with an optimism bias, others compatible a pessimism bias, and another group of studies with neither (Bar-Hillel & Budescu, 1995; Harris et al., 2009; Krizan & Windschitl, 2007).

Empirical data relevant to assess the effects of loss function asymmetries are less extensive, as for example they concern primarily the aversive, rather

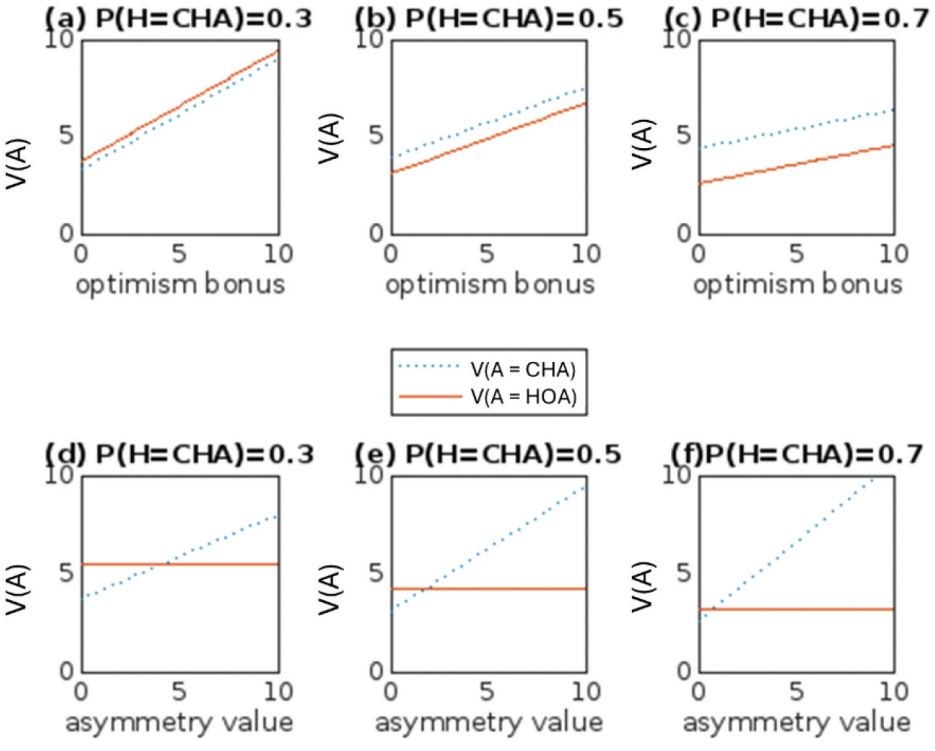

Figure 3. Predictions of the BDMR regarding optimism bias and loss function asymmetry. Note. The simulated scenario is one where an actor arbitrates between the climate change hypothesis (CHA) and the climate hoax hypothesis (HOA) based on an observation concerning whether a major hurricane has occurred in the country during the last year (O = hurr) or not (O = Nohurr). All simulations have the following parameters: O = hurr; P(O = hurr | H = CHA) = 0.7; P (O = hurr | H = HOA) = 0.4. **(a)** This panel assesses the BDMR's predictions in terms of any optimism bias. The scenario is one where an optimism bonus is added to the value of the preferred hypothesis (in this example, the hoax hypothesis) when the hypothesis is true, both in the case that the hypothesis is selected (V(A = HOA, H = HOA)) and in the case that it is not selected (V(A = CHA, H = HOA)). The panel shows that the difference in value between the two hypotheses (V(A = CHA) and (V(A = HOA) remains constant as the optimism bonus changes. This simulation shows that the BDMR does not predict any optimism bias. Here, parameters are as follows: P(H = CHA) = .3; V(A = CHA, H = CHA) = 5; V(A = CHA, H = HOA) = 2 + optimism bonus; V(A = HOA, H = HOA) = 5 + optimism bonus; V(A = HOA, H = CHA) = 2. **(b)** Same as a except for P(H = CHA) = .5. **(c)** Same as a except for P(H = CHA) = .5. **(d)** This panel assesses the BDMR's predictions in terms of any loss function asymmetry. The scenario is one where an asymmetry value is added to the value of the climate change hypothesis if this hypothesis is true and is correctly selected (V(A = CHA, H = CHA)). The panel shows that, as the asymmetry value grows, the value of the climate change hypothesis (V(A = CHA) grows as well, while the value of the climate hoax hypothesis (V(A = HOA) remains constant. This simulation shows that the BDMR predicts effects due to loss function asymmetry. Here, parameters are as follows: P(H = CHA) = .3, V(A = CHA, H = CHA) = 2 + asymmetry value; V(A = CHA, H = HOA) = 5; V(A = HOA, H = HOA) = 8; V(A = HOA, H = CHA) = 2. **(e)** Same as d except for P(H = CHA) = .5. **(f)** Same as d except for P(H = CHA) = .7.

than appetitive, domain (Harris et al., 2009; Weber, 1994). Yet, available evidence indicates that, when loss function asymmetries are manipulated experimentally, an effect on probability judgments emerges. In a study documenting this effect (Harris et al., 2009), participants were presented with a vignette describing an orchard with apple trees, where some of the fruits were poisonous to the point of provoking death. Participants were informed that a child often walked within the orchard and, despite the father's prohibition to eat the apples, picked some and eat them. The task consisted in estimating the probability that the child picked a poisonous apple. Two conditions were compared: a no-control condition, where the father was described as being unable to block the child's behavior; and a high-control condition, where, based on listening to the participant's judgment, the father could decide to erect a fence and thus protect the child. The analyses revealed that participants viewed the possibility that the child eat the poisonous apples as more probable during the high-control compared to the no-control condition. According to an explanation based on loss function asymmetry (Harris et al., 2009), the effect emerges because the cost of rejecting the hypothesis that the child will eat the poisonous apple if the hypothesis is true is negligible in the no-control condition. This is because, in this condition, nothing can be done to prevent the accident. By contrast, the same cost surges in the high-control condition, thus inflating the estimated probability.

In short, research on people's judgments about probability suggests that, while no clear evidence of optimism nor pessimism bias emerges, asymmetries in the loss function appear to count. This is broadly consistent with the computational mechanisms advocated by the BDMR to explain motivated reasoning. Notably, theories that neglect motivated reasoning struggle to explain data documenting effects due to loss function asymmetries. These effects may represent the most compelling evidence in favor of the hypothesis that motivational factors are indeed critical during belief formation.

The empirical literature just reviewed is also relevant for a crucial theoretical point which I shall consider now. The key intuition behind the BDMR is that a computational account of motivated reasoning requires integrating Bayesian inference with value calculations, something the model does by employing the formalism of Bayesian decision. At the outset, however, an alternative model may be advanced to integrate Bayesian inference and value calculations, a model where values are captured by manipulations of the prior probability. It has been suggested that, when a hypothesis is more appealing, the brain may attribute higher prior probability to this hypothesis, resulting in an optimism bias (Kahan, 2016). For instance, the case of a person who desires the climate hoax hypothesis to be true (e.g., the CEO of the oil multinational) may be modeled by boosting the

prior probability associated with that hypothesis. Such biased prior may in turn be employed during Bayesian inference, and the hypothesis with the largest posterior probability (technically, the maximum a posteriori) may end up being the person's conscious belief. Note that, in this alternative model, there is no decision stage involved, but only Bayesian inference. This is because the value component is already incorporated within the prior probability, and thus no subsequent decision is required.

This alternative model is, at the outset, a plausible candidate for explaining motivated reasoning. Is it preferable to the BDMR? To answer this question, consider that the BDMR and the alternative model make distinct predictions regarding judgments about probability. Specifically, the alternative model implies an optimism bias: the favored hypothesis is granted higher prior probability, meaning that one's belief will be biased toward that hypothesis. If one prefers the climate hoax hypothesis, for example, this hypothesis will be attributed a disproportionate posterior probability, resulting in a bias in favor of it. By contrast, as discussed above, the BDMR implies an effect linked with loss function asymmetries. What counts in this model is not which hypothesis is preferred, but what is the value expected by accepting/rejecting the hypothesis if the hypothesis is true/false. As documented above, empirical evidence in support of an optimism bias is controversial (Bar-Hillel & Budescu, 1995; Harris et al., 2009; Krizan & Windschitl, 2007), while evidence of effects due to loss function asymmetries is more consistent (Harris et al., 2009; Weber, 1994). On this basis, the BDMR appears to be a more promising computational model of motivated reasoning compared to an alternative model where the desirability of a hypothesis is modeled by a biased prior probability.

Incoming information

Another empirical domain that should be analyzed with regard to the BDMR pertains the influence of incoming information. It is widely assumed that people's beliefs are shaped by the information detected in the environment. Computational models that disregard motivated reasoning explain the influence of incoming information by assuming that this provides new observations during inference (e.g., Austerweil & Griffiths, 2011; Dasgupta et al., 2020; Griffiths & Tenenbaum, 2006; Hahn & Oaksford, 2007; Jern et al., 2014; Oaksford & Chater, 2007; Tenenbaum et al., 2006; Vul et al., 2014). What is the picture offered by the BDMR on this matter? The present section examines how, according to the BDMR, incoming information influences beliefs and discusses the ensuing predictions in light of empirical data.

In essence, the BDMR postulates two different pathways whereby new information can influence beliefs. Most obviously, and in

agreement with Bayesian inference theories (Oaksford & Chater, 2007; Tenenbaum et al., 2006), incoming information can provide an observation (O) that can be employed to estimate the posterior probability (P(H|O)) during the inference stage. The second pathway is more subtle, but no less important: it concerns information relevant for the value function. To understand what this means, consider a scenario where a person is talking about climate change with a dear friend, at the same time desiring strongly to be admired by the friend. During the conversation, the person realizes that the friend scorns people who are skeptical about climate change. This information may affect the person's value function in such a way that accepting the climate change hypothesis is now viewed as highly rewarding, thus nudging the person toward that hypothesis. Note that information relevant for the value function may have no impact during the Bayesian inference stage. For example, the person may deem the friend to be ignorant on the matter, implying that the friend's opinion has little to no influence during the Bayesian inference stage. Yet, the motivation of securing the friend's admiration may be strong, implying a substantial influence upon the value function and, ultimately, upon beliefs.

The formula presented above describes how new information (O) is used to calculate the posterior probability (P(H|O)) within the Bayesian inference component. For completeness, here I propose an analogous formula to describe how new information influences the calculation of the value function. This is based on a simple Rescorla-Wagner rule (Sutton & Barto, 1998):

$$V(A,H) = V_{initial}(A,H) + \alpha(V_{observed}(A,H) - V_{initial}(A,H))$$

Here $V_{initial}(A,H)$ describes the value function before an event (the conversation with a friend), $V_{observed}(A,H)$ reflects the value function linked with the event, and α indicates a learning rate bounded between zero and one.

In short, the BDMR proposes that incoming information can affect beliefs in two distinct ways: by providing new observations during the Bayesian inference stage and by shaping the value function. In its essence, this picture is no different from the one offered by previous theories of motivated reasoning (e.g., Ditto, 2009; Ditto et al., 2009; Festinger, 1957; Harmon-Jones & Harmon-Jones, 2007; Harvey & Weary, 1984; Heider, 1958; Hogg, 2016; Jost et al., 2004, 2022; Kahan, 2016; Kunda, 1990; Lodge & Taber, 2013; Tajfel & Turner, 1986). Thanks to the employment of a computational framework, the BDMR may contribute to this literature by elucidating the nature of the two distinct effects exerted by incoming information. Below, I shall examine some of the key predictions ensuing

from this view and assess them in the context of two relevant empirical phenomena: the confirmation bias and the backfire effect.

Confirmation bias

The empirical literature indicates that, when presented with new evidence, people tend to update their beliefs differently based on whether the evidence is consistent with their current beliefs or not (e.g., Bronfman et al., 2015; Klayman & Ha, 1987; Mercier, 2017; Mercier & Sperber, 2017; Nickerson, 1998; Talluri et al., 2018). Specifically, the degree of update appears to be stronger when new evidence is consistent with one's beliefs compared to when it is not, a phenomenon referred as confirmation bias. Theories ignoring motivated reasoning have advanced clever explanations of this bias. A compelling one grounded on Bayesian inference posits that, when a new piece of evidence is observed, not only an actor estimates the posterior probability of a hypothesis, but, at the same time, also the reliability of the source (Christensen, 2023; Gentzkow & Shapiro, 2006; Hahn & Harris, 2014; Merdes et al., 2021; Pilgrim et al., 2024). For example, a person reading a report about climate change from an expert may employ this information both to assess the validity of the climate change hypothesis and to establish whether the expert is credible or not. These models explain the confirmation bias because, when the evidence is consistent with prior expectations, the source is attributed higher reliability and thus the evidence is weighted more. By contrast, evidence inconsistent with prior expectations leads to view the source as less reliable, and thus to weight evidence less. Since the BDMR encompasses a Bayesian inference stage, the explanation of confirmation bias just outlined is fully compatible with the model (of course, provided that the model is extended by implementing a more sophisticated inference stage). Nevertheless, in keeping with previous theories of motivated reasoning, the BDMR argues that this and similar explanations are only partial. To understand the phenomenon in all its complexity, the BDMR maintains that the role of motivation and values needs also to be considered.

The literature on motivated reasoning has pinpointed various motivational factors that could be responsible for the confirmation bias. For example, in the political and public opinion realm, people may desire to identify themselves with a specific party (e.g., the conservative party) (Kahan, 2016; Strickland et al., 2011). This, in turn, may motivate people to downplay information that conflicts with the party's stance (e.g., information signaling that immigration boosts the gross domestic product) and to overstress information that aligns with the party's position (e.g., information that signals that immigration increases unemployment). This would

result in a form of confirmation bias which derives from motivated reasoning, in a way that aligns with the BDMR.

Motivated reasoning may produce the confirmation bias also because of people's desire to exhibit confidence. Empirical observations show that, if a person exhibits confidence in her own beliefs, then other people are more likely to be persuaded by the person (Belmi et al., 2020; Murphy et al., 2015; Ronay et al., 2019; Schwardmann & Van der Weele, 2019). On this basis, evolution might have endowed people with an inbuilt motivation for displaying confidence in their beliefs (Butterworth et al., 2022; Trivers, 2011; Von Hippel & Trivers, 2011), even if this typically leads to overconfidence, namely, to the conviction that one's own beliefs are more accurate than they really are (Alicke & Sedikides, 2009; Moore & Healy, 2008). The implication of the desire for displaying confidence is that changing mind may be costly for humans (Acharya et al., 2018; Mullainathan & Shleifer, 2005), because changing mind signals poor confidence. Following the BDMR, it can be suggested that, unconsciously, people consider the cost of changing their mind during the decision stage of belief formation. The implication is that, to avoid this cost, people may downplay evidence that is inconsistent with their current beliefs, therefore exhibiting a confirmation bias.

Formally, the BDMR can implement the motivational factors eliciting the confirmation bias by adding a confirmation bonus (R_{conf}) to the values associated with acceptance of the hypothesis previously endorsed (Figure 4). To illustrate how this works, compare two persons named Jack and James, who both have hitherto expressed a belief in the climate change hypothesis. Assume that Jack does not manifest any confirmation bias while James does. This can be implemented by assuming that Jack's values are V(A = CHA, H = CHA) = 5, V(A = CHA, H = HOA) = 2, V(A = HOA, H = HOA) = 5, and V(A = HOA, H = CHA) = 2, while James' values are equivalent to Jack's except that V(A = CHA, H = CHA) = 5 + R_{conf} and V(A = CHA, H = HOA) = 2 + R_{conf} (with R_{conf} being a positive number). Note that the conformation bonus (R_{conf}) applies to the values associated with accepting the hypothesis hitherto endorsed. The consequence of implementing such bonus is that, when a new observation is recorded, Jack will not show any confirmation bias, while James will.

The empirical literature has highlighted a confirmation bias not only when people *receive* new information, but also when people *seek* new information. Such bias is expressed in a general tendency to seek evidence that confirms one's current beliefs, rather than evidence contrary to them. This has been documented at various levels, for example as a predisposition to produce arguments or to remember events consistent with current beliefs, or as an inclination to attend to sources that are more likely to

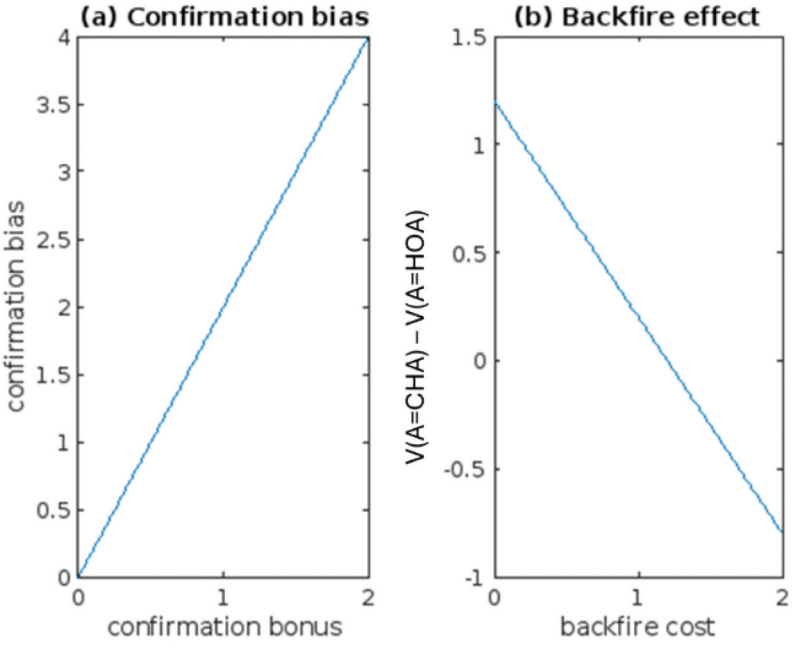

Figure 4. Implementation of the confirmation bias and of the backfire effect according to the BDMR. Note. The simulated scenario is one where an actor arbitrates between the climate change hypothesis (CHA) and the climate hoax hypothesis (HOA) based on an observation concerning whether a major hurricane has occurred in the country during the last year (O = hurr) or not (O = Nohurr). All simulations have the following parameters: P(O = hurr | CHA) = 0.7; P(O = hurr | H = HOA) = 0.3. **(a)** This panel assesses the BDMR's predictions in terms of any confirmation bias. Here two agents are simulated, a baseline agent and a biased agent. The biased agent is one who, in the past, has committed in favour of the hoax hypothesis. As a consequence, the biased agent adds a confirmation bonus to the value of selecting the hoax hypothesis both if this hypothesis is correct (V(A = HOA, H = HOA)) and if this hypothesis is incorrect (V(A = HOA, H = CHA)). Parameters are as follows. For both agents, P(H = CHA) = 0.3. The baseline agent has V(A = CHA, H = CHA) = 5; V(A = CHA, H = HOA) = 2; V(A = HOA, H = HOA) = 5; V(A = HOA, H = CHA) = 2; The biased agent has V(A = CHA, H = CHA) = 5; V(A = CHA, H = HOA) = 2; V(A = HOA, H = HOA) = 5 + confirmation bonus; V(A = HOA, H = CHA) = 2 + confirmation bonus. For each agent, I calculated the quantity Q = [V(A = HOA | O = Nohurr) – V(A = CHA | O = Nohurr)] - [V(A = CHA | O = hurr) – V(A = HOA | O = hurr)]. Intuitively, Q indicates how much an agent prefers the hoax hypothesis compared to the climate change hypothesis across different observations O. The confirmation bias is equal to the Q for the biased agent minus the Q for the baseline agent. The graph shows that, as the confirmation bonus grows, the confirmation bias grows as well. **(b)** This panel assesses the BDMR's predictions in terms of any backfire effect. In this scenario, an agent has been exposed to a message from a disliked source who supports the climate change hypothesis. This results in subtracting a backfire cost to the value of selecting the climate change hypothesis both if the hypothesis is true (V(A = CHA, H = CHA)) and if the hypothesis is false (V(A = CHA, H = HOA)). Here, parameters are as follows: P(H = CHA) = 0.5; O = hurr; V(A = CHA, H = CHA) = 5 - backfire cost; V(A = CHA, H = HOA) = 2 - backfire cost; V(A = HOA, H = HOA) = 5; V(A = HOA, H = CHA) = 2. The graph shows that, as the backfire cost grows, the value of selecting the climate change hypothesis (V(A = CHA)) diminishes in comparison to the value of selecting the climate hoax hypothesis V(A = HOA).

provide information consistent with one's current beliefs (Eagly et al., 1999; Mercier & Sperber, 2017; Vedejová & Čavojová, 2022). As above, the BDMR implies that inferential and motivational factors may both be critical during information seeking. At the inferential level, for instance, the tendency to seek confirmatory evidence may derive from a propensity to attend to sources deemed to be more reliable (e.g., a right-wing supporter may believe that right-wing politicians are more reliable, therefore seeking information from them). At the motivational level, this tendency may arise from the same desires highlighted above, that is, the desire to identify oneself with the group (e.g., by seeking confirmatory evidence, a person can better justify the beliefs shared with the group) and the desire to display confidence (e.g., by seeking confirmatory evidence, a person can better justify her views and thus display confidence).

In short, in keeping with previous motivated reasoning interpretations of the confirmation bias, the BDMR asserts that this bias can arise both from inferential and motivational factors. Whether this prediction is correct remains to be established empirically, as there is debate about whether inferential mechanisms are sufficient to explain all manifestations of the confirmation bias (Christensen, 2023; Gentzkow & Shapiro, 2006; Hahn & Harris, 2014; Merdes et al., 2021; Pilgrim et al., 2024) or whether motivational factors are also important. The BDMR may contribute to this debate by offering a precise computational definition of the motivational processes potentially involved.

Backfire effect

The backfire effect occurs when, after being presented with a message opposing their initial beliefs, people paradoxically increase the confidence on their initial beliefs (Nyhan & Reifler, 2010; Swire-Thompson et al., 2020). For example, after watching a climatologist's interview warning about the risks of climate change, a person initially skeptical about climate change may express even greater skepticism on the matter. Empirical research on the backfire effect has overall produced mixed findings, and there is ongoing debate on how reliable, general, and strong the effect is, with some scholars even questioning its very existence (Nyhan, 2021; Nyhan & Reifler, 2010; Swire-Thompson et al., 2020; Wood & Porter, 2019). Yet, notwithstanding these pertinent calls for caution, the possibility that, at least in a limited number of circumstances, the backfire effect occurs remains open.

The present paper does not aim to evaluate the empirical literature about the backfire effect. Rather, it aims to propose a theoretical interpretation thereof that may help future empirical research. According to models of belief formation that ignore motivated reasoning, no backfire effect is conceivable. This is because, according to this perspective, beliefs should

always move in the direction of new messages, or, at least, remain unmodified; changes in the direction opposite to new messages is hard to explain. On this basis, virtually all previous treatments of the backfire effect have attributed a critical role to motivational factors (Nyhan & Reifler, 2010; Swire-Thompson et al., 2020). Building on these treatments, here I shall analyze how the BDMR explains this effect. To illustrate the argument, consider a conversation about climate change among a group of acquaintances, two of whom named Tom and Jane. Tom, who is initially mildly skeptical about climate change, deeply hates Jane. Hearing Jane explaining why climate change is at present the greatest threat for humanity may lead Tom to attach a large cost to accepting the climate change hypothesis ("I hate agreeing with Jane!"). In turn, this may have the paradoxical effect that, spurred by the desire to disagree with Jane, Tom now moves from a mild to a radical skepticism about climate change. In this example, Jane's message is uninfluential during the inference stage (imagine that Tom does not see Jane as a reliable source), but it affects dramatically the value function. As a consequence, the cost of accepting the climate change hypothesis surges, with the effect of radicalizing Tom's initial position.

A similar logic may be applied to interpret empirical data about the backfire effect. For example, in a study conducted in the U.S.A. (Bail et al., 2018), for one month liberals and conservatives were exposed on social media to messages posted by opinion leaders of the opposite ideological side. After such exposure, participants (especially conservatives) reported to embrace their initial ideology even more strongly. Applying the logic outlined above, it is possible that, in this study, participants did not consider the opinion leaders of the opposite side as reliable, thus ignoring them during the Bayesian inference stage. At the same time, given a desire to identify themselves with a specific ideological stance, participants may have attached a negative value to the possibility of agreeing with the messages propagated by the opposite ideological side. This may have resulted in an even greater opposition to such messages as manifested in an increased radicalization, consistent with a backfire effect.

At the formal level, the BDMR can implement the backfire effect in a way similar to the confirmation bias: in this case, a backfire cost (R_{back}) can be subtracted to the value function associated with accepting the hypothesis advocated by the disliked source (Figure 4). To illustrate how this works, compare two persons named Mary and Sarah who have just heard a speech from a U.S.A. liberal politician warning about the risks of climate change. Imagine that this speech has aroused in Mary, but not in Sarah, a strong motivation to identify herself as conservative in opposition to liberals. For Sarah, the value functions are V(A = CHA, H = CHA) = 5, V(A = CHA, H = HOA) = 2, V(A = HOA, H = HOA) = 5, and V(A = HOA, H = CHA) = 2. Given her desire to oppose liberal ideas, Mary's value functions are

equivalent to Sarah's except that now V(A = CHA, H = CHA) = 5 - R_{back} and V(A = CHA, H = HOA) = 2 - R_{back} (with R_{back} being a positive number). The implication is that, in response to the speech of the liberal politician worrying about climate change, Sarah's belief will move toward the climate change hypothesis (thanks to the inference stage), but Mary's belief will move away from that hypothesis (thanks to the backfire cost) – the latter being an instance of a backfire effect.

To conclude, I shall briefly summarize the discussion of the BDMR regarding the role of incoming information. The model asserts that incoming information plays a double role during belief formation. On the one hand, it drives the Bayesian inference stage. On the other, it molds the value function. This framework provides a computational interpretation of various psychological phenomena where motivational factors are potentially critical in shaping beliefs, including the confirmation bias and the backfire effect.

Previous theories

I what follows, I shall discuss the BDMR in the context of some of the most illustrious theories of belief formation. The number of theories on this topic is enormous, while the space available here is insufficient for considering all. Thus, the discussion will be limited to some of the theories that best help highlighting the distinguishing features of the BDMR.

Argumentative theory

Hugo Mercier and Dan Sperber (2011, 2017) have advanced an influential framework asserting that human reasoning has evolved as a collective, rather than individual, faculty – they call their position *argumentative theory*. Their claim is that, within an evolutionary outlook, the natural context for human reasoning is one where people discuss with one another, and not a context where a single individual thinks in isolation. The theory asserts that conversational settings, in which multiple persons are involved, are the context where reasoning can serve its true function of providing an accurate understanding of reality. By contrast, when enacted by an isolated individual, reasoning is viewed as producing suboptimal or fallacious conclusions.

Given the influence of this perspective, it is important to discuss it in relation with the BDMR. A fundamental difference is that the two frameworks have a distinct focus: while both acknowledge the distinction between deliberative and intuitive reasoning, the argumentative theory focuses exclusively on the former and the BDMR on the latter. Moreover, the broad function attributed to reasoning diverges between the two

frameworks. Argumentative theory ultimately downplays the existence of motivated reasoning, and thus views reasoning as functional to achieve accuracy (although this can be realized only as a collective endeavor). By contrast, the BDMR assigns a central role to motivated reasoning and thus, in an evolutionary sense, views reasoning as functional to persuasion, and not to accuracy. Arbitrating between argumentative theory and the BDMR (more generally, any motivated reasoning theories) requires, therefore, establishing whether motivated reasoning exists or not – as discussed above, consensus on the matter has not been reached.

Another way the two theories diverge concerns the question of whether group thinking is accurate. According to argumentative theory, group thinking should naturally converge on consensus and on a relatively accurate description of reality – after all, this is viewed as the natural context for accurate reasoning to emerge. By contrast, the BDMR implies that consensus and accuracy can emerge within a group only if the actors are highly motivated to be impartial. If this is not the case (as it is typically not in the real world), the BDMR predicts that complex persuasion dynamics will ensue that will produce widespread disagreement and distortions. As this argument illustrates, in contrast with argumentative theory, the BDMR emphasizes the tight link between impartiality and accuracy. The idea is that, in order to reach accurate conclusions, impartiality is needed even during group discussions – while for argumentative theories reasoning as a group is enough for accurate conclusions to be reached.

Error management theory

Another influential proposal relevant to assess the BDMR is Error Management Theory, a framework that advances an evolutionary explanation of various mental biases observed among humans (Haselton & Buss, 2000; Haselton et al., 2005). Originally, the theory was proposed to interpret empirical evidence showing that, compared to women, men tend to overestimate the availability of potential sexual partners. According to Error Management Theory, this effect occurs because, in evolutionary terms, the cost of deeming someone as sexually available if this judgment is wrong is larger for women than men – the larger cost for women is explained as due to women's higher parental investment and to the risk of enduring pregnancy without male long-term support. On this basis, the theory asserts that, over the history of the human species, males who overestimated sexual availability had higher fitness, and thus spread this trait among their offspring. A similar logic has been proposed to explain various mental biases such as one toward religious beliefs and one toward conspiracy theories (Boyer, 2001; Johnson, 2009; Van Prooijen & Van Vugt, 2018).

The analogies between Error Management Theory and the BDMR are substantial. Both theories propose that, to some extent, beliefs derive from calculating the costs and benefits associated with their acceptance/rejection. Nevertheless, there is a fundamental difference between the two perspectives. In Error Management Theory, the costs and benefits are expressed in terms of fitness along the species' evolutionary history; they are not calculated online by the brain. By contrast, the BDMR claims that the calculation of costs and benefits is performed online by the brain. This difference implies divergent predictions. For example, the men's bias toward overestimating sexual availability is, according to Error Management Theory, inbuilt and rigid: men simply possess such bias as a genetic trait. By contrast, the BDMR proposes a different narrative. It suggests that, unconsciously, any person (man or woman) may ponder the benefits and costs of estimating different levels of sexual availability and build her beliefs accordingly. In this view, rather than being preestablished by evolution, any bias toward overestimating sexual availability is interpreted as stemming from an unconscious decision, and thus as a plastic and context-dependent effect. This perspective stresses the role of the context, since it asserts that biases are not fixed, but depend on the values a person pursues in a specific situation. This raises the possbility that cognitive biases such as the one regarding a tendency to overestimate sexual partners are, by and large, cultural rather than genetic – though a precise picture about cross-cultural differences is at present unavailable in this domain.

The difference between the two theories has critical implications for the study of motivated reasoning. According to Error Management Theory, the cost/benefit analysis is performed by evolution and not by the brain; in other words, at the psychological level, value calculations do not play any role in shaping beliefs. This implies that Error Management Theory does not contemplate any role for motivational factors in producing human biases, thus being inapplicable to explain motivated reasoning. By contrast, the BDMR attributes a key role to motivational factors in misjudgment, and it is therefore compatible with the existence of motivated reasoning.

Economic models of culture

The BDMR shares similarities also with economic models that leverage on market principles to explain why people embrace certain cultural belief systems (e.g., Gries et al., 2022; Iannaccone et al., 1998; Stark & Finke, 2000). A domain where these models have been applied is religion, with the aim of investigating why certain faiths have gained popularity in specific times and places (Iannaccone et al., 1998; Stark & Finke, 2000). Religious transactions are interpreted by these models as being akin to market transactions: the supply side is casted in terms of different religious institutions

competing against one other, and the demand side in terms of individuals seeking to maximize utility, the latter encompassing aspects such as prestige, eternal life, and social capital. The question of why people embrace a certain political ideology has also been examined by this framework (Gries et al., 2022). In this context, the market metaphor views, on the supply side, a competition among alternative ideological parties and, on the demand side, citizens arbitrating among the different ideologies based on the latter's ability to satisfy their motives and aspirations.

Economic models of culture are broadly compatible with the BDMR: both stress the role of utility (or value) in explaining why certain cultural narratives are more attractive than others. However, the focus of the two approaches is distinct. Notwithstanding a detailed analysis of the social transactions underpinning cultural dynamics, economic models have neglected the precise psychological processes at play. Specifically, they have focused exclusively on the utility component, downplaying the role of inferential processes. By contrast, the importance of the latter is acknowledged by BDMR thanks to the Bayesian inference component implemented by the model. Moreover, because of their focus on the behavioral domain, economic models have not explicitly addressed the question of how people's subjective beliefs arise. By contrast, the BDMR advances an explicit theory which, rather than focusing on overt behavior, analyses the origin of beliefs as expressed at the conscious level. Thus, all in all, the BDMR can be viewed as extending economic models of cultures by highlighting the importance of analyzing subjective beliefs (and not only overt behavior) and by considering the role of inference (and not only of utility) during belief formation.

Discussion

At present, there is a heated debate among scholars regarding whether, and to what extent, motivation is important during belief formation. An aspect that hinders this debate is that, at the theoretical level, an asymmetry exists when the two opposite camps are compared. On the one hand, the most influential theories that are skeptical about the role of motivation are formulated adopting computational modeling. On the other hand, virtually all theories of motivated reasoning are articulated in a verbal form. This asymmetry is problematic for at least two reasons. First, to the extent that it offers precise theories and predictions, computational modeling is highly regarded in modern cognitive science research. The fact that this approach has rarely been applied to study motivated reasoning implies that the precise computational principles underlying this construct remain poorly understood. Second, a comprehensive comparison between theories of motivated reasoning and theories downplaying motivated reasoning is problematic

because these are expressed in two different forms: verbal and computational, respectively.

To address these issues, the paper proposes the BDMR. This builds on previous models of belief formation based on Bayesian inference by adding computations concerning value. The result is an interpretation of motivated reasoning as being akin to a process reflecting an unconscious Bayesian decision. This framework appears to be broadly consistent with empirical evidence, especially about the effect of loss function asymmetries on probability judgments, about the confirmation bias, and about the backfire effect. Moreover, it is compatible with evolutionary explanations of motivated reasoning that interpret this phenomenon as ensuing from self-deception.

One common line of critique against Bayesian models is that they provide "just so" explanations (Bowers & Davis, 2012) since they can virtually explain everything. Does this criticism apply to the BDMR? Although the BDMR is relatively abstract and flexible, and thus can be partially viewed as a "just so" story, nonetheless it offers a valuable contribution for the following reason. Previous models of motivated reasoning (e.g (Ditto, 2009; Ditto et al., 2009; Festinger, 1957; Harmon-Jones & Harmon-Jones, 2007; Harvey & Weary, 1984; Heider, 1958; Hogg, 2016; Jost et al., 2004, 2022; Kahan, 2016; Kunda, 1990; Lodge & Taber, 2013; Tajfel & Turner, 1986) are all verbal and, as a consequence, relatively ambiguous. From previous models, it is not always clear what the specific computations underlying motivated reasoning are and what mechanisms underpin phenomena like the confirmation bias, loss function asymmetry, or backfire effect. By relying on mathematical modeling, the contribution of the BDMR is to spell out the mechanisms underpinning motivated reasoning in a way that is less ambiguous. Some of the ensuing explanations may appear, so to speak, obvious or trivial. But this is not necessarily a flaw, as it suggests that, if one accepts the assumptions made by the model, then certain phenomena can be interpreted almost as being "obvious." All in all, the BDMR contributes to conceptual clarity on the topic, which is important to understand the processes involved, to interpret empirical evidence, and to generate new empirical hypotheses.

Despite its generality and flexibility, the BDMR is not totally unconstrained and makes specific empirical predictions that distinguish it from other accounts of motivated reasoning. An example is that, contrary to most theories of motivated reasoning, the BDMR does not imply any optimism nor pessimism bias (see above). Because of its specificity, the BDMR can be leveraged to formulate novel empirical predictions. A set of these predictions concern loss function asymmetry effects. The empirical literature on this phenomenon is relatively scarce and there are various aspects that remain to be explored. For instance, the role of prior probability remains to be assessed systematically, and it remains to be examined whether loss

function asymmetry effects concern also the domain of gains (and not only the domain of losses) as well as contexts where both gains and losses are implied. The BDMR can offer guidance to explore these aspects. Another area where the BDMR can inspire empirical enquiry is the backfire effect. As explained above, the BDMR explains the backfire effect as due to experiencing an opposed message conveyed by a disliked source, leading to attributing a negative value to accepting the hypothesis advocated by the disliked source. From this picture, several novel empirical predictions arise. First, the backfire effect should emerge only when the opposed message is professed by a disliked source, and not by a source which is liked. Second, the strength of the backfire effect should be proportional to how much the source is disliked. Third, the BDMR predicts that a message expressed by a disliked source should become less appealing not only when the message is opposed at the beginning, but also when the message is initially neutral or even when it is initially embraced. Finally, other characteristics of the source (e.g., whether the source is reliable) should have no impact on the backfire effect. All these represent novel empirical predictions inspired by the BDMR.

Another potential criticism of the BDMR is that, since this theory is grounded on Bayesian decision, it is more complex than theories based solely on Bayesian inference. It is indeed important to stress that the BDMR is more complex than Bayesian inference theories and that, other things being equal, simpler theories should be preferred. However, the key question is whether Bayesian inference models (the more parsimonious) are sufficient to explain the empirical evidence. If they are not, then more complex models like the BMDR should be preferred (of course, if the latter can instead explain the data). Our paper does not settle the issue of whether empirical evidence supports the (simpler) Bayesian inference models or the (more complex) BDMR, but, by offering a systematic analysis of the BDMR, it can help future research that seeks to compare the two theories at the empirical level.

A related point concerns the philosophical debate about the ontology of mind. Tracing back at least to Hume's writings, an influential view asserts that the mind encompasses two distinct elements, beliefs and desires (Junker et al., 2024). At present, this dualistic position is embraced by various cognitive science frameworks such as rational decision theory and reinforcement learning (e.g., Lewis, 2018; Sutton & Barto, 1998). A monistic perspective, instead, characterizes the more recent Active Inference/Predicting Processing approach, positing that all mental dynamics, including those classically viewed as affective and motivational, can ultimately be reduced to one single process, that of Bayesian inference[6] (e.g., Clark, 2020; Junker et al., 2024; Klein, 2018). Within this debate, the BDMR can be placed among the dualistic theories, since it asserts that conscious beliefs arise from integrating two

distinct elements, Bayesian inference and value. Monistic theories afford greater parsimony than dualistic ones and therefore, other things being equal, should be preferred. However, a key point is whether monistic theories afford parsimony at the expense of explaining empirical evidence. As mentioned above, comparing monistic and dualistic models vis-à-vis empirical evidence is not the purpose of the present manuscript. Instead, here the purpose is to develop a systematic and explicit dualistic model of motivated reasoning in such a way that, in light of empirical evidence, future research can compare this with monistic models.

An open question is whether Active Inference/Predicting Processing (Clark, 2016; Friston et al., 2013, 2015; Hohwy, 2013), which is a monistic theory of brain functioning, can offer an adequate explanation of motivated reasoning. At present, a theory of motivated reasoning grounded on Active Inference/Predicting Processing remains to be developed, and it remains unclear whether Active Inference/Predicting Processing can capture the various nuances characterizing motivated reasoning. Still, this framework has been employed to explain motivation as such (Kiverstein et al., 2025; Miller Tate, 2021), and a promising avenue is to explore its potential to be extended to motivated reasoning. The present paper may provide some guidance to scholars who seek to build an Active Inference/Predicting Processing theory of motivated reasoning: it suggests that, in order to fit empirical evidence, the theory should avoid predicting an optimism bias (i.e., it cannot be simply an account that states that the most desired belief gets the highest prior probability).

In conclusion, the broad contribution of the BDMR is twofold. First, it helps understanding the computational principles behind the concept of motivated reasoning. Second, by operating at the same computational level as theories skeptical about motivated reasoning such as those grounded on Bayesian inference (Austerweil & Griffiths, 2011; Dasgupta et al., 2020; Griffiths & Tenenbaum, 2006; Hahn & Oaksford, 2007; Jern et al., 2014; Oaksford & Chater, 2007; Tenenbaum et al., 2006; Vul et al., 2014), it facilitates the comparison between competing perspectives. This may inform empirical research aimed at establishing the real contribution of motivated reasoning during belief formation.

Notes

1. The BDMR represents a generalization of recent proposals that have employed a Bayesian decision framework to investigate belief formation in specific domains including religion (Rigoli, 2021a, 2023a), political reasoning (Rigoli, 2021b), conspiracy theories (Rigoli, 2022a), delusion (Rigoli et al., 2021), forecasting (Rigoli, 2023b), and social influence (Rigoli, 2022b).

2. The definition of value proposed here is broad and can encompass a plurality of human motives. For instance, these can be self-interested as well as altruistic motives (e.g., a component of value may indicate one's desire to help another individual or the community). Moreover, accuracy motives can be included too, expressed as the desire to achieve an accurate understanding of reality.
3. The Matlab scripts used to generate Figures 2, 3, and 4 are provided as Supplementary Material.
4. Another philosophical school with remarkable similarities with the BDMR is pragmatism, especially as articulated in the work of William James. This author advanced an interpretation of beliefs as arising from considering whether acceptance of such beliefs is good or bad for an actor (James, 1897). He employed this perspective to assess various types of beliefs, from those characteristic of everyday life to those pertaining religion and science.
5. The pragmatic outlook pioneered by Pascal and James has continued to inspire philosophical enquiry (Maher, 1993) and, more recently, even psychological theories (Priniski et al., 2022; Rigoli, 2021a, 2021b, 2022a; 2022b, 2023a, 2023b). Some of these theories (Priniski et al., 2022) have claimed that people consciously decide which beliefs to express based on the costs and benefits predicted by reporting those beliefs to other people. Since, according to these theories, the decision process acts consciously, these theories deny the existence of motivated reasoning, and therefore they are ultimately incompatible with the BDMR. Other theories, by contrast, have proposed that the decision process underlying belief formation acts unconsciously (Rigoli, 2021a, 2021b, 2022a, 2022b, 2023a, 2023b), and therefore these theories argue in favor of the existence of motivated reasoning as the BDMR does. However, their focus has been on restricted domains rather than on the phenomenon of motivated reasoning at large as analyzed by the BDMR.
6. Within the predictive processing literature, some scholars (Junker et al., 2024) have distinguished between *optimistic predictive processing* accounts and *preference predictive processing* accounts. Based on this distinction, only the former accounts are monistic, since they implement desires in terms of optimistic priors, while the latter accounts are dualistic, since they postulate distinct representations for desires.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Acharya, A., Blackwell, M., & Sen, M. (2018). Explaining preferences from behavior: A cognitive dissonance approach. *The Journal of Politics*, *80*(2), 400–411. https://doi.org/10.1086/694541

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. https://doi.org/10.1080/10463280802613866

Armor, D. A., & Taylor, S. E. (1998). Situated optimism: Specific outcome expectancies and self-regulation. *Advances in Experimental Social Psychology*, *30*, 309–379. https://doi.org/10.1016/S0065-2601(08)60386-X

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*(3), 499–526. https://doi.org/10.1111/j.1551-6709.2010.01161.x

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Volfovsky, A., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bar-Hillel, M., & Budescu, D. (1995). The elusive wishful thinking effect. *Thinking and Reasoning*, *1*(1), 71–103. https://doi.org/10.1080/13546789508256906

Barrett, H. (1987). *The sophists: Rhetoric, democracy, and Plato's idea of sophistry*. Chandler & Sharp.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Bayes, R., & Druckman, J. N. (2021). Motivated reasoning and climate change. *Current Opinion in Behavioral Sciences*, *42*, 27–35. https://doi.org/10.1016/j.cobeha.2021.02.009

Belmi, P., Neale, M. A., Reiff, D., & Ulfe, R. (2020). The social advantage of miscalibrated individuals: The relationship between social class and overconfidence and its implications for class-based inequality. *Journal of Personality and Social Psychology*, *118*(2), 254. https://doi.org/10.1037/pspi0000187

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389. https://doi.org/10.1037/a0026450

Boyer, P. (2001). *Religion explained*. Random House.

Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., & Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1810), 20150228. https://doi.org/10.1098/rspb.2015.0228

Butterworth, J., Trivers, R., & von Hippel, W. (2022). The better to fool you with: Deception and self-deception. *Current Opinion in Psychology*, *47*, 101385. https://doi.org/10.1016/j.copsyc.2022.101385

Christensen, L. (2023). Optimal persuasion under confirmation bias: Theory and evidence from a registered report. *Journal of Experimental Political Science*, *10*(1), 4–20. https://doi.org/10.1017/XPS.2021.21

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, *98*(1), 1–15. https://doi.org/10.1080/00048402.2019.1602661

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412–441. https://doi.org/10.1037/rev0000178

Ditto, P. H. (2009). Passion, reason, and necessity: A quantity-of-processing view of motivated reasoning. T. Bayne & J. Fernández (Eds.), *Delusion and self□deception: Affective and motivational influences on belief formation* (pp. 21–39). Psychology Press.

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*(4), 568–584. https://doi.org/10.1037/0022-3514.63.4.568

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*(19), 1120–1132. https://doi.org/10.1177/0146167203254536

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making* (pp. 307–338). Elsevier Academic Press. https://doi.org/10.1016/S0079-7421(08)00410-6

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*(1), 53–69. https://doi.org/10.1037/0022-3514.75.1.53

Druckman, J. (2015). Communicating policy-relevant science. *PS: Political Science & Politics*, *48*(S1), 58–69. https://doi.org/10.1017/S1049096515000438

Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *Proceedings of the National Academy of Sciences*, *114*(36), 9587–9592. https://doi.org/10.1073/pnas.1704882114

Eagly, A. H., Chen, S., Chaiken, S., & Shaw-Barnes, K. (1999). The impact of attitudes on memory: An affair to remember. *Psychological Bulletin*, *125*(1), 64. https://doi.org/10.1037/0033-2909.125.1.64

Erdelyi, M. H. (1974). A new look at the new look: Perceptual defense and vigilance. *Psychological Review*, *81*(1), 1. https://doi.org/10.1037/h0035852

Evans, J. S. B. (1989). *Bias in human reasoning: Causes and consequences*. Lawrence Erlbaum Associates, Inc.

Evans, J. S. B. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Feldman, L., Myers, T. A., Hmielowski, J. D., & Leiserowitz, A. (2014). The mutual reinforcement of media selectivity and effects: Testing the reinforcing spirals framework in the context of global warming. *Journal of Communication*, *64*(4), 590–611. https://doi.org/10.1111/jcom.12108

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214. https://doi.org/10.1080/17588928.2015.1020053

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, *7*, 598. https://doi.org/10.3389/fnhum.2013.00598

Gentzkow, M., & Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, *114*(2), 280–316. https://doi.org/10.1086/499414

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Gries, T., Müller, V., & Jost, J. T. (2022). The market for belief systems: A formal model of ideological choice. *Psychological Inquiry*, *33*(2), 65–83. https://doi.org/10.1080/1047840X.2022.2065128

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Hahn, U., & Harris, A. J. L. (2014). What does it mean to be biased: Motivated reasoning and rationality. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 41–102). Elsevier Academic Press.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732. https://doi.org/10.1037/0033-295X.114.3.704

Harmon-Jones, E., & Harmon-Jones, C. (2007). Cognitive dissonance theory after 50 years of development. *Zeitschrift für Sozialpsychologie*, *38*(1), 7–16. https://doi.org/10.1024/0044-3514.38.1.7

Harris, A. J., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition*, *110*(1), 51–64. https://doi.org/10.1016/j.cognition.2008.10.006

Harvey, J. H., & Weary, G. (1984). Current issues in attribution theory and research. *Annual Review of Psychology*, *35*(1), 427–459. https://doi.org/10.1146/annurev.ps.35.020184.002235

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, *78*(1), 81. https://doi.org/10.1037/0022-3514.78.1.81

Haselton, M. G., Nettle, D., & Andrews, P. W. (2005). The evolution of cognitive bias. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 724–746). John Wiley & Sons, Inc.

Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.

Hogg, M. A. (2016). Social Identity Theory. In S. McKeown, R. Haji, & N. Ferguson (Eds.), *Understanding Peace and Conflict Through Social Identity Theory*. Springer. https://doi.org/10.1007/978-3-319-29869-6_1

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

Iannaccone, L. R., Stark, R., & Finke, R. (1998). Introduction to the economics of religion. *Journal of Economic Literature*, *36*(3), 1465–1495. https://doi.org/10.1111/j.1465-7295.1998.tb01721.x

James, W. (1897). *The will to believe and other essays in popular philosophy*. Longmans Green and co.

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

Johnson, D. D. (2009). The error of God: Error management theory, religion, and the evolution of cooperation. In S. Levin (Ed.), *Games, groups, and the global good. Springer series in game theory* (pp. 169–180). Springer. https://doi.org/10.1007/978-3-540-85436-4_10

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive-motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, *1*(10), 560–576. https://doi.org/10.1038/s44159-022-00093-5

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*(6), 881–919. https://doi.org/10.1111/j.1467-9221.2004.00402.x

Junker, F. T., Bruineberg, J. P., & Grünbaum, T. (2024). Predictive minds can Be humean minds. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/733413

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*(4), 407–424. https://doi.org/10.1017/S1930297500005271

Kahan, D. M. (2015). Climate-science communication and the measurement problem. *Political Psychology*, *36*(S1), 1–43. https://doi.org/10.1111/pops.12244

Kahan, D. M. (2016). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences*. Wiley.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kim, S. Y., Taber, C. S., & Lodge, M. (2010). A computational model of the citizen as motivated reasoner: Modeling the dynamics of the 2000 presidential election. *Political Behavior*, *32*(1), 1–28. https://doi.org/10.1007/s11109-009-9099-8

Kiverstein, J., Miller, M., & Rietveld, E. (2025). Desire and motivation in predictive processing: An ecological-enactive perspective. *Review of Philosophy and Psychology*, *16*(3), 887–907. https://doi.org/10.1007/s13164-024-00757-6

Klaczynski, P. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development*, *71*(5), 1347–1366. https://doi.org/10.1111/1467-8624.00232

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211. https://doi.org/10.1037/0033-295X.94.2.211

Klein, C. (2018). What do predictive coders want? *Synthese*, *195*(6), 2541–2557. https://doi.org/10.1007/s11229-016-1250-6

Kobayashi, K. (2018). The impact of perceived scientific and social consensus on scientific beliefs. *Science Communication*, *40*(1), 63–88. https://doi.org/10.1177/1075547017748948

Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological Bulletin*, *133*(1), 95. https://doi.org/10.1037/0033-2909.133.1.95

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480. https://doi.org/10.1037/0033-2909.108.3.480

Lewis, A. (Ed.). (2018). *The Cambridge handbook of psychology and economic behaviour* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781316676349

Linden, S. V. D., Leiserowitz, A., & Maibach, E. (2018). Scientific agreement can neutralize politicization of facts. *Nature Human Behaviour*, *2*(1), 2–3. https://doi.org/10.1038/s41562-017-0259-2

Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

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. https://doi.org/10.1037/0022-3514.37.11.2098

Ma, Y., Dixon, G., & Hmielowski, J. D. (2019). Psychological reactance from reading basic facts on climate change: The role of prior views and political identification. *Environmental Communication*, *13*(1), 71–86. https://doi.org/10.1080/17524032.2018.1548369

Maher, P. (1993). *Betting on theories*. Cambridge University Press.

Mercier, H. (2017). Confirmation bias—Myside bias. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (2nd ed., pp. 99–114). Routledge/Taylor & Francis Group.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

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

Merdes, C., Von Sydow, M., & Hahn, U. (2021). Formal models of source reliability. *Synthese*, *198*(S23), 5773–5801. https://doi.org/10.1007/s11229-020-02595-2

Miller Tate, A. J. (2021). A predictive processing theory of motivation. *Synthese*, *198*(5), 4493–4521. https://doi.org/10.1007/s11229-019-02354-y

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502. https://doi.org/10.1037/0033-295X.115.2.502

Mullainathan, S., & Shleifer, A. (2005). The market for news. *American Economic Review*, *95* (4), 1031–1053. https://doi.org/10.1257/0002828054825619

Murphy, S. C., von Hippel, W., Dubbs, S. L., Angilletta, M. J., Jr., Wilson, R. S., Trivers, R., & Barlow, F. K. (2015). The role of overconfidence in romantic desirability and competition. *Personality and Social Psychology Bulletin*, *41*(8), 1036–1052. https://doi.org/10.1177/0146167215588754

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Niiniluoto, I., Sintonen, M., & Woleński, J. (2004). *Handbook of epistemology*. Springer.

Norem, J. K. (2001). Defensive pessimism, optimism, and pessimism. In E. Chang (Ed.), *Optimism and pessimism: Implications for theory, research, and practice* (pp. 77–100). American Psychological Association.

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, *118*(15), e1912440117. https://doi.org/10.1073/pnas.1912440117

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Pilgrim, C., Sanborn, A., Malthouse, E., & Hills, T. T. (2024). Confirmation bias emerges from an approximation to Bayesian reasoning. *Cognition*, *245*, 1–15. https://doi.org/10.1016/j.cognition.2023.105693

Priniski, J. H., Solanki, P., & Horne, Z. (2022, October 21). A Bayesian decision-theoretic framework for studying motivated reasoning. https://doi.org/10.31234/osf.io/ngavz

Ricoeur, P. (1974). *The conflict of interpretations: Essays in hermeneutics* (Vol. 1). Northwestern University Press.

Rigoli, F. (2021a). A computational perspective on faith: Religious reasoning and Bayesian decision. *Religion, Brain & Behavior*, *11*(2), 147–164. https://doi.org/10.1080/2153599X.2020.1812704

Rigoli, F. (2021b). Masters of suspicion: A Bayesian decision model of motivated political reasoning. *Journal for the Theory of Social Behaviour*, *51*(3), 350–370. https://doi.org/10.1111/jtsb.12274

Rigoli, F. (2022a). Deconstructing the conspiratorial mind: The computational logic behind conspiracy theories. *Review of Philosophy and Psychology*, *15*(2), 1–18. https://doi.org/10.1007/s13164-022-00657-7

Rigoli, F. (2022b). Belief formation in the social context: A Bayesian decision account. *Changing Societies & Personalities*, *6*(4), 750–763. https://doi.org/10.15826/csp.2022.6.4.201

Rigoli, F. (2023a). The computations underlying religious conversion: A Bayesian decision model. *Journal of Cognition and Culture*, *23*(1–2), 241–257. https://doi.org/10.1163/15685373-12340161

Rigoli, F. (2023b). The intervention bias: People overpredict social problems upon which they believe society can intervene. *American Journal of Psychology*, *136*(4), 415–428. https://doi.org/10.5406/19398298.136.4.07

Rigoli, F., Martinelli, C., & Pezzulo, G. (2021). I want to believe: Delusion, motivated reasoning, and Bayesian decision theory. *Cognitive Neuropsychiatry*, *26*(6), 408–420. https://doi.org/10.1080/13546805.2021.1982686

Ronay, R., Oostrom, J. K., Lehmann-Willenbrock, N., Mayoral, S., & Rusch, H. (2019). Playing the trump card: Why we select overconfident leaders and why it matters. *The Leadership Quarterly*, *30*(6), 101316. https://doi.org/10.1016/j.leaqua.2019.101316

Rota, M. (2017). Pascal's wager. *Philosophy Compass*, *12*(4), e12404. https://doi.org/10.1111/phc3.12404

Samuelson, P. A. (1950). The problem of integrability in utility theory. *Economica*, *17*(68), 355–385. https://doi.org/10.2307/2549499

Schwardmann, P., & Van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, *3*(10), 1055–1061. https://doi.org/10.1038/s41562-019-0666-7

Sharot, T. (2012). *The optimism bias: Why we're wired to look on the bright side*. Hachette UK.

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Journal of Personality and Social Psychology*, *70*(4), 844. https://doi.org/10.1037/0022-3514.70.4.844

Smith, M. K., Trivers, R., & Von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, *63*, 93–101. https://doi.org/10.1016/j.joep.2017.02.012

Stark, R., & Finke, R. (2000). *Acts of faith: Explaining the human side of religion*. Univ of California Press.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Strickland, A. A., Taber, C. S., & Lodge, M. (2011). Motivated reasoning and public opinion. *Journal of Health Politics, Policy and Law*, *36*(6), 935–944. https://doi.org/10.1215/03616878-1460524

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1, No 1, 9–11). MIT press.

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, *9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

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

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relation* (pp. 7–24). Hall Publishers.

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology*, *28*(19), 3128–3135. https://doi.org/10.1016/j.cub.2018.07.052

Tappin, B. M., van der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General*, *146*(8), 1143. https://doi.org/10.1037/xge0000298

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Trivers, R. (2011). *Deceit and self-deception: Fooling yourself the better to fool others*. Penguin UK.

Van Prooijen, J. W., & Van Vugt, M. (2018). Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on Psychological Science*, *13*(6), 770–788. https://doi.org/10.1177/1745691618774270

Vedejová, D., & Čavojová, V. (2022). Confirmation bias in information search, interpretation, and memory recall: Evidence from reasoning about four controversial topics. *Thinking & Reasoning*, *28*(1), 1–28. https://doi.org/10.1080/13546783.2021.1891967

Von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(1), 1–16. https://doi.org/10.1017/S0140525X10001354

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. https://doi.org/10.1111/cogs.12101

Weber, E. U. (1994). From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychological Bulletin*, *115*(2), 228. https://doi.org/10.1037/0033-2909.115.2.228

Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, *65*, 7–19. https://doi.org/10.1016/j.jesp.2016.02.005

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Zmigrod, L. (2022). Mental computations of ideological choice and conviction: The utility of integrating psycho-economics and Bayesian models of belief. *Psychological Inquiry*, *33*(2), 107–116. https://doi.org/10.1080/1047840X.2022.2065134

Zmigrod, L., Burnell, R., & Hameleers, M. (2023). The misinformation receptivity framework: Political misinformation and disinformation as cognitive Bayesian inference problems. *European Psychologist*, *28*(3), 173. https://doi.org/10.1027/1016-9040/a000498