A Bayesian decision-theoretic framework for studying motivated reasoning

A Bayesian decision-theoretic framework for studying motivated reasoning

J. Hunter Priniski*
University of California, Los Angeles

Prachi Solanki*
Michigan State University

Zachary Horne †
University of Edinburgh

Abstract

Psychological, political, cultural, and sociological factors shape how people form and revise their beliefs. An established finding across these fields is that people are motivated to hold onto their beliefs even in the face of evidence by ignoring or reinterpreting information in a way that supports what they think. Although these and similar findings are compelling, the predominantly qualitative theories which guide research in this domain, and the often implicit definitions of motivation that accompany these theories, come at the cost of obscuring the cognitive mechanisms that produce motivated reasoning. Here, we introduce a new Bayesian decision-theoretic framework which describes three key factors necessary for distinguishing between cases of practically rational behavior and motivated reasoning. We demonstrate how the framework works in a series of simulations and argue that it provides guidance about what psychologists need to measure to determine where the errors in people's reasoning are occurring when they fail to revise their beliefs in light of new evidence. We then propose that this framework provides guidance for thinking about the development of interventions aimed at correcting misconceptions.

Keywords: motivated reasoning; Bayesian modeling; decision theory; informational interventions; utility; measurement

---

*Both authors contributed equally
†Corresponding author Zachary.Horne@ed.ac.uk

Contents

Reader's guide

This is a long and often technical manuscript, so we will highlight the main components so that you can focus on aspects of the paper most relevant to you.

In the introduction, we summarize key debates in the literature on motivated reasoning and orient the reader to the goals of the paper. This section is likely relevant to all readers interested in the literature on motivated reasoning.

In Part 1, we introduce some technical jargon and then focus on what researchers have implicitly or explicitly assumed about rationality, bias, and motivation in empirical studies of motivated reasoning. This section of the paper is most relevant to theorists and those interested in the distinction and definition of key concepts invoked in the literature on motivated reasoning. Readers who want a rough lay of the land but are more interested in modeling or the implications of the model can look at Figure 1 and then skip to Part 2.

In Part 2, we show how to model the distinction between motivated reasoning and practical reason (discussed in Part 1), with a particular focus on incorporating utilities into a Bayesian model of decisions to believe hypotheses. This section demonstrates the mechanics of a computational framework that guides the remainder of the paper. This section is most relevant for computational modelers or researchers who want to understand one way motivated reasoning could be computationally modeled. Readers who are not interested in the technical details of the model can read the first section of Part 2 and then skip to Part 3.

In Part 3, we consider two implications of our computational model: measurement and intervention development. Our central claim in this section is that utilities and evidence are not typically quantified but ought to be, because measuring these features allows us to determine "where" directional utilities affect people's decision to believe hypotheses. We highlight how our framework provides a new perspective on interventions for tackling misconceptions. This section is most relevant for psychologists who want a high-level summary of what parameters are typically measured in experiments on motivated reasoning and strategies researchers have come up with to correct misconceptions in the face of motivated reasoning. Readers who are primarily interested in intervention development can read Table 3 and then read the section on intervention development starting on page 43.

Introduction

There is widespread consensus in the scientific community that climate change is exacerbated by human activities, but a recent Gallup poll indicates that almost 35% of Americans do not agree with this claim (Brenan & Saad, 2018). Evolutionary theory is among the most well-supported theories in all of science, yet over 40% of U.S. citizens believe in creationism (Gallup, 2014). Meanwhile, large-scale studies have shown that vaccines are not linked to autism (Jain et al., 2015), and still 10% of the United States population believes that the side effects of vaccines are more dangerous than the diseases that they prevent (Gallup, 2014). Why do people hold these and other beliefs when they are inconsistent with established empirical evidence? Four decades of research have found that psychological, political, cultural, and sociological factors shape how people form and revise their beliefs (e.g., Alker & Poppen, 1973; Emler, Renwick, & Malone, 1983; Fishkin, Keniston, & McKinnon, 1973; Hickling, Wellman, & Dannemiller, 2001; Killen & Stangor, 2001; Schult & Wellman, 1997; Shweder, Mahapatra, & Miller, 1987; Shweder & Sullivan, 1993). A cross-cutting theme in this literature is that people hold onto their beliefs even in the face of evidence by ignoring or reinterpreting information in a way that supports what they think (e.g., Babcock & Loewenstein, 1997; Dawson, Gilovich, & Regan, 2002; Ditto et al., 2018; Gilovich, 1983; Hastorf & Cantril, 1954; Jost, Baldassarri, & Druckman, 2022; Kunda, 1990; Zuckerman, 1979).

The apparent effects of motivation to maintain one's beliefs are pervasive (e.g., Klaczynski, 2000; Kunda, 1990; Lord, Ross, & Lepper, 1979; West & Kenny, 2011). Researchers have observed motivated reasoning in political psychology (e.g., Kahan, 2013; Kunda, 1990; Taber, Cann, & Kucsova, 2009; Taber & Lodge, 2006), in attitudes about climate change (e.g., Dixon, Bullock, & Adams, 2019; Hart & Nisbet, 2012), and in science literacy (e.g., Druckman, 2015; Drummond & Fischhoff, 2017; Pasek, 2018). Even practicing scientists who are aware of the effects of motivation on cognition are not immune to their influence (e.g., Simmons, Nelson, & Simonsohn, 2011; E. C. Yu, Sprenger, Thomas, & Dougherty, 2014). Motivated reasoning is often construed as a bias that is inherent in the decision-making process, and directly related to ideological beliefs, for example, beliefs which signify and promote loyalty to an in-group (e.g., Kahan, 2013).

Social scientists often study motivated reasoning by way of conducting case studies guided by verbal theories (i.e., theories which make only qualitative predictions; Alker & Poppen, 1973; Babcock & Loewenstein, 1997; Dawson et al., 2002; Ditto et al., 2018; Emler et al., 1983; Fishkin et al., 1973; Gilovich, 1983; Hastorf & Cantril, 1954; Hickling et al., 2001; Killen & Stangor, 2001; Schult & Wellman, 1997; Shweder & Sullivan, 1993; Zuckerman, 1979), rather than performing comparisons against computational models that allow for quantification of the factors that ought to impact how people update their beliefs (e.g. Cook & Lewandowsky, 2016; Jern, Chang, & Kemp, 2014; Pilditch, Roozenbeek, Madsen, & van der Linden, 2022). For example, Nyhan and colleagues (2014) found when participants were provided corrective information about the measles, mumps, and rubella vaccine, the intervention failed to correct misconceptions among those most skeptical of vaccination, leading participants to become more hardened in their beliefs. The authors proposed that participants' response to the educational intervention was driven by motivated reasoning (Nyhan et al., 2014). To consider another example, Ditto and colleagues (1992; 2003; 1998)

examined how college students reacted to evidence that they may be unhealthy following a test of their saliva. Specifically, participants were told their saliva would be tested for an enzyme which was claimed to be linked to positive or negative health outcomes. Participants who believed the enzyme was linked to negative health outcomes concluded the test was unreliable, but this was not the case when the enzyme was linked to positive health outcomes. Because participants showed an asymmetric response based on the valence of the evidence, the authors argued that the tendency to engage in motivated reasoning drove people's responses. In both cases, a verbal theory guided the researchers' interpretation of the data.

Although these and similar findings are compelling (e.g., Babcock & Loewenstein, 1997; Dawson et al., 2002; Ditto et al., 2018; Gilovich, 1983; Hastorf & Cantril, 1954; Kunda, 1990; Lord et al., 1979; Zuckerman, 1979), verbal theories and the often implicit definitions of motivation that accompany these theories come at the cost of introducing ambiguity; it is more difficult to determine the cognitive mechanisms producing people's seemingly irrational behavior. An example can illustrate this point. Many political and fiscal conservatives are skeptical that human activities impact Earth's climate, despite near uniform agreement among climate scientists worldwide. One possible explanation of this rejection is that accepting the reality of human-caused climate change would entail radically changing how they must behave, an unwanted outcome. In turn—so the argument goes—conservatives are motivated to reject or reinterpret evidence of human-caused climate change (for evidence this occurs in other domains, see Levy et al., 2022).

However, it is possible a climate skeptic has not accessed the relevant facts in forming their beliefs, instead basing their views on unreliable sources. Conditional on these inaccurate prior beliefs, people may reason "rationally" based on the information at hand (similar phenomena have been observed in other domains, see Jern et al., 2014). This pattern of behavior could, on its face, appear to be evidence of motivated reasoning but the mechanisms underlying the output would be quite different than what researchers have typically suggested. The error would not be in how people make inferences, but rather in the *inputs* to their inferential machinery. There are now well-established alternative hypotheses of effects which were previously presumed to be evidence of motivated reasoning (e.g. Jern et al., 2014; Kahan, 2013), but in much of the research on attitude and belief change, the conceptualization of motivated reasoning-like effects as evidence of a cognitive defect persists (e.g., Dixon et al., 2019; Pennycook & Rand, 2019).

A central problem in much of the work on motivated reasoning is measurement. Prior studies on motivated reasoning have often focused on examining how evidence impacts beliefs that are inextricably linked to aspects of one's identity (e.g., beliefs about politics, religion, or morality). These topics provide a more naturalistic test of the impact of motivation on reasoning, but certain aspects of these topics may obscure the underlying cognitive mechanisms. For example, *what kind* of information is objectively relevant to coming to believe in human-caused climate change, *how much* of it ought a reasoner accommodate, and most importantly, how would we even begin to *quantify* this evidence? The very nature of evidence related to climate change (such as consensus among scientists, mathematical models, severe weather events) will make it difficult to measure the extent to which motivation impacts people's beliefs about climate change because it is unclear how much each piece of evidence *ought* to impact what someone should believe. Or to consider another

example, if we provide people with evidence that vaccines are safe and effective in the form of summary statistics from large-scale trials, how much evidence are participants actually given and, objectively, how ought participants update their beliefs in light of this evidence?[1]

In both cases, even if we could quantify this evidence, a principled mathematical benchmark needs to be assumed to properly understand the extent to which motivation biases people's reasoning, a difficult task when relying on verbal theories. Some researchers now argue that once a benchmark is defined, and human reasoning is tested against optimal models of belief updating instantiating this benchmark, it appears that people approximate these optimal models more than psychologists have often assumed (e.g. Dasgupta, Schulz, Tenenbaum, & Gershman, 2020). For example, Jern and colleagues (2014) argue that the normative standard for reasoning under uncertainty is a kind of probabilistic inference, and therefore should be subject to the axioms of probability theory. Surprisingly, they find that when comparing participants' performance to this normative model, seemingly irrational decisions—such as belief polarization—conform to (approximately) Bayesian reasoning (also see, Little, 2021). Work in several other domains has come to similar conclusions over the last two decades (Austerweil & Griffiths, 2011; Dasgupta et al., 2020; Griffiths & Tenenbaum, 2006; Jin, Jensen, Gottlieb, & Ferrera, 2022; Tenenbaum, Griffiths, & Kemp, 2006; Vul, Goodman, Griffiths, & Tenenbaum, 2014; Wallace, 2020; A. J. Yu & Cohen, 2008; Zimper & Ludwig, 2009). Whether one accepts their interpretation of the data, it highlights the need to delineate the factors that ought to impact the beliefs people form.

A further assumption that is overlooked not only in much of the original research on motivated reasoning but also newer Bayesian alternative explanations of heuristics and biases, is a somewhat technical but nonetheless important point. Namely, early research on motivated reasoning along with contemporary Bayesian interpretations often assume the rationality of the cognitive processes that deliver an output dictate whether this output needs to be corrected. To speak generally, social psychologists' interest in motivated reasoning is not limited to understanding the operations of the mind; rather, motivated reasoning is studied *because* people appear to believe many things that are inconsistent with the facts and their behavior has substantial societal implications – whether for vaccination uptake (e.g. Horne, Powell, Hummel, & Holyoak, 2015), climate change policy support (e.g. Lewandowsky, Oberauer, & Gignac, 2013), or racial justice (e.g. Kraus & Tan, 2015). These beliefs demand correction as a matter of public policy. Assuming that the source of the underlying problem is a bug in our inferential machinery is thus a plausible starting place.

Papers offering alternative Bayesian explanations (Dasgupta et al., 2020; Lieder & Griffiths, 2020) seem to make a similar assumption about the connection between underlying processes, output, and the need for correction, but the conclusions they draw from these assumptions are exactly contrary to those of many practicing social psychologists. Researchers examining the preceding questions from a Bayesian perspective suggest that if the underlying processes producing polarization, or motivated reasoning-like effects are the result of optimal Bayesian inference, there is not much to be done – people are reasoning as optimally as they plausibly could (but see Cook & Lewandowsky, 2016). In fact, we'll argue

[1] While many researchers are concerned about the possibility of backfire effects when correcting misconceptions—where corrective information leads to hardening of the misconceptions—these effects have not been reliably observed in follow-up experiments (e.g., Wood & Porter, 2019). Thus, there seems to be a lack of plausible cognitive mechanisms to explain backfire effects.

that evaluating and attempting to correct the outputs of people's inferential machinery and evaluating the rationality of the generative processes are *separate issues*. Scientists are not (and should not) be interested in motivated reasoning just as a psychological question, but because misconceptions (whatever their cause) can impact society, requiring policy-makers to act (Cook & Lewandowsky, 2016; Pilditch et al., 2022). The last section of this paper aims to demonstrate how identifying the likely source of people's errors, including identifying the rational processes producing misconceptions, may be instructive for developing more effective interventions (see Cialdini, Kallgren, & Reno, 1991; Jachimowicz, Hauser, O'Brien, Sherman, & Galinsky, 2018). The ability to accomplish this task depends on our ability to measure key psychological constructs and to use these measurements in a computational model, or so we will argue.

We can see that there is a confluence of unresolved issues surrounding the interpretation and implications of motivated reasoning. In this paper, we describe three broad factors that normatively ought to affect how people update their beliefs when confronted with evidence contrary to what they think. We define a Bayesian decision-theoretic model which serves as a computational account of how these three factors may be integrated and affect reasoning when motivations are at play. This model serves as a computational framework to better understand the limitations of prior research on motivated reasoning and devise a strategy for measuring and locating the impact of motivation on belief formation and updating. We describe key features of studies that need to be measured and controlled for to distinguish practically rational Bayesian updating from sometimes pernicious, theoretically irrational, motivated reasoning. We describe when people's reasoning ought to be subject to the norms of theoretical reason, and when their beliefs could be directionally-oriented but nonetheless compatible with the norms of practical reason. (The distinction between theoretical and practical reason is unpacked in the next section of the paper). The upshot of these considerations is not just semantic – we will argue that drawing distinctions between pernicious motivated reasoning and practical reason provides guidance about how to develop educational interventions (see Cialdini et al., 1991; Jachimowicz et al., 2018).

The paper is divided into three parts. In the first part of the paper, we introduce some necessary technical jargon and then focus on what researchers have implicitly or explicitly assumed about rationality, bias, and motivation in empirical studies of motivated reasoning. Specifically, we draw a distinction between norms of theoretical reason and norms of practical reason. A central claim in this section is that credences and decisions should be evaluated against different norms – whereas credences ought to directly reflect the way the world is, decisions are acts which ought to integrate the utility of performing those actions. Consequently, we argue that, in some cases, directional utilities *ought* to impact people's decision to believe some hypotheses.

The second part of the paper focuses on how to model the distinction between motivated reasoning and practical reason, with a particular focus on incorporating utilities into a Bayesian model of decisions to believe hypotheses. Specifically, we develop a Bayesian decision-theoretic framework and outline a toy experiment to gain traction on the question of what must be measured and how measuring these features can distinguish practical reason from motivated reasoning. This section demonstrates the mechanics of a computational framework that guides the remainder of the paper.

The third part of the paper focuses on two key implications of the Bayesian decision-

theoretic model: measurement and intervention development. Our central claim in this section is that utilities and evidence are not typically quantified but ought to be, because measuring these features allows us to determine "where" directional utilities affect belief reports. We work through an example of a prior study purportedly demonstrating motivated reasoning which we recast in the Bayesian decision-theoretic framework to demonstrate how failure to measure key ingredients obscures the underlying cognitive mechanisms. We then highlight how our framework provides a new perspective on interventions for tackling misconceptions. A central question we consider is whether we can improve the efficacy of common interventions by measuring and manipulating the utility people see in adopting some beliefs. Finally, we then consider a range of open questions and directions for future research on motivated reasoning.

Part I
How have researchers defined motivated reasoning, and alternative assumptions

Definitions and assumptions

Researchers studying motivated reasoning often appear to assume different accounts of rationality—of how people ought to reason—but a thorough discussion of just what these accounts entail is not always stated in many empirical papers. For example, many researchers seem to assume that so long as people are reasoning in ways consistent with the axioms of probability theory, we can say they are reasoning rationally (e.g. Jern et al., 2014), a kind of *probabilism* (e.g. Pettigrew, 2019). Or, on a more recent account, once we account for the fact that they have limited resources (e.g., time), people reason more rationally than they initially may appear to (e.g. Lieder & Griffiths, 2020). Here, we'll take a broader perspective and focus on different *norms of reason* and their connection to the outputs and internal states of certain cognitive processes. From this broader perspective, we'll see how making decisions which accommodate the axioms of probability theory, but also taking into account directional goals, could be practically rational.

A brief detour is necessary to link questions about rationality to key distinctions between types of motivation. On one influential account, motivation is goal-orientation, and all reasoning is motivated by goals (Kunda, 1990). The kind of goal representation that impacts a belief distinguishes rational reasoning from defective motivated reasoning. For example, *accuracy goals* shape interpretations of evidence to cohere with states of the world. In contrast, *directional goals* shape inferences to cohere with conclusions one may wish to draw about states of the world. On this account, accuracy goals ought to motivate how we update our beliefs, but directional goals ought not to.

We take a different perspective that two broad factors ought to shape people's belief reports. First, doxastic considerations (e.g., one's priors, data, and the like). Second, practical considerations, in this case, an assessment of the utility associated with believing a particular proposition, including an assessment of the utility of one's belief being true and an assessment of the *consequences* of holding that belief (Radcliffe, 1999; Schoenfield, 2018; Von Neumann & Morgenstern, 2007). We'll give some reasons why we think this shortly, but a further bit of jargon is necessary. We distinguish between two types of doxastic states, first-order and second-order beliefs, which track different internal representations and integrate different sets of information. *First-order beliefs* are constructed from doxastic considerations alone. They are beliefs as people commonly understand them. For example, the belief that the distance between Los Angeles and New York City is less than 3,000 miles. When beliefs of this sort fail to represent the way the world really is based on the evidence at hand, we'll assume they're defective (though this claim is itself a matter of debate because on revisionist accounts of rationality first-order beliefs may be "resource-rational", Lieder & Griffiths, 2020). To our knowledge, motivated reasoning as it is typically construed in the literature is a strictly *first-order* account because it assumes directional goals shape beliefs as they are constructed by biasing how evidence is sampled and posteriors are computed. On this view, researchers have inferred that directional influences bias how evidence is sampled (i.e., such that they diverge from Bayesian updating; for discussion of this account, see Little, 2021).

However, directional goals may influence beliefs in other ways — for instance, by realigning doxastic states with practical considerations after a posterior is initially computed (e.g., how reporting a belief can hurt one's chances of obtaining a goal). *Second-order*

beliefs, as we define them, are belief states that incorporate not only doxastic considerations, but practical considerations as well (see Cialdini et al., 1991; Jachimowicz et al., 2018), which we operationalize as goal representations, or more precisely, a utility-calculus.[2] Goal representations shape beliefs during this second-order inference step: after people have properly integrated evidence and determined the utility of taking certain actions. This is a second-order influence on doxastic states, which is why we call them second-order beliefs. Thus, second-order beliefs can incorporate directional influences and may therefore diverge from the verdict of a process that only relies on probabilistic and causal information from the environment. For example, a Republican may be uncertain human activities affect the climate based on the evidence they are aware of, but decide to believe our activities do not impact the climate after recognizing the consequences of believing that they do (e.g., feeling obligated to take fewer international flights etc.)

Is it ever rational—on any widely accepted account of rationality—for one's beliefs to incorporate a directional influence in a way that diverges from probabilistic information? One might immediately balk at the suggestion that utilities should affect what one believes *at all*; in Kunda's (1990) terminology, it is perfectly rational for reasoning to be motivated by accuracy, but directional goals are fundamentally at odds with rational inference. Indeed, one implicit assumption in much of the research on motivated reasoning is that the utility of holding a belief—which is directional—is *irrelevant* to what one ought to believe (Little, 2021). What makes motivated reasoning pernicious is the fact that utilities directionally impact people's beliefs in the first place. Contrary to this assumption, Bayesian epistemologists have argued that there are situations where the evidence is logically compatible with multiple hypotheses, presenting a decision problem where people can assent or dissent to a hypothesis on the grounds of expected utility (Maher, 1993). We unpack this idea below.

We can understand how rationality and motivated reasoning relate, as well as how these issues contrast with the account we develop below, by briefly reviewing work on the distinction between the norms of theoretical and practical reason (Wallace, 2020). We must acknowledge at the outset that a complete discussion of the work on the norms of reason is beyond the scope of the paper, so our treatment of these issues will be coarse-grained.

Let us introduce a bit of terminology to sharpen how we think about the issues under consideration: *theoretical reason* concerns the relationship between the acceptance of a proposition and its truth value. This kind of reasoning concerns the evidence for and truth of propositions absent the consequences of believing them. *Practical reasoning* has an "end" that is not truth per se. Rather, the aim of practical reason concerns the value of taking actions (where actions need not be literal physical acts; Wallace, 2020). The "reasons" that are relevant here are things that address actions being good or worthwhile (where goodness is not necessarily morally valenced). Thus, theoretical reasoning concerns changes in sets of beliefs and practical reasoning concerns reasons that give rise to actions, including intentions (Bratman, 1987; Harman, 1986; Wallace, 2020).

Work on theoretical and practical reason also aims to address the *norms of reasoning*. Theoretical and practical reasoning can both occur irrationally – for theoretical reasoning, it is irrational to update one's beliefs in a way that is incompatible with the evidence once we recognize an inconsistency. Here, a norm of theoretical reasoning is that we should believe

[2] To be clear, second-order beliefs should not be confused with how developmental psychologists have operationalized them in research on theory-of-mind (Perner & Wimmer, 1985).

propositions that are true. For example, a credible scientist tells me that the climate is changing, something I did not know before. All else being equal, it would be irrational not to update my belief to reflect their testimony. In contrast, ethicists typically suggest that irrationality in practical reason is close to a kind of weakness of will (Wallace, 2020). For example, I form a plan to exercise more, I ascribe high utility to exercise, but I fail to execute on my plan – this is practically irrational behavior. In this case, a norm of practical reasoning is that I execute actions (e.g., decisions, plans, etc.) that accord with (1) what I believe about the world and (2) the utilities I ascribe to those actions. Thus, the primary outputs of theoretical and practical reasoning are distinct. Theoretical reasoning outputs doxastic states or credences that aim to fit with the facts about the world. Credences can be thought of as complex, fine-grained beliefs about the world – they vary in degree and roughly correlate with the subjective probability that some proposition is true (Jackson, 2019). Practical reasoning outputs *intentions or acts*, which incorporate information about the world but will also include an aim to realize some plan, broadly defined (see Figure 1).[3]

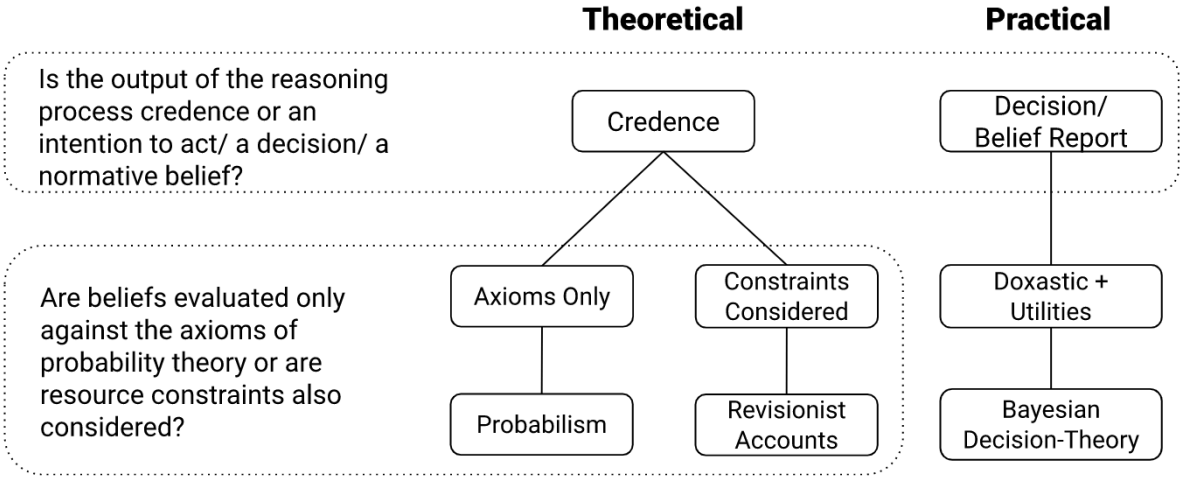

Figure 1. Tree diagram showing different accounts of reasoning.

Bearing these definitions in mind, it is tempting to assume that we ought to assess the beliefs people report in motivated reasoning experiments against the norms of theoretical reason. People should shift their beliefs to reflect the evidence they're presented with. We're going to cast doubt on the idea this assumption always holds. For example, psychologists measure whether participants' doxastic states are corrupted or contaminated by information which *ought not* bear on the beliefs people form (e.g. Ditto & Lopez, 1992; Ditto et al., 1998; Nyhan & Reifler, 2010; Nyhan et al., 2014). Psychologists need to assess whether there is a mismatch between what people believe and the facts at hand caused by a directional goal (e.g., Ditto & Lopez, 1992). However, in typical experiments on motivated reasoning, psychologists do not have unmitigated access to people's beliefs and how they reflect the world. And they would not expect to. Rather, psychologists measure *belief reports* which they hope reflect an underlying belief directly, but are the outputs of a complex psychological

[3]We should also note that we've drawn a sharp distinction between theoretical and practical reason, but this distinction may not be so clear. Our beliefs may well *always* include some action-taking component.

system which could include decision-procedures (Kahan & Braman, 2006; Maher, 1993). Thus, while we would want to assess inaccessible internal credences against the norms of theoretical reason, we'll argue that belief-reports may instead have altogether different normative constraints. Beliefs are strictly doxastic representations. Belief reports extend belief representations to jointly encode an associated action-plan, and thus the utility of taking certain actions. When people produce a belief report (i.e., a second-order belief) it can diverge from a first-order belief. In Maher's terminology, this is a kind of a "cognitive utility function" because it assigns utilities to cognitive consequences to rationally accepting hypotheses in ways that may diverge from probabilism (Maher, 1993). For example, Jones initially thinks they have enough gas in their car to reach their destination (a first-order belief), but when they realize it's particularly important their belief is true because there are no gas stations for the next 300 miles (a second-order belief with an accuracy goal), they decide to stop for gas anyway.

It is an open question whether psychologists ought to assess belief reports from typical studies on motivated reasoning against the norms of theoretical or practical reason. An example will clarify why this is so: Consider a situation in which a participant reasonably thinks the probability of H is completely uncertain, but they have a second-order belief that H would yield an enormous reward. For example, a participant thinks the evidence for the hypothesis, "The number of atoms in the universe is an even number" is completely uncertain based on mixed evidence they've been presented within an experimental study. However, the participant is told an eccentric benefactor has proposed that if, by chance, it turns out that the number of atoms is even *and* the participant *forms the belief* that the number is even, then they'll receive a million dollars. Psychologists now ask the participant whether they believe the number of atoms in the universe is an even number. The participant uses a Likert scale to indicate they "Strongly agree" the number of atoms is even. This team of psychologists has not measured anything else.

There are two possibilities: One possibility is that the million-dollar prize has caused the participant to think the evidence for the even-atoms proposition is very strong. This is a clear case of directionally-motivated reasoning as it is traditionally conceived, and (arguably) subject only to the norms of theoretical reason – the output, a credence, dictates the norms of evaluation. The participant has sampled from evidence in a way so as to discount evidence against the proposition and in line with the reward.

A second possibility is the participant understands that the amount of evidence for the even-atoms proposition is uncertain, but after assessing the evidence and weighing the utility of forming the belief, they have *formed the intention to believe* that the number of atoms in the universe is an even number. They select a response option on a Likert scale to match this decision. In this second case, it is uncontroversial that the norms of reasoning will involve not just an assessment of doxastic features of the situation, but also those related to the utility of holding a given belief. Further, it is *practically rational* to form the intention to believe the number of atoms in the universe is even because the evidence is compatible with multiple hypotheses, the reward for one hypothesis is large, and they need to make a decision (e.g., Briggs, 2019; Maher, 1993). We cannot distinguish these cases when we've only measured belief reports without quantifying the evidence, without measuring a participant's understanding of the evidence, and without measuring the utilities people assign to a belief report.

The distinctions we've drawn might seem esoteric, but they have substantial consequences. For example, suppose a participant thinks the evidence for a wonky policy aimed at reducing carbon emissions is uncertain based on the limited evidence they could reasonably have access to, but they later learn that the majority of their political party supports the policy. The participant doesn't think that their party supporting the policy *necessarily* provides evidence they're correct, but the participant is well-aware that holding a view consistent with their broader party has high utility (Kahan & Braman, 2006). An experimenter conducts a study examining how presenting mixed-evidence for the wonky policy affects people's beliefs with different party affiliations and finds that, sure enough, political affiliation appears to impact the belief the participant reports – Democrats seem to focus on one aspect of the mixed evidence and Republicans the other competing aspect of the evidence (Lord et al., 1979). Just as in the even-atoms case described above, this behavior is completely compatible with participants respecting the norms of practical reason. Their assessment of the evidence is not inaccurate (the evidence is unclear), but there is utility in forming a belief report that's party-compatible, so their belief report is in line with the perceived utility of holding this belief (see Cialdini et al., 1991; Kahan & Braman, 2006; Sunstein, 2000). There is experimental evidence this occurs: people see utilities in forming beliefs, and anticipate the consequences of those beliefs (Falk & Zimmermann, 2016; Golman & Loewenstein, 2018; Jachimowicz et al., 2018; Levy et al., 2022).

Our view is that it is likely people sometimes act in (1) ways inconsistent with the norms of theoretical reason, and (2) ways consistent with the norms of practical reason despite having a directional goal. In the former case, we'll assume people engage in motivated reasoning and we should assess them by the norms of theoretical reason (but see Lieder & Griffiths, 2020). Our point here is that the ostensible errors, those which are directionally goal-oriented, but nonetheless rational, are also likely to occur. We need a computational framework to distinguish these cases.

Further considerations and implications

Weighing accuracy and directional utility. One immediate question concerns situations where it's practically rational for directional-goals to influence one's belief reports. For example, when the evidence is logically compatible with different hypotheses, assent or dissent on a hypothesis could be determined based on an expected utility calculation (Maher, 1993). However, as Kunda (1990) argues, the motivations in question can take the form of both accuracy goals and directional goals. This raises the question of *how much* utility people ascribe to forming true beliefs and how people weigh this calculation against directional goals. Even under the norms of practical reason, participants' utility function over accuracy and direction needs to be accounted for to determine how we should assess the compatibility of their belief report with a mathematical benchmark which dictates the rationality or irrationality of their behavior. One possibility is that once we take the utility of accuracy into account as well, it may well be that people's belief reports should primarily be driven by accuracy goals rather than directional goals. It is an empirical question but we'd conjecture, for many beliefs, participants are not particularly worried about being inaccurate because they do not care about the truth of a given proposition per se, in the same way practicing psychologists are unlikely to care that they're right about the even-atoms proposition (Falk & Zimmermann, 2016; Kahan & Braman, 2006).

Instead, participants are invested in the direction of a hypothesis insofar as they perceive the consequences of forming that belief for their future behavior or recognize how holding that belief would change how people perceive them (Falk & Zimmermann, 2016; Kahan & Braman, 2006). Clearly, people also have strong convictions; they have beliefs which they are invested in (Skitka, 2010). For example, it's clear people have strong views about many general claims, like the claim human-caused climate change is real. However, their beliefs about very specific climate policies are unlikely to take this form (Kahan & Braman, 2006); we think these situations are likely the norm rather than the exception. We should note, though, that there will be individual differences in the utilities assigned to accuracy and directional goals. For example, as a practicing scientist, we might wholly be concerned about accuracy, so much so that directional goals are perceived as being irrelevant to the question of what we believe (but see Maher, 1993).

When and where do credences and belief reports match. A fundamental assumption in our discussion of first and second-order beliefs concerns the possibility of people's internal credences and belief reports can diverge. What is the empirical evidence for this claim? In fact, there is rich literature which is specifically focused on these and related questions (Orne, 2017). For example, the creation of one of the most famous paradigms in psychology—the Implicit Associations Test—rests on the idea that internal credences about topics like racism, sexism, and the like can diverge from people's belief reports. It is a secondary question whether tests like the Implicit Associations Test accurately capture anything meaningful. But, the fact that these and other more implicit measures exist suggests that psychologists are well-aware first-order and second-order beliefs could diverge for a host of reasons (e.g. Greenwald et al., 2020; Orne, 2017). To consider another example, it is well-known that political polling is not so straightforward that pollsters can simply ask who one plans to vote for without considering any other factors; people opt out of studies when they are asked about certain topics (non-response bias; Groves & Peytcheva, 2008), or might withhold stating their attitude about a polarizing topic for fear of the consequences of taking a stance (e.g., Harrison & Startin, 2013). This doesn't mean that second-order beliefs are always strategic, deceptive, or insincere. They could result from a deliberative process of weighing their certainty about the evidence, their investment in the truth of the proposition, and the perceived utility of forming the belief. And these computations could happen either explicitly or implicitly. [4]

The current situation we've outlined might lead one to think that we are skeptical credences could *ever* be measured in a way to distinguish motivated reasoning from practically rational behavior. This is not a claim we're committed to. Rather, in many cases belief reports surely reflect people's credences, particularly in cases where there is no utility at all in forming a belief. For example, suppose that a participant is asked whether they think the distance between New York City and Los Angeles is greater than 3,000 miles. We'd expect whatever the report here to directly reflect their credence. However, we think the situation is altogether different for many of the typical topics social psychologists have aimed to study. Cases where social psychologists test for motivated reasoning are often the very cases where belief reports are *least likely* to reflect credences directly because participants

[4]Second-order beliefs could result from metacognitive readjustments of a first-order belief constructed using lower-level perceptual and evidence quantification processes (e.g., Maniscalco & Lau, 2012). In this sense, a second-order belief is explicitly held while a first-order belief is not.

can anticipate the consequences of forming a belief about that topic (Falk & Zimmermann, 2016); the topic matters to them and so they consider the consequences in their deliberation (Horne, Powell, & Hummel, 2015). This means that researchers must go beyond measuring belief reports simpliciter. Researchers must measure the utility participants assign to holding some beliefs, and participants' understanding of the evidence for those beliefs.

Revisionist definitions of rationality. Our discussion of practical and theoretical reason might raise questions in the reader's mind about how our account relates to well-established research questions surrounding the definition of rationality, or, for instance, attempts to reconstrue heuristics and biases as instances of rational decision-making (e.g. Dasgupta et al., 2020; Hahn & Oaksford, 2007; Horne & Livengood, 2017). The focus of this paper is to provide a computational framework for comparing motivated reasoning against *established* norms of reasoning, like practical reason (Wallace, 2020). We are not proposing a *revision* of the definition of theoretical rationality to account for features like the adaptiveness of people's behavior in complex environments (e.g. Lieder & Griffiths, 2020; Mercier & Sperber, 2011).

Our decision to focus on comparing behavior against established norms is, in part, due to the impasse in the literature about when and where people violate the norms of *theoretical* reasoning. Whether the cases where credences are incompatible with norms of theoretical reasoning are actually adaptive (Lieder & Griffiths, 2020) and thus "practically" rational in another sense of the word, is beyond the scope of this paper. Instead, we are focused on the utility of forming certain beliefs and cases where it is uncontroversial that utility ought to affect people's decisions. More broadly, we are aiming to provide a computational account of how belief reports are generated to compare behavior against established norms and speak to the literature on motivated reasoning on its own terms.

We should also mention a related issue in the psychological literature on rationality, namely, the revisionist idea that rationalization is rational (Cushman, 2020). Theories of rational action assume beliefs and desires give rise to actions. An action is rational if it maximizes one's desires conditioned on their beliefs. Rationalization inverts this process: actions give rise to beliefs and desires constructed *post-hoc* to justify an action. Recently, Cushman (2020) has argued that rationalization is in fact a rational process because it serves a necessary psychological function – it makes information from non-rational influences on behavior available during reasoning (e.g., instincts, social norms). Whether rationalization is in fact rational is a question we won't consider here. Our account of second-order beliefs is part of a theory of rational action rather than rationalization; actions are computed from credences (beliefs) and utility-representations (desires). A utility-calculus can, for instance, encode information about social norms, which is provided to a second-order belief and action plan. This is distinct from a situation in which utility information is subsequently used to provide justification for an action a reasoner has *already decided* (Audi, 1985). In this latter case, social norms constrain actions to a subset of the action space, and people construct beliefs and desires to justify why they performed the action conditioned on the norm. People do both of these things (Botzen & van den Bergh, 2012; Cushman, 2020; Falk & Zimmermann, 2016; Maher, 1993), but when and how to distinguish them is an important question.

An example will clarify the issue: Even though Democrats strongly believe in human-caused climate change, they may contribute more to carbon emissions from air travel than

Republicans. Democrats sometimes justify their actions by citing the fact that major companies are the primary contributors of climate change, but this justification seems to be *post-hoc* to their decision to fly. This is a kind of rationalization because the *actual* reason they choose air travel is that they want to and it's easier, but the reason *they state* is the fact that major companies are primarily responsible for emissions – the actual reason and the stated reason diverge (Audi, 1985). We can see then that rationalization is quite different than incorporating utilities in one's decision about a hypothesis when, for example, the evidence is compatible with two competing hypotheses (Maher, 1993).

To conclude, distinguishing theoretical reason from practical reason helps us sharpen how we talk and think about directionally-motivated belief reports. Directional goals can rationally influence second-order beliefs in a way distinct from prior accounts, which assume any directional influences are evidence of problematic motivated reasoning (Kunda, 1990). In the next section, we develop a Bayesian decision-theoretic framework to guide researchers' measurement of key cognitive constructs and distinguish motivated reasoning from practically rational behavior.

Part II
How we distinguish practical rationality from motivated reasoning

Computational models of motivated reasoning and Bayesian decision-theory

The first factor is the evidence people update their beliefs on. The evidence people have and how it is weighted could take a number of forms. For instance, people believe things because of the testimony of their friends or experts. Likewise, they believe things based on their own observations about the world. Reasoners assign some weight to each kind of evidence – the extent to which they think it supports a hypothesis. For example, some evidence, like anecdotes from friends, might be perceived as relatively weak evidence compared to testimony from experts. Though, of course, trust in expert testimony will dictate the weight evidence is likely to be assigned (Imundo & Rapp, 2022).

The second factor is what people believed, and the information available to them, before considering new data. This could be in the form of known base-rate information, or a mere hunch about the credibility of a hypothesis. We'd expect some of our priors to be very weak, perhaps even flat, whilst others to exert a strong effect on what we believe even after we confront new data. For example, our prior that the number of atoms in the universe is an even number might be completely uncertain, but our prior that extrasensory perception is not real might be extremely strong. In the former case, even a little bit of new information could shift our credence in the proposition, but a lot more data would be needed to materially impact our belief in extrasensory perception.

The third factor we'll focus on is the utility of holding a belief. This is the least clearly defined. Here, we will assume that a reasoner who is motivated to believe a hypothesis associates some *utility* with the prospect of holding that belief. For instance, utility could be associated with the social benefits attained from displaying in-group support of climate change skepticism (Kahan, Landrum, Carpenter, Helft, & Hall-Jamieson, 2017; Sidanius & Pratto, 1999). People may perceive some hypotheses as having extremely large utilities (e.g., eternal salvation) or little to no utility at all (e.g., believing the number of atoms in the universe is an even number).

We'll now discuss the computational mechanisms of motivated reasoning and Bayesian decision-theory. We first describe the typical mechanisms for generating first-order beliefs, namely, using Bayesian updating. The motivated reasoning model and Bayesian decision-theoretic model both perform Bayesian updating. What distinguishes these models is that the motivated reasoning model computes posteriors using a directional prior, a prior which encodes utility information about hypotheses. In contrast, the Bayesian decision-theoretic model computes posteriors via Bayesian updating, after which utility information is integrated to produce a second-order belief report. Readers who are not interested in the details of our computational framework can skip to Part 3 of the paper.

First-order processes

Bayesian updating: Computing first-order beliefs with normative priors. Bayesian updating is the standard model for computing first order beliefs from data. Both the motivated reasoning model and the Bayesian-decision theoretic model generate beliefs via Bayesian updating. People hold beliefs about hypotheses and these hypotheses include a subjective likelihood attributed to each hypothesis being true. A hypothesis could be a logical, numerical, or a natural language proposition that explains some feature of a (logical, statistical, or real-world) system. A single hypothesis, h_i, comes from a larger

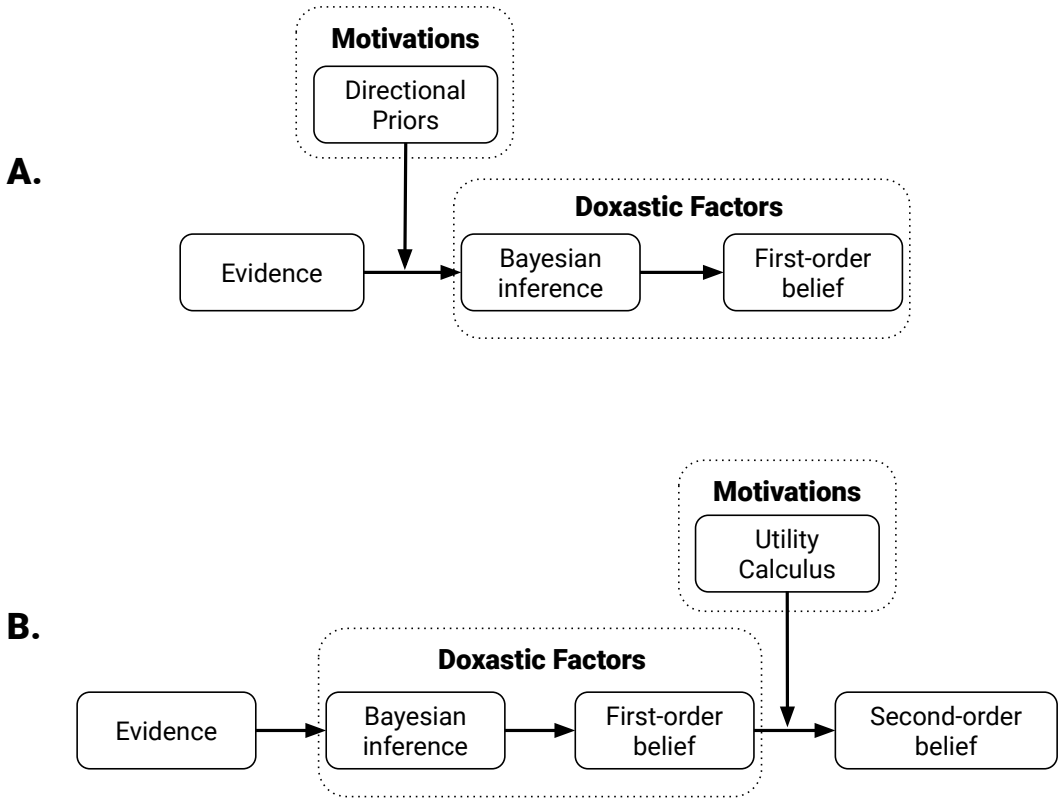

Figure 2. Flowchart showing two ways motivations can shape belief reporting. Diagram A is the motivated reasoning account. The dashed arrow captures direct influence of motivations on inferential machinery. Previous research doesn't specify how this happens. Diagram B is the Bayesian decision-theoretic account. It shows how utilities can exert influence on first-order doxastic representations to generate a second-order belief.

set of hypotheses called a hypothesis space, $\mathcal{H}$. The hypothesis space expresses a set of hypotheses which partition a set of *possible worlds* as either true (i.e., actual) or false, where "possible worlds" just means different ways the world could be. Let $h_i \in \mathcal{H}$ be a hypothesis from a hypothesis space.

People deliberate about hypotheses given data, which needs to be reflected in the framework. We need to computationally define how evidence is integrated with prior beliefs. Bayesian updating is a normative framework for doing this. Bayes' Theorem describes how to update the perceived probability of certain hypothesis, h_i, given new data. It is composed of three parts: a prior $\mathbb{P}(h_i)$, a likelihood $\mathbb{P}(d|h_i)$, and a normalizing constant given the observed data $\mathbb{P}(d)$. When a prior is not conditioned on utility-information, we say it's a *normative prior*. Normative priors are the basis of the Bayesian decision model which we describe at the end of this section. The motivated reasoning model uses non-normative priors because they are dependent on utility information.

Let $d \in \mathcal{D}$ be a representation of some data d sampled from a "data space" $\mathcal{D}$. For

instance, if this summer is warmer, d, than previous years $\mathcal{D}$, a reasoner would need to update their beliefs about how hot summers are likely to be. Data is evidence that can be integrated to generate doxastic states. We can express the updated probability in their hypothesis h_i given d using Bayes' Theorem:

$$\mathbb{P}(h_i|d) = \frac{\mathbb{P}(d|h_i)\mathbb{P}(h_i)}{\mathbb{P}(d)} \tag{1}$$

where $\mathbb{P}(d)$ is a normalizing constant, that scales the computed posterior values so they sum to one and form a probability measure. Modelers often disregard the probability of the data and leave the posterior unscaled because doing so leaves the relative credences unchanged and simplifies the computations. The posterior calculation can therefore also be simplified to:

$$\mathbb{P}(h_i|d) \propto \mathbb{P}(d|h_i)\mathbb{P}(h_i) \tag{2}$$

Bayesian updating is the normative process (that is, it satisfies the axioms of probability theory) for integrating information into belief states. It is uncontroversial that beliefs updated using this method satisfy the norms of theoretical reason (Vineberg, 2022). We'll denote doxastic states that are only updated on the data as *first-order beliefs*. They are outputs of the Bayesian updating module from the computations in Equation 2 above.

Motivated reasoning: Directional priors affect evidence integration. In the Bayesian-decision model, which we'll detail after this section, normative priors—priors which are not impacted by motivation—dictate how evidence is integrated to form a first-order belief. This is in contrast to the common account of motivated reasoning, which assumes that people are theoretically irrational when motivations shift how they compute their first-order beliefs. In this paper, we develop a motivated reasoning model which we think best characterizes how social psychologists have qualitatively described motivated reasoning. We assume motivation biases the construction of first-order beliefs by directing prior credences to sample data for hypotheses ascribed higher utility. To instantiate this assumption in a computational theory, we define *directional priors*, prior credences conditioned on utilities. Directional priors weight the prior probability of hypotheses by their utilities. They can be defined more generally as $\mathbb{P}^*$, where one's prior credence in a hypothesis h_i is conditioned on its cognitive utility u_{h_i}:

$$\mathbb{P}^*(h_i) = \mathbb{P}(h_i|u_{h_i}) \tag{3}$$

We can turn to Equation 2 to understand how directional priors lead to motivated beliefs. Evidence updates a posterior by multiplying the prior and the likelihood. The likelihood quantifies the amount of information a piece of evidence has, conditioned on a hypothesis being true. In other words, how much learning should happen given data under *each* hypothesis. Priors, on the other hand, encode the *a priori* probability a reasoner attributes to each hypothesis being true. The likelihood cannot be interpreted independently of the prior because the prior dictates the extent to which the likelihood function will shape a first-order, posterior belief. Credences sampled from $\mathbb{P}^*$ violate the norms of theoretical reason because typical accounts of motivated reasoning assume that directional utilities should not shape first-order beliefs (Kunda, 1990). Note that revisionist accounts

of rationality may disagree with this claim (Lieder & Griffiths, 2020). From a modeling perspective, we need to specify a $\mathbb{P}^*$ that best suits a given domain. This requires establishing some mathematical relation between utilities and prior credences. To illustrate, in the next section we define, implement, and simulate predictions from Bayesian models with "optimal" and "motivated priors." We show how $\mathbb{P}^*$ can be defined as a Beta distribution. In this motivated reasoning model, an expectation is equal to the expected relative utility of a hypothesis. We simulate predictions with varying utilities to demonstrate how they impact the model's credences.

Thereafter, we provide a toy example of a study which could distinguish motivated reasoning from practical reason where we measure the parameters discussed above. We relate two models, the motivated reasoning model and a Bayesian decision model, to this toy study to show how a study structured this way may be able to distinguish motivated reasoning and practical reason.

Second-order processes

Computing the expected utility of beliefs. The framework we present describes how utility information can alter doxastic states and produce second-order beliefs to generate belief reports. We now describe an expected utility account of how second-order belief states can be derived from a posterior computed using normative priors. Note that theoretical rationality says nothing about second-order beliefs, so this step only applies to the Bayesian-decision model. We will first describe how expected utilities can be computed for hypotheses and then describe how to incorporate Bayesian inference in this model.

We'll write the expected utility of reporting h_i as:

$$EU(h_i) = \mathbb{P}(h_i)\,\mathbb{U}(h_i) \tag{4}$$

where $\mathbb{P}(h_i)$ is the probability that a given hypothesis h_i is true, and $\mathbb{U}(h_i)$ is the utility the reasoner attributes to h_i being true. For instance, adopting a new belief can be sensitive to consequences (Williams, 2021), such as maintaining or severing ties to one's social group (Kahan, 2013). Social consequences broadly, and the maintenance of the ties we form to our social groups particularly, have utilities. Consider a reasoner forming a belief about climate change. The consequences of believing in human-caused climate change is the summation of the consequences of believing that proposition is true. What does that mean? Some consequences could be increased business regulation, gasoline taxes, and being discouraged from air travel unless it is necessary. There are further consequences, of course, such as the social relationships we maintain as a result of taking a side on a "controversial" issue.

How may we operationalize this? A reasoner may encounter a set of K outcomes (or prospects) when believing h_i, which can be expressed as $c_{h_i} \subseteq \mathcal{C}(\mathcal{H})$. We use c to denote outcomes because they can understood as real-world consequences the reasoner attributes to a hypothesis being true. These symbols say the outcome of the hypothesis h_i is within the set of all possible outcomes. Individual outcomes associated with h_i are further indexed by $c^k_{h_i} \in \mathcal{C}_{h_i}$. The notation for $c^k_{h_i}$ can be understood as follows: the subscript, h_i denotes which hypothesis a given outcome belongs to (here, h_i) and the superscript, k denotes the indexing within the set of outcomes belonging to the hypothesis. Therefore, $c^k_{h_i}$ means the

k^{th} outcome given hypothesis h_i. The reasoner needs to index a hypothesis' outcomes in order to compute $\mathbb{U}(h_i)$.

Outcomes occur with some degree of uncertainty and this uncertainty can vary as a function of the credence in a hypothesis. For instance, a pro-vaccination parent and an anti-vaccination parent may attribute similar costs to potentially negative consequences of a vaccine (e.g., their child getting sick from a vaccine dose), but anti-vaccination parents may assume the negative consequence are more likely than the pro-vaccination parent (Horne, Powell, Hummel, & Holyoak, 2015). We need to encode this uncertainty with a probability measure. In English, the expression below means the joint probability of h_i being true and the probability of possible consequences of h_i manifesting as a result (being caused by the actuality) of h_i. We can rewrite $\mathbb{P}(h_i)$ as the probability of the conjunction of h_i and its associated outcomes as c_{h_i} as the joint probability $\mathbb{P}(h_i, c_{h_i})$. We will expand out this joint probability in the equation below using a conditional probability.

For generality, we can say that h_i produces a set of k outcomes with computable utilities. By extending Equation 4 above, we can express the utility of h_i as a scaled proportion of its credence and credence of its consequences. More concretely, given a set of k outcomes to h_i, $c^k_{h_i} \in c_{h_i}$, we can then write the expected utility of a hypothesis as a function of the utilities associated with consequences conditioned on the hypothesis, which are scaled by the probability that the hypothesis is true and that the consequences obtain:

$$EU(h_i) = \mathbb{P}(h_i, c_{h_i})\, \mathbb{U}(c_{h_i}) = \mathbb{P}(c_{h_i}|h_i)\mathbb{P}(h_i)\, \mathbb{U}(c_{h_i}) \tag{5}$$

where we iterate over the possible outcomes and their associated utilities:

$$\sum_{k=1}^{K} \mathbb{P}(c^k_{h_i}|h_i)\mathbb{P}(h_i)\, \mathbb{U}(c^k_{h_i}) = \mathbb{P}(h_i) \sum_{k=1}^{K} \mathbb{P}(c^k_{h_i}|h_i)\, \mathbb{U}(c^k_{h_i}) \tag{6}$$

The preference of an outcome depends on the values of other possible outcomes (Von Neumann & Morgenstern, 2007); all else being equal, people prefer outcomes which yield the greatest relative expected utility. We need to express this as a relative expected utility calculation (the expected utility of a hypothesis *relative* to the other hypotheses in the hypothesis space). Rather than having a utility calculation for a hypothesis, we need to normalize it for the sum of the consequences for all other hypotheses. This will allow us to think more clearly about the utility between options (which option dominates the other). We use the function $z(h_i)$ to define the *relative utility* for a hypothesis h_i given the utilities of the other hypotheses. The relative utility can be thought of as a scaled utility value, where utilities for all outcomes are normed to be between the values of zero and one. We define a relative utility function because it will be helpful when explaining computations in the cognitive models.

There are various ways a relative utility function can be computed as the shape of this equation depends on the preference space (e.g., Regenwetter, Dana, & Davis-Stober, 2011) and other psychological mechanisms (e.g., loss aversion; Tversky & Kahneman, 1991). Here, we describe the simplest relative utility function which is essentially a value of the summation of outcomes for each hypothesis scaled by the summation of utilities for all other outcomes for all other hypotheses. We are not saying this is the computational mechanism for preference formation, but use this equation as an example to illustrate our framework.

Let $K_i = |C_{h_i}|$ where $|C_{h_i}|$ denotes the total number of outcomes (or the size of the set of consequences) for h_i, then:

$$z(h_i) = \frac{EU(h_i)}{\sum_{h_j \in \mathcal{H}} EU(h_j)} = \frac{\mathbb{P}(h_i) \sum_{k \in K_i} \mathbb{P}(c^k_{h_i} | h_i) \, \mathbb{U}(c^k_{h_i})}{\sum_{h_j \in \mathcal{H}} \mathbb{P}(h_j) \sum_{k' \in K_j} \mathbb{P}(c^{k'}_{h_j} | h_j) \, \mathbb{U}(c^{k'}_{h_j})} \tag{7}$$

Here, $z(h_i)$ is the relative expected utility function which will compute the utility of a hypothesis *relative* to the other hypotheses under considerations (i.e., $h_j \in \mathcal{H}$). This is why in the second part of the expression (i.e., $\frac{\mathbb{U}(h_i)}{\sum_{h_j \in \mathcal{H}} \mathbb{U}(h_j)}$) we compute the utility of h_i in the numerator (i.e., $\mathbb{U}(h_i)$) and divide it by the summation of utilities for the remainder of the hypotheses in the denominator. This is somewhat handwavy as a hypothesis in the abstract doesn't have a utility per se, rather it is represented as being evidence for possible outcomes that do have a real-world, computable utilities. So in the third part of the equation, we expand out the equation to make this fact explicit. This is an application of Equation 6, where the expected utility of h_i is equal to the summation (from 1 to K outcomes) of the joint probability of a hypothesis and its outcome, multiplied by the utility of those consequences. This is just an application of Bayes' rule: the joint probability $\mathbb{P}(h, c_{h_i})$ is equal to the prior probability of h_i ($\mathbb{P}(h_i)$) multiplied by the probability of the outcome c_{h_i} conditional on the credence in hypothesis h_i. We'll assume for simplicity that people often attempt to hold hypotheses that are maximally rewarding. Stated another way, people generally aim to maximize the expected utility they perceive as being a consequence of holding a hypothesis. We are not specific when describing how to compute a utility function, because there are various ways to do so (Maher, 1993). For instance, a reasoner may consider a trade-off between a concern for accuracy (which leads to accepting hypothesis with high probabilities) versus informativeness (lower probability). A utility function capturing preference for one desideratum over the other is a *cognitive utility function*, as it ascribes utility-values to the outcomes of a hypothesis being true.

In its current form, the framework shows how utility information can be incorporated into the selection of hypotheses. When a hypothesis is selected given this processes, we call it a second-order belief. We interpret this process as a second-order sampling of a posterior scaled by a utility calculus. Utility information can lead to divergences in belief reports (when conditioned on identical data) in one of two ways: (1) When reasoners set different utility values to different outcomes and (2) when reasoners infer different likelihoods of consequences obtaining.

Consequently, the hypothesis h^* (the sampled second-order belief) is that which maximizes the expected relative utility calculation in Equation 7:

$$h^* = \max_{h_i \in \mathcal{H}} z(h_i) \tag{8}$$

The expected utility of a hypothesis is by definition a function of the likelihood of that hypothesis being true, scaled by the utility of its outcomes being true. Therefore, as Equation 8 states, the hypothesis with the largest relative expected utility is the one that maximizes these two constraints (i.e, the likelihood of being true scaled by the utility of its trueness). To be precise, the actual belief a reasoner reports is the degree of belief

(subjective probability of truth) they assign to the maximizing hypothesis h^* being true, $\mathbb{P}(h^*)$.

Incorporating Bayesian inference when computing second-order beliefs. By integrating Bayesian updating in our expected utility framework, we get the Bayesian decision framework, with the three factors stated above (i.e., the likelihood, the prior, and the utility):

$$EU(h_i|d) = \mathbb{P}(h_i|d)\,\mathbb{U}(h_i) \propto \mathbb{P}(d|h_i)\mathbb{P}(h_i)\,\mathbb{U}(h_i) \tag{9}$$

Equation 9 is an extension of Equation 5 in which $\mathbb{P}(h_i)$ is replaced with the posterior of h_i given d. Therefore, the expected utility of a hypothesis is updated via Bayes' Rule to incorporate learning new data.

We can then expand Equation 9 to its fully expressed form:

$$EU(h_i|d) = \mathbb{P}(h_i|d) \sum_{k \in K_i} \mathbb{P}(c^k_{h_i}|h_i)\,\mathbb{U}(c^k_{h_i}) \propto \mathbb{P}(d|h_i)\mathbb{P}(h_i) \sum_{k \in K_i} \mathbb{P}(c^k_{h_i}|h_i)\,\mathbb{U}(c^k_{h_i}) \tag{10}$$

This equation states that we will update the utility in line with Bayesian updating. Further, people will report the belief which maximizes utility in this equation, given a utility calculation and an integration of data where the reasoner reports the belief h^* given data. We calculate that value with this expression:

$$h^* = \max_{h_i \in \mathcal{H}} z(h_i|d) \tag{11}$$

This function maximizes the relative utility over the hypothesis space, given observing a new piece of data d. This is how a reported belief is calculated in our framework.

An implication of our proposal is that reasoners weigh the Bayes factor of a hypothesis against the relative utility of an opposing hypothesis. In effect, weighing evidence against utility. This result has important implications (and makes testable predictions) for how we think about concrete examples like how people form beliefs about climate change, vaccines, and the like. Consider a case where a reasoner has to decide what to believe between two hypotheses h_1 and h_2 given data d and utilities of the consequences of $\mathbb{U}(c_{h_1})$ and $\mathbb{U}(c_{h_2})$. In this framework, $EU(h_1) = \mathbb{P}(h_1|d)\,\mathbb{U}(c_{h_1})$ and $EU(h_2) = \mathbb{P}(h_2|d)\,\mathbb{U}(c_{h_1})$. And our central claim is that if $EU(h_1) > EU(h_2)$ then the reasoner will decide on h_1. This is because the reported belief h^* is the hypothesis h_i with the largest value. We discuss this result in the model simulation section.

How do motivation-representations affect beliefs?

As shown in Figure 2, a reasoner evaluates evidence for or against their beliefs in generating a belief report. One way directional motivations could impact belief updating is by shaping evidence representations in the Beta-Bernoulli model. This is the view tacitly assumed by much of the research on motivated reasoning. In contrast to this view, directional motivations can alter one's decisions *after* integrating the evidence to construct a first-order belief. Motivations shape second-order belief representations after a Bayesian

model optimally transforms evidence to a belief. In the motivated reasoning model, first-order beliefs encode directional information. In the Bayesian decision model, first-order beliefs do not encode directional information — this is a key difference between the models.

In the former case, directional goals "corrupt" first-order beliefs (but again, see Lieder & Griffiths, 2020). In situations in which this occurs, it is possible then, for example, that encouraging people to attend to the accuracy of headlines they read could lead to better sampling of the evidence and in turn better judgements about the truth or falsity of the news (Pennycook, 2022; Pennycook & Rand, 2021). However, in situations where directional goals affect belief reports *after* evidence integration, these interventions may not be as effective – we return to this point in more detail in the *Implications for developing interventions aimed at correcting misconceptions* section of the paper. These are not mutually exclusive possibilities in that the effects of motivation could manifest both at the stage of updating and after evidence representations are formed.

These two paths of influence have different implications from the standpoint of rational updating: Directionally motivated reasoning could be a violation of theoretical reason, where rational norms demand our first-order belief states match the way the world really is (once we know a mismatch is present). Alternatively, it may be a belief report is actually a decision that is the result of utility-sensitive Bayesian updating, and this decision conforms to the norms of practical reason because it is uncontroversial that *decisions* should take into account doxastic and utility considerations (Maher, 1993; Von Neumann & Morgenstern, 2007; Wallace, 2020). Both of these cases are realizable in the computational models we described above.

In Table 1, we summarize the assumptions and consequences of the Bayesian decision-theoretic framework outlined above.

Table 1
Summary of assumptions and consequences of the Bayesian Decision-Theoretic Framework.

Assumptions	Consequences
– The factors that shape people's beliefs include priors, likelihoods and utilities	– Bayesian decision-theoretic framework generates belief reports as a function of these three factors
– Utility is computed separately from posteriors given data	– Representations of uncertainty are separated from utility calculations
– Directional motivations can impact evidence representations directly or indirectly	– Directional motivations can violate norms of theoretical reason or conform with norms of practical reason

Model simulations distinguishing motivated reasoning from practical rationality in a toy experiment

How can researchers design experiments to study motivated reasoning in light of the Bayesian decision-theoretic framework? The details of researchers' questions will dictate the answer, but we think it's still possible to offer some general guidance.

One way to measure the impact of directional goals on the sampling of evidence and the construction of a second-order belief is to (1) sequentially present evidence for a hypothesis and (2) ask participants about their memory for the amount of evidence which supports one of two hypotheses, where a distinguishing feature of one hypothesis is it yields a reward. There are established designs in psychophysics where the experimenter systematically varies decision thresholds by assigning different reward-values (i.e., utilities) to the outcomes (e.g., Wickens, 2001). In this design, the experimenter induces a directional goal. A virtue of inducing a directional goal is that participants will not have predetermined expectations about the likelihood of the data under each hypothesis, so the effects of participants' priors will be more easily quantified. That is, inducing a directional goal would make it easier to assume that the prior distribution is approximately flat.

In the design described above, researchers could distinguish practical reasoning from motivated reasoning based on inferences about people's decisions. For example, if their memory for the evidence is accurate (or not systematically biased towards a directional goal), but their decisions nonetheless indicate they've incorporated the utility of holding the hypotheses with the larger reward, this could indicate they are reasoning in a practically rational way. In Table 2 we list a series of features of experimental designs which both capture typical ways people form beliefs about the world and would be necessary to incorporate to distinguish motivated reasoning from practical reason.

In the remainder of this section, we spell out this proposed toy experiment further. The purpose of this experiment is to highlight how the Bayesian decision-theoretic framework differentiates motivated reasoning and practical reason in a constrained setting. Our goal is not to provide empirical evidence that—as a matter of fact—people are *reliably* practically rational rather than engaged in motivated reasoning, or make any other similarly

Table 2
Key features of belief formation in naturalistic settings.

- People accumulate evidence about a hypothesis before they make a decision. This means one's memory of the evidence can impact what they believe.
- People might perceive evidence as being consistent or inconsistent with what they want to believe. Evidence can also be ambiguous under hypotheses they are considering. Either way, people can make decisions despite their level of uncertainty.
- Individual pieces of evidence are not necessarily definitive. That is, they may support a hypothesis, but rarely outright confirm or reject a hypothesis.
- People do not receive immediate feedback about whether their judgements are correct, but they do anticipate the consequences of assent or dissent towards a hypothesis.
- For many beliefs of interest to people, there is some utility associated with endorsing a hypothesis.
- People's directional goal for an outcome could be relatively weaker or stronger (i.e., the utility associated with a given outcome could vary), which suggests the need to investigate how different rewards impact the accuracy of people's beliefs.

broad claims. This question can only be answered with respect to a specific context once researchers have measured people's priors, their utility functions, quantified the evidence, and compared their behavior against a mathematical benchmark based on principled reasons.

Experimental setup

We built two computational models that simulate credences in situations where utilities and evidence can impact beliefs. The *motivated reasoning* model uses directional priors, which encode utility information for the hypotheses. The *Bayesian decision* model incorporates utilities after sampling credence values in an additional belief-report step based on an expected utility calculation. We run these models through a toy motivated reasoning experiment. We discuss how each model can yield a similar decision even though the underlying representations which produce this decision differ.

As we've discussed, psychologists often measure participants' prior beliefs towards a hypothesis h_i before they're presented with evidence. Alternatively, participants can be presented with base-rate information so the prior distribution is easier to quantify. In this toy experiment, our models are presented with a series of facts, some of which support a hypothesis h_i and some of which support an alternative hypothesis. We treat the number of facts (i.e., evidence count) supporting h_i as a discrete random variable, ev_{world}, sampled uniformly from values ranging from zero to five. We represent this mathematically as $ev_{world} \sim unif[0:6]$. We sample from a uniform distribution because in the imagined current experiment, the models have no prior expectations about the distributions of the evidence under each hypothesis.

In studies of motivated reasoning, participants are motivated to believe one hypothesis

over another. For example, they are motivated to believe that their political party has the right economic policy (Caddick & Rottman, 2021). We quantify this motivation using a utility function, which maps world states (e.g., the actuality of a hypothesis h_i) to utility values. We set the utility of h_i being true equal to three, $\mathbb{U}(h_i) = 3$ and the utility of $\neg h_i$, equal to one, $\mathbb{U}(\neg h_i) = 1$. The numbers themselves are somewhat arbitrary; the point is only that the perceived utility of one hypothesis being true is higher than the utility of its negation being true. Consequently, in the present toy experiment, each model's prior expectations are manipulated independently of the prior distribution – a feature that is atypical of most experiments examining motivated reasoning (e.g. Ditto & Lopez, 1992; Lord et al., 1979; Nyhan et al., 2014), but a feature which simplifies our ability to quantify the unique impact of both the prior distribution and the utility function.

Assume a reasoner is deciding between reporting they believe a hypothesis H or not H (denoted as $\neg H$). Both hypotheses have some probability of being correct. If the reasoner reports the correct hypothesis, they get a reward. The probability and reward structure for these hypotheses follows:

$$H : (p_H = 0.5, u_H = 3)$$
$$\neg H : (p_{\neg H} = 0.5, u_{\neg H} = 1)$$

Assume the reasoner encounters a discrete count of evidence d_i for H being true. What should their posterior credence in H be? Rational choice theory says agents should maximize their expected utility, but what are the agent's internal credences? The two accounts we discuss differ in how d_i impacts credence in H, and provides competeting answers. The Bayesian decision model incorporate utilities after first-order beliefs are constructed in a second-order inference step. For this example, we define the Bayesian-decision prior in H as the normative prior, or simply, p_H. In the motivated reasoning model, priors are directed by the perceived utility of a hypothesis.

There are many ways to model how utilities could influence the prior distribution, but we will consider the simplest possible case, where P^* equals the expected relative utility:

$$\mathbb{P}^*(H) = \frac{p_H u_H}{p_H u_H + p_{\neg H} u_{\neg H}} = \frac{0.5 \times 3}{0.5 \times 3 + 0.5 \times 1} = \frac{3}{4} \tag{12}$$

The directional prior in $\neg H$ is simply $1 - \mathbb{P}(H)$, or its relative utility $z(\neg H)$. In this simple case, increasing the utility of reporting H linearly updates – or directs – the prior credence in H.

The expectation of $\beta(a, b)$ is $a/(a + b)$. Consequently, setting $a_{prior} = u_H$ and $b_{prior} = u_{\neg H}$ allows us to define a probability distribution over credences with the desired expected value. This modeling choice allows us to not only include uncertainty when sampling credence, but also demonstrates a path for developing statistical models that can be fit to behavioral data. Directly equating a and b parameters to utilities is only possible when utilities are integers (the β distribution takes integers as inputs) and when the normative prior is uniform. We will not cover here how we might solve systems of equations for finding a and b values for other utilities and priors.

While this model may be limited, it serves as a useful illustrative example. To help gain an intuition for why instantiating a directional prior as a β distribution may make sense, consider the shape of the distribution for different combinations of a and b. When $a = b = 1$,

the utilities are equal, and the β is uniformly distributed. When $a < b$, $u_H > u_{\neg H}$, and credences favoring H are more probable. Conversely, when $a > b$, $u_H < u_{\neg H}$, and credences favoring $\neg H$ are more probable. In the Bayesian-decision model $a_{prior} = b_{prior} = 1$.

Because the β distribution is conjugate to the Binomial distribution, we can update beliefs using Bayes rule analytically given a discrete count of evidence i (out of n trials) for H. By updating the parameters of the β distribution to $a_{posterior} = i$ and $b_{posterior} = n-i+1$, we can then sample a posterior credence (first-order belief).

Results from toy experiment

In Figure 3 and Figure 4, we simulate predictions for the motivated reasoning model and the Bayesian decision model. These simulations show that the motivated reasoning model has a boost in credence for H for each evidence count. In contrast, in the Bayesian decision model, because utilities affect second-order beliefs (i.e., after a credence is constructed), the credence function is normative. Stated another way, utilities do not impact credences directly. Divergences from this line shows how much a utility impacts a credence at a given evidence count.

After a credence is constructed, we assess an output against the norms of practical rather than theoretical rationality. In behavioral tasks, psychologists measure second-order, belief reports which represent a combination of both a credence and a decision to assent to a hypothesis (Maher, 1993). The Bayesian-decision theoretic model describes how internal credences produce belief reports. In Figure 4, we plot the probability of deciding on H (the higher utility hypothesis), which shows that the utility of the hypothesis affects the model's decision to choose it. Even though the probability of choosing a hypothesis varies as a function of utility, as shown in Figure 3, the internal credence—the posterior probability representing the model's first-order belief—is unaffected. These simulations show that a Bayesian decision model can generate credences which conform to the norms of theoretical reasoning while executing decisions that can, on the surface, look like prototypical instances of motivated reasoning. Thus, we see that the model's credences can diverge from its belief reports, a situation which can be practically rational when the evidence for the hypotheses are uncertain, but a reasoner needs to make a decision (Maher, 1993).

These models can therefore serve as a framework for distinguishing motivated reasoning and practical reason. As we discuss in the ***section on implications for intervention development***, distinguishing cases of motivated reasoning from practical reason can affect how we choose to correct misconceptions. But as should be clear, the ability to distinguish these cases requires measuring and quantifying the prior distribution, the evidence participants are considering, and the utilities people assign to accuracy and directionality. As we discuss in the next section, the bulk of the previous research on motivated reasoning doesn't measure these key representations which obscures the underlying cognitive processes.

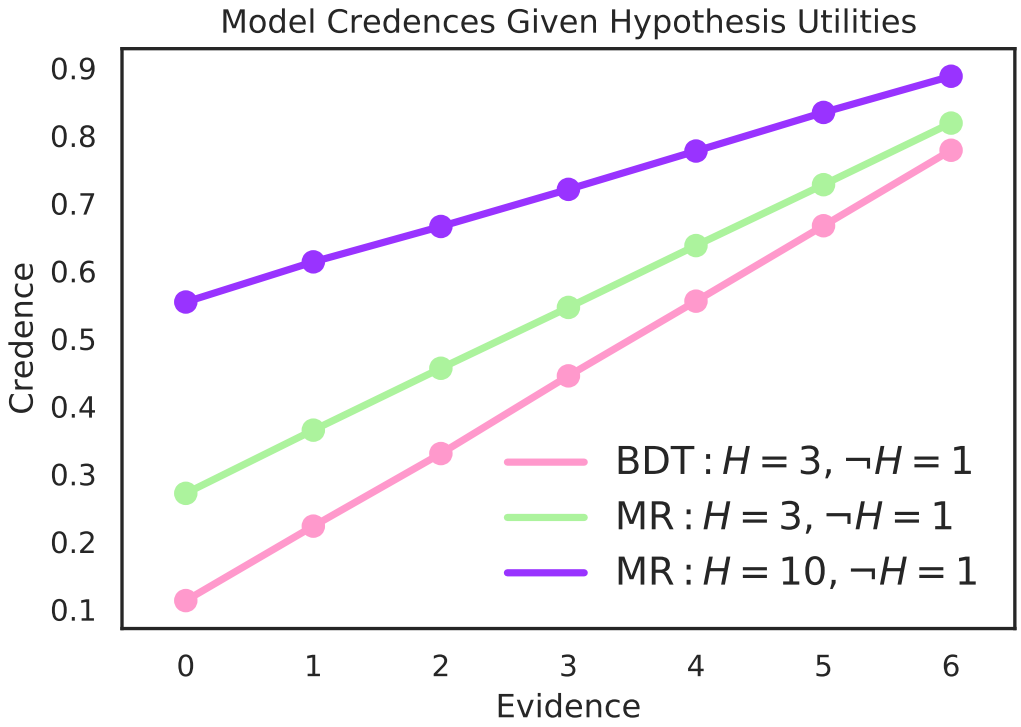

Figure 3. Credences sampled from three models. Two motivated reasoning (MR) models and one Bayesian-decision theoretic (BDT) model. The motivated reasoning models sample credences in favor of higher utility hypotheses when compared to the (normative) Bayesian decision model. In the MR model, as the $u(H)$ grows, so does its credence.

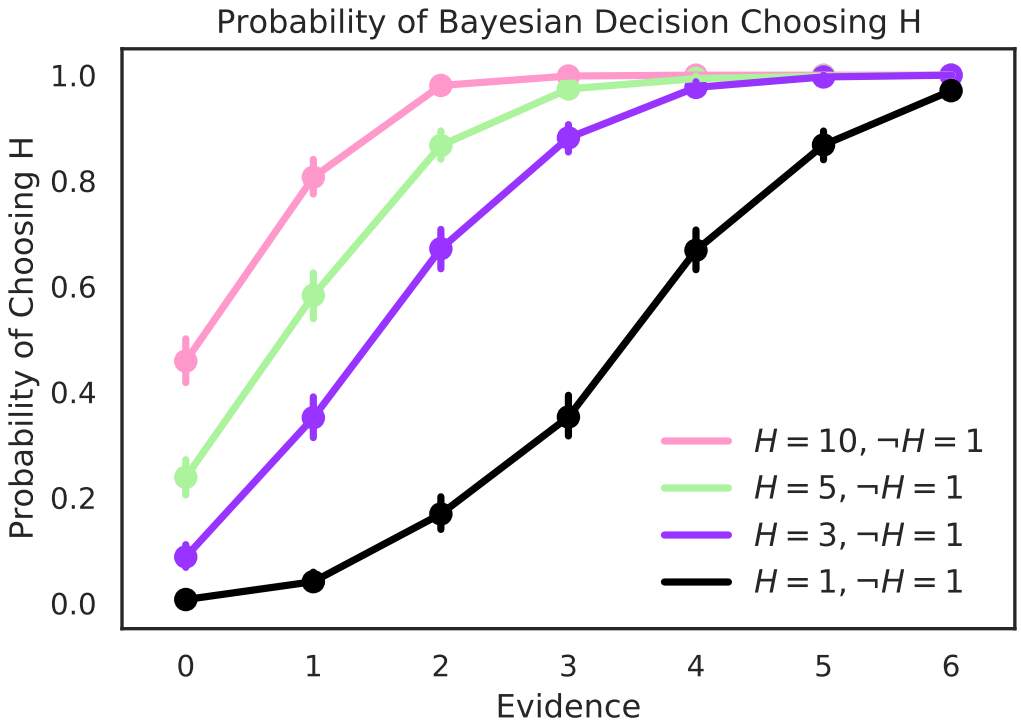

Figure 4. The probability the Bayesian decision model chooses the higher-valued hypothesis given utility and evidence threshold.

Part III
Implications of the Bayesian decision framework

Measurement implications: A new perspective on measurement of motivated reasoning

In this section, we aim to provide a snapshot of what researchers studying motivated reasoning have typically measured to guide the development and modeling of future studies examining motivated reasoning. Let's start by making a few general claims about work on motivated reasoning. One claim in the enormous and varied literature on motivated reasoning is that motivations exert "direct effects" on people's ability to represent information for or against their beliefs, and as a consequence, people fail to update their beliefs as they should (e.g., Hart & Nisbet, 2012; Kunda, 1990; Little, 2021; Taber & Lodge, 2006). It is worth noting that much of the research on motivated reasoning has been conducted in social psychology, which primarily relies on verbal theories which examine the relation between the inputs (data) and outputs (belief reports), while treating the processing that links the two as a difficult-to-quantify black box (but see Austerweil & Griffiths, 2011; Cook & Lewandowsky, 2016; Jansen, Rafferty, & Griffiths, 2021; Jern et al., 2014; Little, 2021). These links are sometimes instantiated in mediation or structural equation models, and researchers make informed conjectures on the basis of these models, but we think it is fair to say the conjectures rarely rise to the level of making predictions which could be compared against a computational model. We are going to use our framework to discuss how their inferences about the effects of motivation on belief updating have not been adequately tested, in part, because what has been measured in their experimental designs and the modeling strategies used in this research preclude the possibility of doing so (Little, 2021).

We will consider a series of experiments where participants were presented a fictitious health outcome that had positive or negative health implications depending on the experimental manipulation (Ditto & Lopez, 1992; Ditto et al., 2003, 1998). We alluded to this series of experiments in the introduction of the paper. Ditto and colleagues (1992; 2003; 1998) told participants they were testing their saliva for a "TAA enzyme," which was supposedly linked to favorable or unfavorable health outcomes. In the TAA-negative condition, participants were told that having this enzyme would make it less likely for them to develop pancreatitis (a favorable health outcome), whereas in the TAA-positive condition participants were told they were more susceptible to pancreatitis (an unfavorable health outcome). The experimental procedure included asking participants to put their saliva on a color-changing strip of paper which tested for the presence of a TAA deficiency.

In one version of this experimental setup, participants were provided base-rate information about the prevalence of TAA in the general population. This manipulation entailed telling participants base-rate information either before or after they self-administered the enzyme test. In the high probability condition, 1 in 3 (33%) people had the TAA enzyme whereas in the low probability condition, 1 in 20 (5%) had the enzyme. The authors found that the base-rate information (which they interpreted as the prior probability) impacted participants' beliefs in expected directions — those in the high probability condition reported greater likelihood of having the enzyme and vice versa. However, participants who believed TAA enzymes to be deleterious to their health reported that it was less likely they had the enzyme than participants who believed TAA enzymes would lead to favorable health outcomes. This finding suggested the impact of directional motivation on people's beliefs because participants interpreted the outcome of the TAA-deficiency test to maintain

the belief that they are, in fact, healthy.

In another version of the experimental setup, the researchers manipulated the likelihood of false positives. When describing the test, participants were told that unusually high or low blood sugar levels could affect the reliability of the test. Participants in the TAA-negative condition showed no sensitivity to the probability of blood sugar as an alternative explanation for their test result, but participants in the TAA-positive condition did. The use of this alternative explanation (Powell, Weisman, & Markman, 2022) in some cases but not in others again suggests participants were engaged in motivated reasoning.

How should we understand these results, including what's been measured, in the context of a Bayesian decision-theoretic framework? First, we see that although participants were given base-rate information, the sample was primarily undergraduate students. While this is not a unique problem to this series of studies, it presents a unique problem in the current experiment. For example, undergraduate students are mostly healthy and thus might correctly balk at the purported evidence that is inconsistent with what they know about their health. Namely, they are young and thus very likely to be healthy. Consequently, it would be surprising to find out they were secretly unhealthy. In effect, although base-rate information has been communicated to participants, this isn't the *entirety* of the prior which would inform how one would expect participants to interpret data they are presented with. For example, the authors would also need to measure the participants' perceptions of their own health, in addition with the base-rate information, to quantify the extent to which the participants' posterior reflects the evidence they are provided, base-rate information, and the more complete prior distribution.

Thus far, we've focused on how differences in an undergraduate's prior in their own health could correctly impact their interpretation of a test result. For simplicity, we will assume that the amount of evidence provided by the test itself is quantifiable, and so under the Bayesian decision-theoretic framework we laid out above, the remaining open question is how the *utility* of holding a certain belief could alter the interpretation of the results of the studies.

It is clear in these experiments that the utility functions over "being in good health" or "possibly developing pancreatitis" have gone unmeasured. This fact alone makes it difficult to determine the extent to which the utility has been inappropriately incorporated into the decision to discount the test result. Belief reports are decisions. To evaluate these decisions we need to compare perceived utilities of the outcomes scaled by the posterior distribution against a mathematical benchmark. Can we infer these utilities? People differ in the utility they ascribe to being healthy: Some people smoke and some do not, despite the evidence that smoking is a strong predictor of poor health outcomes. Some people exercise and eat nutritious food, and some do not. The point is the utilities about *how healthy one cares to be* vary, which reflects the utility function over that outcome. On our account, this must be measured to understand how and whether utilities are changing the way evidence is sampled or being rationally integrated into a belief report.

Our aim here is not to single out any particular set of research, but rather demonstrate how measurement of specific information is necessary to understand the underlying mechanisms producing a belief report. We've argued psychologists need to measure people's prior beliefs, how people weigh the evidence presented (and this needs to be quantified), and the utility they assign to holding certain beliefs. In our minds, it is exceedingly likely

people engage in pernicious forms of motivated reasoning that lead them to sample from the relevant evidence inappropriately, and the above studies may well *suggest* this is true as well. However, to more firmly establish this claim, we need to measure and precisely quantify several key ingredients to distinguish cases of motivated reasoning as they have been traditionally construed from cases where participants show practically rational sensitivity to the utility of holding a particular belief.

In the next section, we go beyond considering a handful of cases. We summarize what's been measured in a larger number of papers on motivated reasoning from the past four decades (1970-2022). To be clear, an exhaustive meta-analysis of work on motivated reasoning is not the main focus here. Rather, we hope that this snapshot of the current state of the motivated reasoning literature can help shape future work in the field.

A snapshot of what is measured in research on motivated reasoning

We conducted two rounds of searches on Google Scholar, the first used keywords such as "motivated reasoning" or "motivated cognition", "belief polarization" and "backfire effects" or "boomerang effects" to find the most widely cited experiments across various fields (e.g., social psychology, cognitive science, economics, communication, political science, marketing) in peer-reviewed journals between the years 1990-2020. The second search used keywords like "persuasion" and "biased reasoning" for papers prior to 1990, because researchers more typically used this terminology then.

For both searches, our inclusion criteria was that a paper must have one or more experiments to test the presence of motivated reasoning. We did not set any explicit benchmarks for the number of participants, research area, and type of research design. We created a summary table which included author names, paper titles, primary findings, authors' explanations for their findings, and any additional relevant information for each study. This information is located in the supplement. Of central interest was what most research on motivated reasoning has measured. Under the Bayesian decision-theoretic framework, researchers need to quantify prior beliefs, evidence, and the utility associated with accuracy and directional goals to distinguish motivated reasoning from practically rational behavior.

We operationalized priors across these studies as the beliefs held by people before participating in the experiment (e.g., pretest beliefs). We conceptualized the quantification of evidence as the numeric weight assigned to the evidence presented within each experiment. For instance, we marked an experiment as having quantified evidence if participants were presented evidence in the form of statistical data or percentages (e.g., Baekgaard, Christensen, Dahlmann, Mathiasen, & Petersen, 2019). Accuracy goals motivate people to seek out relevant evidence to form a true belief, whereas directional goals motivate people to evaluate evidence to support existing beliefs and perspectives. We noted these constructs as having been measured if the researchers either measured or manipulated constructs that even roughly *approximated* utility. To be maximally charitable in our coding of what researchers measured, we marked a paper as "successfully" measuring a given parameter if in any of the tasks, experiments, or conditions claimed the parameter to have been measured.

To be clear from the outset, we are not attempting to argue that purported cases of motivated reasoning are in fact instances of practically rational Bayesian decision making. Instead, Table 3 is meant to help assess whether current literature can test this question at all.

Table 3

A summary of whether utility related to accuracy and direction goals, prior beliefs, and evidence presented were quantified in prior research on motivated reasoning.

Authors	**Year**	**Accuracy Utility**	**Directional Utility**	**Prior**	**Evidence**
Cialdini, Braver, and Lewis	1974			×	
Hass	1975				
Hass and Mann	1976				
Lord et al.	1979				
Petty and Cacioppo	1979				
Petty, Cacioppo, and Goldman	1981				
Hunt, Domzal, and Kernan	1982				
Heesacker, Petty, and Cacioppo	1983				
Petty and Cacioppo	1984				
Wu and Shaffer	1987				
DeBono and Harnish	1988				
Sanitioso, Kunda, and Fong	1990				
Sharma	1990				
Ditto and Lopez	1992				
Bickart	1993				
Babcock and Loewenstein	1997	×			
Buehler, Griffin, and MacDonald	1997				
Ditto et al.	1998				
Fischle	2000			×	
Redlawsk	2002			Flat	
Ditto et al.	2003				
Taber and Lodge	2006			×	
Taber et al.	2009			×	
Nyhan and Reifler	2010		×		
Hart and Nisbet	2011				
Kahan	2013			×	
Bolsen, Druckman, and Cook	2014	×	×		

Table 3

A summary of whether utility related to accuracy and direction goals, prior beliefs, and evidence presented were quantified in prior research on motivated reasoning.

Authors	**Year**	**Accuracy Utility**	**Directional Utility**	**Prior**	**Evidence**
Horne, Powell, Hummel, and Holyoak	2015				
Horne, Powell, and Hummel	2015			×	
Peter and Koch	2016			×	
Drummond and Fischhoff	2017				
Kahne and Bowyer	2017	×	×	×	×
James and Van Ryzin	2017			×	×
Pasek	2018		×	×	
Baekgaard et al.	2019			×	×
Dixon et al.	2019			×	
Pennycook and Rand	2019				
Wood and Porter	2019				
Tangari, Bui, Haws, and Liu	2019				
M. Stanley, Henne, Yang, and De Brigard	2020			×	
Bayes, Druckman, Goods, and Molden	2020	×	×		
Guess and Coppock	2020			×	
Christensen and Moynihan	2020		×	×	×
Madson and Hillygus	2020			×	×
Porumbescu, Moynihan, Anastasopoulos, and Olsen	2020		×		×
Horne, Rottman, and Lawrence	2021			×	
Molleman, Gradassi, Sultan, and van den Bos	2021	×	×	×	×
Persson, Andersson, Koppel, Västfjäll, and Tinghög	2021		×		×
Painter and Fernandes	2022			×	

Note: Studies were given a mark if *any* of the tasks, experiments, or conditions in the paper quantified utility, priors, or evidence, respectively.

Table 3 provides a snapshot of the factors that are often measured in research on motivated reasoning. We're reluctant to draw firm conclusions based on this review alone, but we can conjecture that it is rare for papers to measure all of the parameters in our framework (i.e., prior, evidence, utilities). It also appears that researchers measure people's priors and the evidence more frequently now than they once did. Consequently, against the backdrop of our framework, the current state of what's measured in the literature and their findings will make it difficult to distinguish not only Bayesian updating from motivated reasoning (Jern et al., 2014; Little, 2021), but also practical rationality from motivated reasoning. This snapshot indicates that more needs to be measured to understand when and where people are engaging in motivated reasoning.

Guidance on measuring priors and utilities

We should acknowledge that it's an open question how to measure each of the parameters in our framework. Entire psychometric careers are made on trying to pin down the best ways to capture people's priors in a task, or attempting to measure people's utility functions. Researchers have made progress on these questions, but an extensive discussion of each of these issues is beyond the scope of the paper. Nonetheless, we'll point the reader to sets of papers which have addressed the estimation of prior distributions and utility functions. We leave it up to researchers to determine how best to quantify the evidence they present participants.

Measuring a prior distribution. Psychologists have typically measured prior distributions in one of two ways. First, they measure pretest attitudes (or a network of pretest attitudes) in a longitudinal design using Likert scales (Horne, Powell, Hummel, & Holyoak, 2015; Miske, Schweitzer, & Horne, 2019; Powell, Weisman, & Markman, 2018). Second, they provide participants with base-rate information. These are reasonable starting places for measuring a prior distribution, but some care needs to be exercised. For example, beliefs are often related to each other coherently, so it's likely researchers need to measure more than a single focal belief (Jern et al., 2014; Powell et al., 2018). It's been demonstrated that, in some cases, people update their beliefs more optimally than it may initially appear once we understand how a network of beliefs is structured (e.g. Cook & Lewandowsky, 2016; Gershman, 2018; Jern et al., 2014; Powell et al., 2018). Likewise, providing base-rate information could be sufficient so long as this exhausts the information participants bring to bear in the study; that is, people have no other unmeasured expectations about the likelihood of the data in the task. A third, though less typically relied on method in research on motivated reasoning, is to create a task where researchers can assume a participant's prior distribution is flat. This might be possible by inducing a motivated reasoning context rather than studying a domain where a prior distribution will be necessarily strong (e.g., vaccines or climate change; Cook & Lewandowsky, 2016; Kahan et al., 2017; Nyhan et al., 2014). This was the approach we proposed in the toy experiment described in the modeling section of the paper. Whatever the strategy a researcher chooses to employ, for researchers to compare behavior against a mathematical benchmark, we need to explicitly quantify the prior distribution and this could be comparatively more difficult using different measurement strategies (e.g., relying on Likert scales vs. providing participants with base-rate information). Finally, we must acknowledge that there are longstanding disputes about how to measure or specify a prior distribution – known as the "problem of priors" (e.g.,

Easwaran, 2011). We cannot overlook these issues when choosing a measurement strategy.

Measuring a utility function. People assign different utilities to different outcomes. For instance, if we place a bet in a game of chance and win, all else being equal, this outcome has higher utility than if we placed a bet and lost. The utilities people associate with the same consequences also vary – for instance, an ultra-wealthy person might be approximately indifferent to winning a bet whereas we would not (academics are not ultra-wealthy). Economists and psychologists have measured utility functions in several ways, but they bear some important similarities (see Alchian, 1953; Farquhar, 1984). For example, economists have measured preferences by having participants compare hypothetical choices, including choices where the certainty of the outcome is manipulated (Halter & Mason, 1978; Mellers, Schwartz, Ho, & Ritov, 1997; Robison, 1982). A key goal across these measurement approaches is the estimation of *indifference points*, where participants are asked to modify their responses until they indicate indifference between the given alternatives (e.g., Luce, 1991). Economists have developed multiple methods for determining participants' indifference curves. First, the *probabilistic equivalence* method is a method where researchers ask a decision maker to determine the probability for which they are indifferent to obtaining a certain outcome given two alternatives. Second, the *value equivalence method* is where researchers ask a decision maker to modify the value they assign to one of two outcomes until they are indifferent about the outcomes. Third, the *certainty equivalence method* is where researchers ask a decision maker to provide the amount of payoff they are certain would make them indifferent between two alternatives (Hershey & Schoemaker, 1985).

We should also consider several factors that can impact participants' preferences (see Halter & Mason, 1978). For instance, an estimate of risk aversion can be informative as some risk properties indicate the implausibility of certain utility functions (e.g., Farquhar, 1984; Tversky & Kahneman, 1974). Other issues in utility measurement stem from researchers using descriptive theories of decision-making without first testing the validity of these theories. For instance, contextual differences can affect people's preferences for certain outcomes and people can display irregularities in their choices towards outcomes (see Birnbaum, 1992; Regenwetter et al., 2011). Researchers conducting a preliminary analysis of participants' utility functions in a new domain might begin with questions about monotonicity, boundedness, differentiability, continuity, risk attitudes, and other properties of utility functions (Bleichrodt, Abellan-Perpiñan, Pinto-Prades, & Mendez-Martinez, 2007; Farquhar, 1984).

The cases we've been particularly focused on are situations where people see the formation of a belief as having a kind of utility (Maher, 1993); whether it means support from one's in-group or recognizing the implications of adopting a belief for one's future behavior (e.g., I think climate change is exacerbated by human activities and realize that means I should take fewer international flights). While we don't think it's controversial that adopting beliefs can have utilities, historically, behavioral economists and psychologists have not measured these sorts of utilities as often. Still, techniques behavioral economists have typically used to estimate utility functions can be adapted to assess the utilities of forming a belief. For example, Botzen and colleagues (2012) measured risk beliefs and the demand for low-probability, high-impact flood insurance, and found that people did not behave according to a traditional expected utility model. Here, the researchers modeled participants' beliefs about their homes' sensitivity to flood damage and indirectly measured

their utility function using a willingness-to-pay measure (in this case, their willingness to buy flood insurance). Similar measurement strategies could be adopted for beliefs with applied outcomes (e.g., vaccine uptake), but also more abstract outcomes. For instance, a researcher could measure people's perceived utility of secrets being disclosed to one's friends (Slepian, 2022), or of acquiring new information more generally (Falk & Zimmermann, 2016; Golman & Loewenstein, 2018).

We can see that there are several decision points when measuring a prior distribution and estimating the perceived utility of holding a belief. The appropriateness of any of the methods is context and topic dependent. Nonetheless, to resolve where in the reasoning process directional goals influence people's beliefs, we need to measure these parameters, a difficult task but one which has seen substantial psychological advances.

This section highlighted a set of implications for the measurement of motivated reasoning and provided us with a snapshot of the current state of research on the topic. In the next section, we discuss further implications of the Bayesian decision-theoretic framework, ones which we think have been particularly overlooked by modelers providing alternative, rational accounts of motivated reasoning – the development of interventions aimed at correcting misconceptions.

Implications for intervention development

Motivated reasoning, as we've defined it, is a situation where evidence representations are biased by directional-goals. Researchers often appear to assume evidence representations must be distorted by directional-goals when participants' belief reports do not shift in ways they seemingly ought to following intervention. On its face, this suggestion is plausible because across many experiments, participants appear to hold onto their beliefs in the face of evidence contrary to what they think. However, this research often leaves many of the core representations unmeasured (see Table 3), and is thus unable to distinguish between situations where people may have properly integrated the evidence *before* taking the utility of holding that belief into account from situations in which motivations directly impacts how people sample from the evidence. These are distinct situations that implicate different sets of cognitive processes, which may demand altogether different kinds of interventions.

The distinction between practically rational behavior and motivated reasoning is not merely a semantic issue but impacts how we may develop interventions that correct misconceptions. For example, it may well be that someone respects the norms of practical reason and nonetheless believes that climate change is a hoax. Thus, even if a person's reasoning is rational under some set of norms, there could be a need for intervention as a matter of public policy. As a consequence, effective interventions may take very different forms when misconceptions are due to defects in one's inferential machinery (i.e., so called "first-order interventions") versus when misconceptions arise because of second-order influences from directional goals (i.e., so called "second-order interventions"). The goal of this section is to explore how the Bayesian decision-theoretic framework may be applied to distinguish between these two cases in order to guide intervention development.

We distinguish between two classes of interventions, first-order and second-order interventions. First-order interventions aim to revise inputs to the belief-decision process, typically (but not only) focusing on doxastic factors. For example, they could take the form of a nudge to encourage participants to more accurately encode information (Bago, Rand, & Pennycook, 2020; Pennycook & Rand, 2019), or by increasing trust in established bodies of knowledge (Van der Linden, Leiserowitz, & Maibach, 2019; Van der Linden, Leiserowitz, Rosenthal, & Maibach, 2017). First-order interventions operate on first-order beliefs and aim to correct the alignment between internal credences and the way the world really is. However, they need not focus exclusively on doxastic factors. For instance, interventions highlighting utilities may incentivize reasoners to use accuracy goals when interpreting evidence during reasoning (Kunda, 1990). This is a case where a utility focused information is designed to revise inputs to a doxastic representation.

Second-order interventions, on the other hand, target second-order beliefs (and consequently, belief reports), often by revising people's perceived utilities associated with holding a belief (although, second-order interventions need not always operate on utilities). For example, education researchers have found that stating the real-world utility associated with learning a piece of information is a strong driver of students' eagerness to learn and remember that information (Hulleman & Harackiewicz, 2009; Soicher & Becker-Blease, 2020). Second-order, utility focused interventions highlight the empirical consequences of a reasoner's decision to hold a belief, and thus are designed to realign second-order beliefs *after* doxastic information has been integrated. Second-order interventions aim to make salient

how decisions to hold certain beliefs can facilitate achieving one's goals.

First-order interventions are tested more frequently than second-order interventions. The research paradigms most typically used in social psychology partially explain this trend: Psychologists developing interventions often only measure first-order doxastic representations, focusing on how motivation biases sampling during the construction of an initial credence. However, in cases where first-order interventions fail to shift people's beliefs, it is nonetheless possible there are other means of shifting people's beliefs (Priniski & Horne, 2019). For instance, a more effective intervention might be developed by appreciating the role second-order utility plays in affecting belief reports (Kahan & Braman, 2006; Kahan et al., 2017).

To illustrate how the Bayesian decision-theoretic framework can assist researchers in developing second-order, utility-value interventions, we will discuss two frequently-cited first-order interventions, *accuracy nudges* and *inoculation tactics*, which intervene on the separate processes of evidence representation and evidence integration, respectively. We will then discuss how they may be adapted to target second-order, directional-goals.

Example 1: Enhancing Accuracy-nudge paradigms with second-order utility information. Nudges are a popular approach in the decision sciences literature for influencing people's behavior and beliefs (Thaler & Sunstein, 2009). Recently, many researchers have been interested in applying nudges to mitigate the spread and uptake of misinformation (e.g. Bago et al., 2020; Bronstein, Pennycook, Bear, Rand, & Cannon, 2019; Fazio, 2020; Fazio, Rand, & Pennycook, 2019; Pennycook et al., 2021; Pennycook, McPhetres, Zhang, Lu, & Rand, 2020; Pennycook & Rand, 2019, 2020, 2021). So called "accuracy-nudges" are designed to get people to think more carefully about the accuracy of a piece of information before engaging with it (e.g., updating their beliefs, sharing the information on social media). The central idea is that a reasoner's reflective, and deliberative "System 2" processes are better suited at detecting a piece of information's veracity than their effortless, intuitive "System 1" processes (Bago & De Neys, 2017, 2019; Pennycook & Rand, 2019; Thompson, Turner, & Pennycook, 2011). Accuracy nudges are designed to subtly nudge people to use System 2 processing when interpreting potentially false information.

There are various ways these nudges may be instantiated. For example, they could require participants to explain why a headline is true or false before gathering accuracy perceptions (Fazio, 2020). Other more implicit tacts have been taken, like varying the time participants are allotted to decide if they would share a piece of information on social media (Bago et al., 2020). Some researchers have used direct methods to nudge people towards accuracy, like reminding participants at the beginning of a misinformation detection task to consider accuracy (Pennycook et al., 2020). The authors of many of these papers report that these effects can inform us about the underlying mechanics of reasoning and belief revision (Brashier, Pennycook, Berinsky, & Rand, 2021; Pennycook et al., 2021; Pennycook & Rand, 2019). In particular, a consistent claim across many of these studies is that because political affiliation does not correlate with participants' ability to detect or share misinformation, and accuracy nudges do not interact with political affiliation, reasoners are lazy rather than "motivated" when engaging with potentially false information (e.g., Pennycook et al., 2021, 2020; Pennycook & Rand, 2019).[5]

[5]This point is contentious, however (Roozenbeek, Freeman, & van der Linden, 2021). For instance, a

However, there are a few things worth considering when interpreting the results and assessing the viability of accuracy-nudge paradigm. The first thing to consider is that the effects of accuracy nudges (and nudges more generally) are small (e.g., correlations between responses on the Cognitive Reflection Task and ability to detect false information range from 0.1 to about 0.25; Pennycook & Rand, 2019). A recent meta-analysis found that nudge-based approaches are likely to exert extremely small effects once publication bias is accounted for (Maier et al., 2022). It is an empirical question, but the magnitude of these effect sizes may also suggest that the source of the tendency to share misinformation is not predominantly a function of a problem with people's doxastic representations. Instead, it is possible practically rational considerations are shaping people's decisions. Thus, it might be fruitful to use other equally-scalable intervention tactics that leverage the perceived utility of a belief report.

The second concern focuses on researchers' failure to measure constructs that resolve how exactly accuracy-nudges are supposed to work, and to what extent they will generalize beyond the domains they have investigated. A considerable body of research on accuracy nudges analyzes behavior involving news headlines, and consequently aims to measure representations relevant to how news headlines are interpreted and shared (e.g., Brashier et al., 2021; Pennycook et al., 2021, 2020; Pennycook & Rand, 2019, 2020). For instance, while these studies often include a measure of political ideology (which can be seen as a broad correlate of prior beliefs), they also lack precise measurement of a participant's priors for a specific belief at hand (but of course, there are exceptions; see Pennycook et al., 2020). A reasoner's prior belief about the information in a specific news headline should have considerable influence not only on their perceptions of accuracy, but also on their intention to share that information.[6] There is sure to be substantial item-level variation, because the perception of a news headline will interact with a participant's unique prior belief about that issue. This is not accounted for in many of the current designs and statistical models in the accuracy nudge literature. Further, most authors have failed to measure utility information at all. For example, group-based utility information may change how news headlines are interpreted (e.g., how do members of my group interact with this type of information; Kahan & Braman, 2006). After all, the very reason *sharing* is of interest to researchers is because of the assumption that information shared by one's in-group will beget further sharing. Consequently, it is possible that deploying utility-value information as an intervention could increase the efficacy of nudges that have predominantly focused on accuracy, but have not yielded particularly large behavioral effects (Bago & De Neys, 2017; Pennycook et al., 2020; Pennycook & Rand, 2019; Roozenbeek et al., 2021). To make this more explicit, we will now discuss how researchers could design an experiment that incorporates utility information into an accuracy-nudge paradigm.

Incorporating second-order information in accuracy-nudge paradigms. People often read misinformation in online environments where interactions between people's social networks can serve as a rich source of second-order utility information. Consequently, making salient how a piece of information was engaged with in one's social networks

motivation for accuracy is a motivation nonetheless (Kunda, 1990).

[6]Although perceived accuracy and intention to share may be weakly correlated in some situations, reasoners might still willingly share misinformation (e.g., Altay, de Araujo, & Mercier, 2022; Chen, 2016; Chen, Sin, Theng, & Lee, 2015; Laato, Islam, Islam, & Whelan, 2020; Madrid-Morales et al., 2020)

could help the reader assess the social costs associated with sharing a piece of information. For example, researchers could manipulate the social consequences associated with sharing a piece of potentially false information, which could be in the form of predicted engagement with a piece of information. In this imagined experiment, a high-group support condition could inform a participant that in-group members are more likely to share and favorite a headline based on past sharing behavior of friends in their network. Researchers can then quantify these effects by varying engagement metrics to more completely define a utility space (e.g., one retweet and one "like" is a smaller group-based reward than 50 retweets, 50 likes, and five new followers). Researchers could then compare participants' intentions to share and perceptions of accuracy to participants in a low-group support condition – a condition in which participants are told that in-group members are unlikely to share and favorite a given headline, or are more likely to unfollow, mute, or block people who have shared similar content. To control for the influence of trust (a first-order, doxastic consideration), a researcher could also measure how engagement metric information shifts people's evaluation of the evidence itself, allowing them to observe the unique effects of second-order, utility-value information.

Misinformation may encourage misperceptions of consensus – false beliefs about what one's group uniformly agrees to believe. These misperceptions can influence people to report beliefs that are possibly out of sync with their internal credences, in turn establishing a social norm that prevents people from updating their reported beliefs about polarizing social issues. For instance, nearly three-fourths of Americans support climate change mitigation policies, but the general public believes that the proportion is closer to one-third (Sparkman, Geiger, & Weber, 2022). Many people—specifically conservatives—may therefore believe that climate change is real and would support policies designed to deal with it, but fear potential social consequences of reporting so. Assuring conservatives that there would be little-to-no social consequence in reporting support for climate change policies (because, in fact, the majority of their party also believes climate change is real and would support mitigation policies), can be one way to realign people's first-order and second-order beliefs (e.g., Constantino et al., 2022; Lewandowsky & van der Linden, 2022).

We can now see a possible benefit of developing a computational framework for distinguishing practically rational behavior from motivated reasoning: it allows us to assess what's gone unmeasured in current studies and thus suggests a way forward in the creation of new intervention strategies. The benefits of developing our framework are not unique to research on nudges. We'll now consider a second example.

Example 2: Inoculating gateway beliefs with preventative arguments. Most people are not trained to interpret scientific evidence. People's attitudes towards topics in science hinge on factors most scientists do not view as objectively relevant to how we should understand the evidential quality of a piece of scientific research. For example, it has been established that perceptions of scientific consensus shape how people evaluate a piece of science (Cook & Lewandowsky, 2016; Imundo & Rapp, 2022; Roozenbeek, van der Linden, Goldberg, Rathje, & Lewandowsky, 2022; Van der Linden et al., 2019). "Gateway beliefs" to climate attitudes and perceptions of consensus play a pivotal role in shaping how evidence for climate change is integrated into people's belief systems. Researchers have found that Bayesian networks of climate change attitudes that include a "perception of consensus" node (such that high-trust in consensus among climate scientists predicts

trusting scientific evidence for climate change) can simulate belief polarization following exposure to evidence of climate change. Because Bayesian networks are able to capture polarization effects, researchers have concluded that the cognitive processes underlying these effects are rational (in that they do not violate the axioms of probability theory, e.g., Cook & Lewandowsky, 2016). These results suggest that climate change misconceptions persist because of inputs to people's inferential machinery rather than the operations of the machinery itself. Consistent with this claim, researchers have successfully strengthened beliefs in climate attitudes by presenting participants with "consensus arguments" that change people's gateway beliefs.

Misinformation, particularly about science, often targets gateway beliefs by casting doubt on established sources of knowledge (Enders, Uscinski, Klofstad, & Stoler, 2020; Lewandowsky, Cook, Fay, & Gignac, 2019; Pierre, 2020; J. Stanley, 2015). The effects of misinformation which undercut established sources of knowledge can have a continued influence on people's beliefs, even after correction (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). Researchers have therefore also argued for preemptive approaches to countering misinformation by designing interventions that establish "defensive priors" by exposing people to weakened forms of misinformation – a kind of "belief vaccine" that inoculates people against misinformation (Compton, van der Linden, Cook, & Basol, 2021; McGuire, 1964; Traberg, Roozenbeek, & van der Linden, 2022; Van der Linden et al., 2017; Vivion et al., 2022). For example, an inoculation intervention for climate change could contain information that preemptively refutes common arguments claiming that there is no consensus among climate scientists about the realities of human-caused climate change (Pilditch et al., 2022; Traberg et al., 2022).

In the cases described above, the interventions that have been developed have focused on first-order, typically doxastic, information to correct misconceptions (or strengthen otherwise uncertain beliefs). But there is no doubt that misconceptions about climate change persist (e.g., Brenan & Saad, 2018; Lewandowsky et al., 2012), so it is worth considering how researchers can improve on existing messaging by going beyond intervening exclusively on first-order, doxastic considerations. Below, we consider how incorporating second-order, utility-value information could augment existing interventions.

Second-order inoculation to guard against conspiracy narratives. Conspiracy theories are a common type of misinformation targeting beliefs about science. They discredit scientific findings and recommendations by linking scientists to malevolent (oftentimes political) characters (Bodner, Welch, & Brodie, 2020). During the COVID-19 pandemic, for example, conspiracies quickly emerged online – generally claiming that public health experts were conspiring alongside left-wing politicians and tech elites to exert control over the public (Douglas, 2021). Conspiracy narratives are integral to climate change denialism as well. These narratives, which are often echoed on conservative news outlets, aim to construe climate science as politically motivated individuals working with left-wing politicians to limit free-market capitalism (Hornsey, Harris, & Fielding, 2018; Uscinski, Douglas, & Lewandowsky, 2017).

One way to build a second-order intervention to guard against the effects of conspiracy theories is to consider how conspiracy theories exploit group-based biases. For instance, conspiracy narratives are often structured with an in-group bias, in that they are designed to negatively construe events (e.g., public health measures during a pandemic) by

positing that they are against the goals of the group (e.g., out-group totalitarian control). This is why conspiracy theories were so effective at inflaming political tensions during the COVID-19 pandemic (Douglas, 2021; Jolley, Douglas, & Sutton, 2018). Conspiracies leveraging in-group bias also allow for people to rationally reinterpret contradictory evidence by dismissing the source of evidence as unreliable (Gershman, 2018; Jern et al., 2014).

To explore how we can develop a second-order *inoculation* intervention, we will focus on the case of vaccination attitudes, particularly in situations where we aim to increase vaccine uptake during disease outbreaks. Researchers have found that these situations are often accompanied by misinformation (Midgon, 2022). Misinformation inoculation follows two steps: forewarning of tactics and refutation of tactics. In the forewarning phase, researchers could provide participants with information about conspiracy theories and why they're used. For example, researchers could inform participants that conspiracy theories are designed to facilitate trust in certain people. Researchers can then provide participants with examples of how conspiracy theories have been used for harm in the past (e.g., Nazi Germany, Rwandan Genocide). Thereafter, researchers can provide participants with concrete examples of how conspiracy theories were used to undermine public health efforts during the COVID-19 pandemic (e.g., how vaccines were planted with tracking devices). Researchers can then survey people's beliefs about a novel, ongoing outbreak (e.g., Monkeypox) and measure their susceptibility to the narrative offered by emerging conspiracy theories. This will allow them to test the effects of a *norms-based message* to promote more positive vaccine attitudes (e.g., see Constantino et al., 2022; Lewandowsky & van der Linden, 2022).

Inoculation tactics usually adopt a refutation step where they logically or statistically undercut a misinformation claim researchers predict participants will have been exposed to. However, refutation interventions target first-order beliefs when second-order beliefs may play a larger role than researchers have typically assumed (Jachimowicz et al., 2018). One way researchers could take second-order beliefs into account is by providing participants with icon arrays of vaccination rates of a participant's community or a participant's political party, which would establish the norms among a participant's in-group. Indeed, it's been found that participants view anecdotal information about people as similar to themselves influences vaccine attitudes (Horne, Powell, Hummel, & Holyoak, 2015). Similar effects have been observed in other domains: For example, Jachimowicz and colleagues (2018) observed that second-order normative beliefs predicted participants curbing their energy use even though participants' first-order credences did not. Under the Bayesian decision-theoretic framework, we would predict that in domains where people perceive the evidence is uncertain but the utilities favor one theory over the other, intervening on second-order, utility oriented beliefs may be a more effective intervention tactic than interventions exclusively focused on first-order, doxastic features.

Open questions and conclusion

Open questions

Utility and rationality. One open question concerns the formation of the utility representations that can impact people's belief reports. We proposed a simple utility function that can be used to compare outcomes and make a decision. We assumed a predefined utility calculus (e.g., a specified mapping of states to reward values) that guides reasoning. Although this is a common assumption in the reinforcement learning literature because the values of world-states are often assumed and used as input for an agent, our framework doesn't discuss how utility representations are computed in the first place (for a discussion of related issues, see Bostrom, 2009). It is possible that directional goals influence utility representations just as they can impact how evidence is sampled. In cases where directional goals impact people's utility function, second-order influences on doxastic states could arguably be a case of motivated reasoning.

Along similar lines, earlier we noted that irrationality with respect to theoretical reason concerns the mismatch between one's beliefs and the way the world really is. Practical irrationality, in contrast, is instead a failure to execute an action plan consistent with your utility function and knowledge of the world. In this latter case, when can we conclude a *belief report* is practically irrational? An open question is how exactly this occurs. First, we'd expect that people might have utilities that they don't correctly join with information they believe, or utility functions which deviate from rational-actor models (Loewenstein & Molnar, 2018; Tversky & Kahneman, 1991). Second, people could fail to act in ways which are compatible with the utilities themselves (like when we plan to exercise but fail to), or fail to update their utilities when they are provided with new information (Edwards, 1954; Kalis, Mojzisch, Schweizer, & Kaiser, 2008; Stetzka & Winter, 2021; Wiggins, 1978). All of these possibilities distinguish the Bayesian decision framework from revisionist accounts of theoretical rationality, like the idea of resource rationality proposed by Lieder and colleagues (2020). Under the Bayesian decision framework, irrationality could take the form of inaccurate assessment of the relevant utilities, failing to act in accordance with one's utility function, or failure to integrate credences and one's utilities appropriately. These issues are wholly distinct from (important) questions regarding how time constraints, limited cognitive resources, and the like place upper-bounds on otherwise Bayesian reasoners.

Another open question is whether practically rational belief reports could *beget* credences at odds with the norms of theoretical rationality. It's possible that initially practically rational behavior could eventually lead people to sample from the evidence improperly – that is, engage in motivated reasoning. As we've discussed, belief reports are in effect a decision to report a credence in a way that takes into account both the credence itself and the utility of forming that belief. This suggests that a belief report itself may not be a stable attitude. Instead, it is the product of perception of the environment at a particular time and context. And it is known that the perceived utility of some actions change as other states of the world change (e.g., see Holton, 1999; Meyer & Hundtofte, 2022; Portmore, 2013; Stetzka & Winter, 2021). One question then is whether an initially practically rational decision could lead to a stable credence at odds with the evidence. We think this is both possible and likely. Researchers have demonstrated that once people make a decision, they subsequently adjust their beliefs to cohere with the decision they've made (e.g.,

Simon, Pham, Le, & Holyoak, 2001; Simon, Stenstrom, & Read, 2015). Thus, to return to our prior example, we can imagine that while a participant initially is well-calibrated to the evidence, their decision to report a belief in accordance with its utility could subsequently lead them—perhaps unknowingly—to shift their beliefs about the strength of the evidence for the theory they've reported believing. In this way, initially practical rational behavior could produce internal credences which no longer reflect the facts about the world.

The idea that people's decisions can impact their stable credences could have beneficial consequences as well. For example, imagine we've implemented a second-order, utility focused intervention to shift climate change behaviors because first-order doxastic interventions have been ineffective. If people's credences are shifted because they aim to maintain coherence between their behavior and their beliefs, then a second-order, utility intervention could eventually lead people to form stable credences which are also appropriately calibrated to the evidence (Jachimowicz et al., 2018; Simon et al., 2015). Our framework highlights a promising direction of future research on this topic.

Using results from game theory to understand second-order belief reporting. People may report a second-order belief *because* it is perceived to be widely-held by in-group members (e.g., families, political or religious groups, nationalities) . Reporting beliefs that align with one's group requires coordinating with others (Kashima, Perfors, Ferdinand, & Pattenden, 2021). What mechanisms underlie coordination and how may their operation guide second-order belief revision?

Game theoretic models extend rational choice theories to study interactive decision-making, where optimal decision strategies hinge not only on one's own understanding of the world, but also on one's beliefs about other people's beliefs and desires (Colman, 2003). The computational mechanics of this process are the subject of extensive research and thus cannot be rehearsed here, but we can speak broadly about how games may relate to belief reporting in social networks. Games describe how rational agents would act given a space of states and associated utilities. Optimal decision-strategies, or Nash equilibria, optimize the expected value of an action given these constraints. When another agent's actions are uncertain, optimal decision-strategies are probabilistic, called mixed-strategy Nash equilibria. In mixed strategy Nash equilibria, more than one strategy could be optimal, but it ultimately depends on how other's will behave. One direction for future research could focus on how to manipulate people's understanding of how other behave to shape second-order beliefs. For example, consider the "stag hunt" game, one of the most widely studied cases in game theory and philosophy (Skyrms, 2004). In a stag hunt, two hunters must decide without communication to hunt either a stag or a hare. Both hunters can only kill the stag with the other hunter's help. Each hunter could kill a hare hunting alone. However, a hare yields less meat and less reward. The hunters must coordinate to capture the meaty stag, but coordinating is uncertain and depends on an infinite chain of expectations about how the other hunter is expected to act. Mathematical models of this game can describe how social structures (Skyrms, 2004) and norms (Bicchieri, 2005) emerge, as well as how people infer another person's motivations and belief-states during strategic interaction (van Baar, Nassar, Deng, & FeldmanHall, 2022). Rational decision-making in these games can have negative consequences

A key aspect of second-order interventions that change perceived utilities is to update probability matrices so agents decide to report more socially-positive beliefs. The misper-

ception of lack of climate change support among Americans demonstrates this idea. By manipulating perceptions of the ubiquity of a belief in a group, the optimal reporting strategy changes. A game-theoretic treatment of this process may not describe how first-order beliefs become second-order beliefs (the focus of this paper), but it may guide better how social sampling and utility calculations shape second-order belief reporting. Future work should explore ways to merge game theory with Bayesian cognitive modeling to understand the interplay between the top-down (coordination) and bottom-up (evidence integration) processes guiding second-order belief reporting.

Recasting previous psychological measures as measures of utility. We have argued that psychologists need to measure the perceived utility of accuracy and directionality to distinguish between motivated reasoning and practical reason. Although utility is a key construct in decision-making research, economics, and neighboring fields, it's less clear that psychologists have explicitly measured the utilities people associate with forming certain beliefs. However, one direction for future research could examine how psychological constructs which have featured prominently in research on motivated reasoning could, in fact, be measures of utility. For example, psychologists have measured participants' Need for Cognition in studies of motivated reasoning (Arceneaux & Vander Wielen, 2013; Caddick, 2016; Caddick & Feist, 2021; Nir, 2011), which we could recast within our framework as a measure of the utility people attribute to engaging in effortful cognitive tasks (Cacioppo & Petty, 1982; Petty, Briñol, Loersch, & McCaslin, 2009). It may be that higher scores on measures of Need for Cognition are akin to attributing higher utility to forming accurate beliefs. Similarly, psychologists have studied Need for closure and Personal need for Structure (Moskowitz, 1993; Neuberg, Judice, & West, 1997; Neuberg & Newsom, 1993; Webster & Kruglanski, 1994), which several researchers have reported as being related to motivated reasoning (Ask & Granhag, 2005; Dijksterhuis, Van Knippenberg, Kruglanski, & Schaper, 1996; Kruglanski, Pierro, Mannetti, & De Grada, 2006; Kundra & Sinclair, 1999; Sinatra, Kienhues, & Hofer, 2014). As is the case for Need for Cognition, these measures could be recast as measures of the utility of getting quick answers, or alternatively, the utility of being accurate. Thus, researchers interested in applying our framework may be able to use existing, validated measures to measure one half of the utility function which can impact second-order belief reports (viz., accuracy goals).

Conclusion

The present research provides a Bayesian decision-theoretic framework for understanding when and where motivated reasoning occurs and highlights the measurement and intervention implications of developing this framework. We first considered different norms of reason, drawing a distinction between theoretical and practical reason. Our aim was to sharpen how motivated reasoning is discussed and ground this discussion in an established literature on rational norms. We then developed a framework for quantifying motivated reasoning and demonstrated the Bayesian decision-theoretic framework with a toy experiment. The model simulations highlighted several key features of experimental designs for distinguishing motivated reasoning from practical rationality. The payoff of these developments is that the framework adds tools to psychologists' intervention toolkit – where first-order interventions may be comparatively ineffective, second-order interventions may be more promising. Altogether, this work highlights a different perspective on motivated

reasoning and relates this perspective to the measurement and development of interventions aimed at correcting misconceptions.

References

Alchian, A. A. (1953). The meaning of utility measurement. *The American Economic Review*, *43*(1), 26–50.

Alker, H., & Poppen, P. (1973). Personality and ideology in university students. *Journal of Personality*, *41*(4), 653–671.

Altay, S., de Araujo, E., & Mercier, H. (2022). "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news. *Digital Journalism*, *10*(3), 373–394.

Arceneaux, K., & Vander Wielen, R. J. (2013). The effects of need for cognition and need for affect on partisan evaluations. *Political Psychology*, *34*(1), 23–42.

Ask, K., & Granhag, P. A. (2005). Motivational sources of confirmation bias in criminal investigations: The need for cognitive closure. *Journal of investigative psychology and offender profiling*, *2*(1), 43–63.

Audi, R. (1985). Rationalization and rationality. *Synthese*, 159–184.

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*(3), 499–526.

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives*, *11*(1), 109–126.

Baekgaard, M., Christensen, J., Dahlmann, C., Mathiasen, A., & Petersen, N. (2019). The role of evidence in politics: Motivated reasoning and persuasion among politicians. *British Journal of Political Science*, *49*(3), 1117–1140.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109.

Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299.

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613.

Bayes, R., Druckman, J. N., Goods, A., & Molden, D. C. (2020). When and how different motives can drive motivated political reasoning. *Political Psychology*, *41*(5), 1031–1052.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bickart, B. (1993). Carryover and backfire effects in marketing research. *Journal of Marketing*, *30*(1), 52–62.

Birnbaum, M. H. (1992). Issues in utility measurement. *Organizational Behavior and Human Decision Processes*, *52*(3), 319–330.

Bleichrodt, H., Abellan-Perpiñan, J. M., Pinto-Prades, J. L., & Mendez-Martinez, I. (2007). Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. *Management Science*, *53*(3), 469–482.

Bodner, J., Welch, W., & Brodie, I. (2020). *Covid-19 conspiracy theories: Qanon, 5g, the new world order and other viral ideas*. McFarland.

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior*, *36*(2), 235–262.

Bostrom, N. (2009). Pascal's mugging. *Analysis*, *69*(3), 443–445.

Botzen, W. W., & van den Bergh, J. C. (2012). Risk attitudes to low-probability climate change risks: Wtp for flood insurance. *Journal of Economic Behavior & Organization*, *82*(1), 151–166.

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5), e2020043118.

Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.

Brenan, M., & Saad, L. (2018). *Global warming concern steady despite some partisan*

shifts. Retrieved from https://news.gallup.com/poll/231530/global-warming-concern-steady-despite-partisan-shifts.aspx.

Briggs, R. A. (2019). Normative Theories of Rational Choice: Expected Utility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/rationality-normative-utility/.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117.

Buehler, R., Griffin, D., & MacDonald, H. (1997). The role of motivated reasoning in optimistic time predictions. *Personality and Social Psychology Bulletin*, *23*(3), 238–247.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116-131.

Caddick, Z. A. (2016). Evaluating contradicting and confirming evidence: A study on beliefs and motivated reasoning. *ProQuest Dissertations and Theses*, 85. Retrieved from http://ezproxy.msu.edu/login?url=https://www.proquest.com/dissertations-theses/evaluating-contradicting-confirming-evidence/docview/1867574098/se-2

Caddick, Z. A., & Feist, G. J. (2021). When beliefs and evidence collide: psychological and ideological predictors of motivated reasoning about climate change. *Thinking & Reasoning*, 1–37.

Caddick, Z. A., & Rottman, B. M. (2021). Motivated reasoning in an explore-exploit task. *Cognitive Science*, *45*(8), e13018.

Chen, X. (2016). The influences of personality and motivation on the sharing of misinformation on social media. *IConference 2016 Proceedings*.

Chen, X., Sin, S.-C. J., Theng, Y.-L., & Lee, C. S. (2015). Why do social media users share misinformation? In *Proceedings of the 15th acm/ieee-cs joint conference on digital libraries* (pp. 111–114).

Christensen, J., & Moynihan, D. P. (2020). Motivated reasoning and policy information: Politicians are more resistant to debiasing interventions than the general public. *Behavioural Public Policy*, 1–22.

Cialdini, R. B., Braver, S. L., & Lewis, S. K. (1974). Attributional bias and the easily persuaded other. *Journal of Personality and Social Psychology*, *30*(5), 631–637.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in Experimental Social Psychology* (Vol. 24, pp. 201–234). Elsevier.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, *26*(2), 139–153.

Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602.

Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., ... Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action. *Psychological Science in the Public Interest*, *23*(2), 50-97.

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*.

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

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin*, *28*(10), 1379–1387.

DeBono, K. G., & Harnish, R. J. (1988). Source expertise, source attractiveness, and the processing of persuasive information: A functional approach. *Journal of Personality and Social*

Psychology, *55*(4), 541–546.
Dijksterhuis, A., Van Knippenberg, A., Kruglanski, A. W., & Schaper, C. (1996). Motivated social cognition: Need for closure effects on memory and judgment. *Journal of Experimental Social Psychology*, *32*(3), 254–270.
Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., ... Zinger, J. F. (2018). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, *14*(2), 273–291.
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.
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.
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.
Dixon, G., Bullock, O., & Adams, D. (2019). Unintended effects of emphasizing the role of climate change in recent natural disasters. *Environmental Communication*, *13*(2), 135–143.
Douglas, K. M. (2021). Covid-19 conspiracy theories. *Group Processes & Intergroup Relations*, *24*(2), 270–275.
Druckman, J. (2015). Communicating policy-relevant science. *PS: Political Science & Politics*, *48*(S1), 58–69.
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.
Easwaran, K. (2011). Bayesianism ii: Applications and criticisms. *Philosophy Compass*, *6*(5), 321–332.
Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*(4), 380–417.
Emler, N., Renwick, S., & Malone, B. (1983). The relationship between moral reasoning and political orientation. *Journal of Personality and Social Psychology*, *45*(5), 1073–1080.
Enders, A. M., Uscinski, J. E., Klofstad, C., & Stoler, J. (2020). The different forms of covid-19 misinformation and their consequences. *The Harvard Kennedy School Misinformation Review*.
Falk, A., & Zimmermann, F. (2016). *Beliefs and utility: Experimental evidence on preferences for information*. CESifo Working Paper Series.
Farquhar, P. H. (1984). State of the art—utility assessment methods. *Management Science*, *30*(11), 1283–1300.
Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2).
Fazio, L., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, *26*(5), 1705–1710.
Fischle, M. (2000). Mass response to the lewinsky scandal: Motivated reasoning or bayesianupdating? *Political Psychology*, *21*(1), 135–159.
Fishkin, J., Keniston, K., & McKinnon, C. (1973). Moral reasoning and political ideology. *Journal of Personality and Social Psychology*, *27*(1), 109–119.
Gallup. (2014). *Evolution, creationism, intelligent design.* Retrieved from https://news.gallup.com/poll/21814/evolution-creationism-intelligent-design.aspx
Gershman, S. (2018). How to never be wrong. *Psychonomic Bulletin and Review*, *26*(1), 1–16.
Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology*, *44*(6), 1110–1126.
Golman, R., & Loewenstein, G. (2018). Information gaps: A theory of preferences regarding the

presence and absence of information. *Decision*, *5*(3), 143–164.
Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J., Friese, M., ... others (2020). The implicit association test at age 20: What is known and what is not known about implicit bias.
Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.
Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, *72*(2), 167–189.
Guess, A., & Coppock, A. (2020). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, *50*(4), 1497–1515.
Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732.
Halter, A. N., & Mason, R. (1978). Utility measurement for those who need to know. *Western Journal of Agricultural Economics*, 99–109.
Harman, G. (1986). *Change in view: Principles of reasoning*. The MIT Press.
Harrison, L., & Startin, N. (2013). *Political research: An introduction*. Routledge.
Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, *39*(6), 701–723.
Hass, R. G. (1975). Persuasion or moderation? Two experiments on anticipatory belief change. *Journal of Personality and Social Psychology*, *31*(6), 1155–1162.
Hass, R. G., & Mann, R. W. (1976). Anticipatory belief change: Persuasion or impression management? *Journal of Personality and Social Psychology*, *34*(1), 105–111.
Hastorf, A. H., & Cantril, H. (1954). They saw a game; A case study. *Journal of Abnormal and Social Psychology*, *49*(1), 129–134.
Heesacker, M., Petty, R. E., & Cacioppo, J. T. (1983). Field dependence and attitude change: Source credibility can alter persuasion by affecting message-relevant thinking. *Journal of Personality*, *51*(4), 653–666.
Hershey, J. C., & Schoemaker, P. J. (1985). Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science*, *31*(10), 1213–1231.
Hickling, A., Wellman, H., & Dannemiller, J. L. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology*, *37*(5), 668–683.
Holton, R. (1999). Intention and weakness of will. *The Journal of Philosophy*, *96*(5), 241–262.
Horne, Z., & Livengood, J. (2017). Ordering effects, updating effects, and the specter of global skepticism. *Synthese*, *194*(4), 1189–1218.
Horne, Z., Powell, D., & Hummel, J. (2015). A single counterexample leads to moral belief revision. *Cognitive Science*, *39*(8), 1950–1964.
Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112*(33), 10321–10324.
Horne, Z., Rottman, J., & Lawrence, C. (2021). Can coherence-based interventions change dogged moral beliefs about meat-eating? *Journal of Experimental Social Psychology*, *96*, 104160.
Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). Relationships among conspiratorial beliefs, conservatism and climate scepticism across nations. *Nature Climate Change*, *8*(7), 614–620.
Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*(5958), 1410–1412.
Hunt, J. M., Domzal, T. J., & Kernan, J. B. (1982). Causal attributions and persuasion: The case of disconfirmed expectancies. *ACR North American Advances*.
Imundo, M. N., & Rapp, D. N. (2022). When fairness is flawed: Effects of false balance reporting and weight-of-evidence statements on beliefs and perceptions of climate change. *Journal of Applied Research in Memory and Cognition*, *11*(2), 258.
Jachimowicz, J. M., Hauser, O. P., O'Brien, J. D., Sherman, E., & Galinsky, A. D. (2018). The

critical role of second-order normative beliefs in predicting energy conservation. *Nature Human Behaviour*, *2*(10), 757–764.

Jackson, E. G. (2019). Belief and credence: Why the attitude-type matters. *Philosophical Studies*, *176*(9), 2477–2496.

Jain, A., Marshall, J., Buikema, A., Bancroft, T., Kelly, J. P., & Newschaffer, C. J. (2015). Autism occurrence by mmr vaccine status among us children with older siblings with and without autism. *Jama*, *313*(15), 1534–1540.

James, O., & Van Ryzin, G. G. (2017). Motivated reasoning about public performance: An experimental study of how citizens judge the affordable care act. *Journal of Public Administration Research and Theory*, *27*(1), 197–209.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756–763.

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

Jin, Y., Jensen, G., Gottlieb, J., & Ferrera, V. (2022). Superstitious learning of abstract order from random reinforcement. *Proceedings of the National Academy of Sciences*, *119*(35), e2202789119. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2202789119 doi: 10.1073/pnas.2202789119

Jolley, D., Douglas, K. M., & Sutton, R. M. (2018). Blaming a few bad apples to save a threatened barrel: The system-justifying function of conspiracy theories. *Political Psychology*, *39*(2), 465–478.

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, 1–17.

Kahan, D. (2013). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision Making*, *8*, 407–424.

Kahan, D., & Braman, D. (2006). Cultural cognition and public policy. *Yale Law and Policy Review*, *24*, 149-172.

Kahan, D., Landrum, A., Carpenter, K., Helft, L., & Hall-Jamieson, K. (2017). Science curiosity and political information processing. *Political Psychology*, *38*, 179–199.

Kahne, J., & Bowyer, B. (2017). Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. *American Educational Research Journal*, *54*(1), 3–34.

Kalis, A., Mojzisch, A., Schweizer, T. S., & Kaiser, S. (2008). Weakness of will, akrasia, and the neuropsychiatry of decision making: An interdisciplinary perspective. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 402–417.

Kashima, Y., Perfors, A., Ferdinand, V., & Pattenden, E. (2021). Ideology, communication and polarization. *Philosophical Transactions of the Royal Society B*, *376*(1822), 20200133.

Killen, M., & Stangor, C. (2001). Children's social reasoning about inclusion and exclusion in gender and race peer group contexts. *Child Development*, *72*(1), 174–186.

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.

Kraus, M. W., & Tan, J. J. (2015). Americans overestimate social class mobility. *Journal of Experimental Social Psychology*, *58*, 101–111.

Kruglanski, A. W., Pierro, A., Mannetti, L., & De Grada, E. (2006). Groups as epistemic providers: need for closure and the unfolding of group-centrism. *Psychological review*, *113*(1), 84-100.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Kundra, Z., & Sinclair, L. (1999). Motivated reasoning with stereotypes: Activation, application, and inhibition. *Psychological Inquiry*, *10*(1), 12–22.

Laato, S., Islam, A., Islam, M. N., & Whelan, E. (2020). Why do people share misinformation

during the covid-19 pandemic? *arXiv preprint arXiv:2004.09600*.

Levy, A. G., Thorpe, A., Scherer, L. D., Scherer, A. M., Drews, F. A., Butler, J. M., ... Fagerlin, A. (2022). Misrepresentation and nonadherence regarding covid-19 public health measures. *JAMA Network Open*, *5*(10), e2235837–e2235837.

Lewandowsky, S., Cook, J., Fay, N., & Gignac, G. E. (2019). Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Memory & Cognition*, *47*(8), 1445–1456.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). Nasa faked the moon landing—therefore,(climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, *24*(5), 622–633.

Lewandowsky, S., & van der Linden, S. (2022). Interventions based on social norms could benefit from considering adversarial information environments: Comment on constantino et al. (2022). *Psychological Science in the Public Interest*, *23*(2), 43-49.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Little, A. T. (2021). Directional motives and different priors are observationally equivalent. *University of California-Berkeley (Unpublished Manuscript)*, 1–31.

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, *2*(3), 166–167.

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.

Luce, R. D. (1991). Rank-and sign-dependent linear utility models for binary gambles. *Journal of Economic Theory*, *53*(1), 75–100.

Madrid-Morales, D., Wasserman, H., Gondwe, G., Ndlovu, K., Sikanku, E., Tully, M., ... Uzuegbunam, C. (2020). Motivations for sharing misinformation: a comparative study in six sub-saharan african countries. *International Journal of Communication*.

Madson, G. J., & Hillygus, D. S. (2020). All the best polls agree with me: Bias in evaluations of political polling. *Political Behavior*, *42*(4), 1055–1072.

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

Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31), e2200300119.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422–430.

McGuire, W. J. (1964). Inducing resistance to persuasion. some contemporary approaches. *CC Haaland and WO Kaelber (Eds.), Self and Society. An Anthology of Readings, Lexington, Mass.(Ginn Custom Publishing), 1981*, 192-230.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*(6), 423–429.

Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Meyer, A., & Hundtofte, C. S. (2022). The longshot bias is a context effect. *Available at SSRN*.

Midgon, B. (2022). Experts worry monkeypox disinformation will harm LGBTQ+ community. *The Hill*.

Miske, O., Schweitzer, N., & Horne, Z. (2019). What information shapes and shifts people's attitudes about capital punishment? *PsyArXiv*. https://doi.org/10.31234/osf.io/z6cxd.

Molleman, L., Gradassi, A., Sultan, M., & van den Bos, W. (2021). Partisan biases in social

information use.

Moskowitz, G. B. (1993). Individual differences in social categorization: The influence of personal need for structure on spontaneous trait inferences. *Journal of Personality and Social Psychology*, *65*(1), 132-142.

Neuberg, S. L., Judice, T. N., & West, S. G. (1997). What the need for closure scale measures and what it does not: Toward differentiating among related epistemic motives. *Journal of personality and social psychology*, *72*(6), 1396.

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of personality and social psychology*, *65*(1), 113-131.

Nir, L. (2011). Motivated reasoning and public opinion perception. *Public Opinion Quarterly*, *75*(3), 504–532.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: a randomized trial. *Pediatrics*, *133*(4), e835–e842.

Orne, M. T. (2017). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. In *Sociological methods* (pp. 279–299). Routledge.

Painter, D. L., & Fernandes, J. (2022). "The big lie": How fact checking influences support for insurrection. *American Behavioral Scientist*, 00027642221103179.

Pasek, J. (2018). It's not my consensus: Motivated reasoning and the sources of scientific illiteracy. *Public Understanding of Science*, *27*(7), 787–806.

Pennycook, G. (2022). A framework for understanding reasoning errors: From fake news to climate change and beyond. *PsyArXiv*.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of Experimental Child Psychology*, *39*(3), 437–471.

Persson, E., Andersson, D., Koppel, L., Västfjäll, D., & Tinghög, G. (2021). Covid-19 and vaccine hesitancy: A longitudinal study. *Cognition*, *214*(104768).

Peter, C., & Koch, T. (2016). When debunking scientific myths fails (and when it does not): The backfire effect in the context of journalistic coverage and immediate judgments as prevention strategy. *Science Communication*, *38*(1), 3–25.

Pettigrew, R. (2019). Epistemic Utility Arguments for Probabilism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/epistemic-utility/.

Petty, R. E., Briñol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior.* The Guilford Press.

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*,

37(10), 1915–1926.

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, *46*(1), 69–81.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, *41*(5), 847–855.

Pierre, J. M. (2020). Mistrust and misinformation: A two-component, socio-epistemic model of belief in conspiracy theories. *Journal of Social and Political Psychology*, *8*(2), 617–641.

Pilditch, T. D., Roozenbeek, J., Madsen, J. K., & van der Linden, S. (2022). Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. *Royal Society Open Science*, *9*(8), 211953.

Portmore, D. W. (2013). Perform your best option. *The Journal of Philosophy*, *110*(8), 436–459.

Porumbescu, G. A., Moynihan, D., Anastasopoulos, J., & Olsen, A. L. (2020). Motivated reasoning and blame: Responses to performance framing and outgroup triggers during covid-19. *arXiv preprint arXiv:2009.03037*.

Powell, D., Weisman, K., & Markman, E. (2022). Modeling and leveraging intuitive theories to improve vaccine attitudes.

Powell, D., Weisman, K., & Markman, E. M. (2018). Articulating lay theories through graphical models: A study of beliefs surrounding vaccination decisions. *Annual Proceedings of the Cognitive Science Society*, 906–911.

Priniski, J. H., & Horne, Z. (2019). Crowdsourcing effective educational interventions. In *41st Annual Meeting of the Cognitive Science Society* (pp. 2599–2605).

Radcliffe, E. S. (1999). Hume on the generation of motives: Why beliefs alone never motivate. *Hume Studies*, *25*(1), 101–122.

Redlawsk, D. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, *64*(4), 1021–1044.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*(1), 42–56.

Robison, L. J. (1982). An appraisal of expected utility hypothesis tests constructed from responses to hypothetical questions and experimental choices. *American Journal of Agricultural Economics*, *64*(2), 367–375.

Roozenbeek, J., Freeman, A. L., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al.(2020). *Psychological Science*, *32*(7), 1169–1178.

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), eabo6254. Retrieved from https://www.science.org/doi/abs/10.1126/sciadv.abo6254 doi: 10.1126/sciadv.abo6254

Sanitioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social Psychology*, *59*(2), 229–241.

Schoenfield, M. (2018). Permissivism and the value of rationality: A challenge to the uniqueness thesis. *Philosophy and Phenomenological Research*, *99*(2), 286–297.

Schult, C. A., & Wellman, H. M. (1997). Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition*, *62*(3), 291–324.

Sharma, A. (1990). The persuasive effect of salesperson credibility: conceptual and empirical examination. *Journal of Personal Selling & Sales Management*, *10*(4), 71–80.

Shweder, R. A., Mahapatra, M., & Miller, J. G. (1987). Culture and moral development. *The Emergence of Morality in Young Children*, 199–283.

Shweder, R. A., & Sullivan, M. A. (1993). Cultural psychology: Who needs it? *Annual Review of Psychology*, *44*(1), 497–523.

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and*

oppression. New York: Cambridge University Press.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.
Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1250–1260.
Simon, D., Stenstrom, D. M., & Read, S. J. (2015). The coherence effect: Blending cold and hot cognitions. *Journal of Personality and Social Psychology*, *109*(3), 369–394.
Sinatra, G. M., Kienhues, D., & Hofer, B. K. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist*, *49*(2), 123–138.
Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, *4*(4), 267–281.
Skyrms, B. (2004). *The stag hunt and the evolution of social structure.* Cambridge University Press.
Slepian, M. L. (2022). A process model of having and keeping secrets. *Psychological Review*, *129*(3), 542–563.
Soicher, R. N., & Becker-Blease, K. A. (2020). Utility value interventions: Why and how instructors should use them in college psychology courses. *Scholarship of Teaching and Learning in Psychology*.
Sparkman, G., Geiger, N., & Weber, E. U. (2022). Americans experience a false social reality by underestimating popular climate policy support by nearly half. *Nature Communications*, *13*, 4779.
Stanley, J. (2015). How propaganda works. In *How propaganda works.* Princeton University Press.
Stanley, M., Henne, P., Yang, B., & De Brigard, F. (2020). Resistance to position change, motivated reasoning, and polarization. *Political Behavior*, *42*(3), 891–913.
Stetzka, R. M., & Winter, S. (2021). How rational is gambling? *Journal of Economic Surveys*.
Sunstein, C. (2000). Essay: Deliberative trouble? Why groups polarize. *Yale Law Journal*, *110*, 71–83.
Taber, C., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, *31*(2), 137–155.
Taber, C., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.
Tangari, A., Bui, M., Haws, K., & Liu, P. J. (2019). That's not so bad, i'll eat more! backfire effects of calories-per-serving information on snack consumption. *Journal of Marketing*, *83*(1), 133–150.
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.
Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness.* Penguin.
Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140.
Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The Annals of the American Academy of Political and Social Science*, *700*(1), 136–151.
Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.
Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.
Uscinski, J. E., Douglas, K., & Lewandowsky, S. (2017). Climate change conspiracy theories. In *Oxford research encyclopedia of climate science.*

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, *6*(3), 404–414.

Van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). The gateway belief model: A large-scale replication. *Journal of Environmental Psychology*, *62*, 49–58.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008.

Vineberg, S. (2022). Dutch Book Arguments. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/dutch-book/.

Vivion, M., Anassour Laouan Sidi, E., Betsch, C., Dionne, M., Dubé, E., Driedger, S. M., ... others (2022). Prebunking messaging to inoculate against covid-19 vaccine misinformation: An effective strategy for public health. *Journal of Communication in Healthcare*, 1–11.

Von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior (commemorative edition)*. Princeton University Press.

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.

Wallace, R. J. (2020). Practical reason. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Stanford University. https://plato.stanford.edu/archives/spr2020/entries/practical-reason/.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of personality and social psychology*, *67*(6), 1049-1062.

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, *118*(2), 357–378.

Wickens, T. D. (2001). *Elementary signal detection theory*. Oxford University Press.

Wiggins, D. (1978). Weakness of will commensurability, and the objects of deliberation and desire. In *Proceedings of the aristotelian society* (Vol. 79, pp. 251–277).

Williams, D. (2021). Socially adaptive belief. *Mind & Language*, *36*(3), 333–354.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163.

Wu, C., & Shaffer, D. R. (1987). Susceptibility to persuasive appeals as a function of source credibility and prior experience with the attitude object. *Journal of Personality and Social Psychology*, *52*(4), 677–688.

Yu, A. J., & Cohen, J. D. (2008). Sequential effects: Superstition or rational behavior? *Advances in neural information processing systems*, *21*.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*(2), 268–282.

Zimper, A., & Ludwig, A. (2009). On attitude polarization under Bayesian learning with non-additive beliefs. *Journal of Risk and Uncertainty*, *39*(2), 181–212.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality*, *47*(2), 245–287.