Personality and Social Psychology Review
1–25

DOI: 10.1177/10888683241251520
pspr.sagepub.com
Sage

The Case for Heterogeneity in Metacognitive Appraisals of Biased Beliefs

Corey Cusimano[1]

Abstract

Academic Abstract

Prominent theories of belief and metacognition make different predictions about how people evaluate their biased beliefs. These predictions reflect different assumptions about (a) people's conscious belief regulation goals and (b) the mechanisms and constraints underlying belief change. I argue that people exhibit heterogeneity in how they evaluate their biased beliefs. Sometimes people are blind to their biases, sometimes people acknowledge and condone them, and sometimes people resent them. The observation that people adopt a variety of "metacognitive positions" toward their beliefs provides insight into people's belief regulation goals as well as insight into way that belief formation is free and constrained. The way that people relate to their beliefs illuminates why they hold those beliefs. Identifying how someone thinks about their belief is useful for changing their mind.

Public Abstract

The same belief can be alternatively thought of as rational, careful, unfortunate, or an act of faith. These beliefs about one's beliefs are called "metacognitive positions." I review evidence that people hold at least four different metacognitive positions. For each position, I discuss what kinds of cognitive processes generated belief and what role people's values and preferences played in belief formation. We can learn a lot about someone's belief based on how they relate to that belief. Learning how someone relates to their belief is useful for identifying the best ways to try to change their mind.

Keywords

belief, metacognition, naive realism, motivated reasoning, lay ethics of belief, bias blind spot, illusion of objectivity

> "I suppose they try and make you believe an awful lot of nonsense?"
>
> *Charles*
>
> "Is it nonsense? I wish it were. It sometimes sounds terribly sensible to me."
>
> *Sebastian*
>
> "But my dear Sebastian, you can't seriously believe it all. . . about Christmas and the star, and the ox and the ass, and the three kings."
>
> *Charles*
>
> "Yes, I believe all that—it's lovely!"
>
> *Sebastian*
>
> *Brideshead Revisited* (1945, p. 44)

Between Sebastian and Charles in this dialogue, Sebastian better represents the common believer. Like Sebastian, many people believe in magical and spiritual forces (Atran & Norenzayan, 2004; Pennycook et al., 2012; White & Norenzayan, 2019). People also hold unrealistic beliefs about romance, relationships, and their health (Baker & Emery, 1993; Murray & Holmes, 1997; Sprecher & Metts, 1989; Taylor & Brown, 1994; Trémolière & Djeriouat, 2019). And in general, people's values, preferences, and identity bias the beliefs that they form (Porot & Mandelbaum, 2021). Sebastian also resembles many people when he introspects and evaluates the quality of his belief (in this case, as sensible and lovely). This kind of

[1]Yale University, New Haven, CT, USA

Corresponding Author:
Corey Cusimano, School of Management, Yale University, 165 Whitney Avenue, New Haven, CT 06511, USA.
Email: corey.cusimano@yale.edu

metacognition—monitoring and evaluating one's thinking—is an integral part of belief formation (Jost et al., 1998; Nelson & Narens, 1994). People analyze competing thoughts that come to mind, check their thoughts for errors, and engage, monitor, and evaluate steps to change their mind (Bhatia, 2017; De Neys & Pennycook, 2019; Wegener et al., 2012; Wilson & Brekke, 1994).

These two observations raise the central question of this article: *What do people think about their biased beliefs?* This question does not have an easy answer. It is commonplace wisdom that people always think of themselves as "rational" in the sense that their current beliefs always reflect a dispassionate weighing of evidence. But it is also commonplace wisdom that religious individuals believe "on faith" in the sense that they believe in God without thinking that they have evidence to do so. But people cannot think that the same belief is a product of faith *and* evidence; and likewise, people cannot treat their belief like a dispassionate observation *and* as a passionate commitment. This tension in everyday talk about belief mirrors controversy in psychological theory. Some theories reject the notion that people ever believe something on faith because people must always view their beliefs as rational or supported by evidence (e.g., Kruglanski, 1996; Kunda, 1990; Pronin et al., 2004). Other work argues that people can readily admit that their beliefs lack evidence (e.g., Cusimano & Lombrozo, 2023). And still others argue that, even if people do not always view their beliefs as rational, they should at least view them as valuable (e.g., Abelson, 1986). In this article, I will describe different accounts of how people evaluate their biased beliefs. Each account describes a kind of "metacognitive position," in other words, a collection of beliefs about one's belief.

We might naturally be drawn to a compromise view such that people sometimes evaluate their beliefs in different ways—in other words, that people at different times or for different beliefs adopt different metacognitive positions. And indeed, this is the view at which I will arrive. I will catalog empirical support for a variety of positions, and argue that, in principle, each position could describe how any person evaluates any of their beliefs. However, this conclusion requires grappling with conflicting ideas about what kinds of beliefs people desire and conflicting ideas about whether, and how, reasoning is constrained. Finally, I argue that it is useful to identify how someone thinks about their belief: Metacognitive positions signal what psychological processes proximately trigger and sustain belief, and for this reason, different metacognitive positions recommend different strategies for changing belief.

Defining Bias, Belief, and Metacognitive Position

The central claim of the current article is that people can adopt a variety of different metacognitive positions toward their biased beliefs. Here is what I mean by these terms:

Belief

A belief is a feeling that some idea (such as a proposition, model, or image of the world) is true. To believe an idea is for that idea to be impressive or forceful in one's mind such that it affects one's deliberation, emotions, and choices as if it were true (Hume, 1793/2017). This definition could be restated in colloquial terms such that people "believe" ideas when those ideas "feel real" to them. This investigation into belief is restricted to a particular set of beliefs, namely beliefs about *matters of fact*. Beliefs about matters of fact can be evaluated as accurate or inaccurate based on whether they correspond to reality. For instance, beliefs such as "God exists," "My friend committed a crime," and "I will win the lottery" count as matters of fact because they may be true (or not) depending on whether God really does exist, whether my friend really did commit a crime, and whether I really will win the lottery.

Biased Belief

Narrowing this investigation further, I am concerned with the subset of beliefs that are *biased*. In this essay, a belief is "biased" when it is not fully explainable as a result of impartial, evidence-based reasoning (Cusimano & Lombrozo, 2021b). So, one way that a belief can be biased is when it is the result of *partial* reasoning. People reason in a partial way when they accept or reject a proposition based in part on the influence of their values or preferences. This happens, for instance, when people think it is risky or undesirable to adopt a belief and so demand especially strong evidence for that belief before accepting it (e.g., Ditto & Lopez, 1992; Gilovich, 1991; Trope & Liberman, 1996). People's reluctance to accept bad news and eagerness to accept good news are good candidates for biased belief in this sense.

Beliefs are also biased when the person who holds it lacks sufficient evidence for it. A useful way to think about whether someone lacks evidence for their belief is to imagine what would happen if this person called to mind everything they have experienced and learned and then reformed their belief in a way that reflected all this information (and all the rules of inference that they also endorse, e.g., consistency, parsimony, etc.; Foley, 1987). If they would change their mind, then their belief was not supported by their evidence. I will sometimes label these biased beliefs as "irrational" or "epistemically irrational." Many superstitious and magical beliefs are good candidates for being biased or irrational in this sense. Many people who believe in magic and superstition often also endorse the arguments and principles that recommend a more scientific outlook. If these individuals reformed their beliefs while incorporating everything they knew, then they would stop believing in magic. Likewise, people who hold overly rosy beliefs about themselves or their relationships sometimes do so because they have neglected evidence that recommends less rosy beliefs. These beliefs do not reflect all of their evidence, would change if they did, and so these individuals are properly thought of as biased.

Evidence

Evidence comprises experiences, arguments, and other kinds of information that are diagnostic of the truth or falsity of a proposition. Evidence in this sense need not be data obtained from an experiment or something published in *Science*. With respect to the existence of God, for instance, relevant evidence may include testimony from parents and pastors affirming God's existence, apparently mystical experiences, exposure to the ontological argument for God's existence, and so on. Whether someone's belief about God is unbiased depends on whether consideration of all their evidence provides sufficient support for their belief (again, assuming inferential rules that the reasoner also endorses, including consistency, appeal to the best explanation, etc.). It is possible, on this definition, for someone to hold an unbiased belief in God and for someone to hold a biased belief in God. This observation applies to most of the beliefs that I will discuss throughout this article, including religious, magical, and superstitious beliefs.

Metacognitive Beliefs and Metacognitive Positions

Metacognitive beliefs are beliefs about one's beliefs. There are potentially many different beliefs that a person can make about their belief—I will focus on only a few. The first two metacognitive beliefs correspond to the two kinds of bias introduced above. These are meta-beliefs about (a) whether one's belief is sufficiently supported by evidence, and (b) whether one has weighed or evaluated one's evidence in an impartial way. If people believe that they have reasoned impartially, and that their belief is supported by sufficiently strong evidence, then they believe that they formed the belief "objectively." Someone who believes that they have been objective in this sense believes that other people who have the same evidence that they do, and who also think about that evidence impartially, would hold the same belief (Ross & Ward, 1996).

The final metacognitive belief this paper will discuss is (c) an all-things-considered evaluation of the value of the belief. I will use the term "justified" to refer to this kind of overall assessment of the belief. The term "justified" sometimes has a technical meaning, but here it means the same thing as whether the belief satisfies someone's standards or goals for belief. In other words, it means essentially the same thing as "valuable" or "good." This is the sort of judgment that, when it is positive, leads someone to want to keep their belief, and when it is negative, leads someone to want to change it. This judgment is an all-things-considered judgment because it potentially incorporates both considerations of the belief's epistemic qualities (such as whether the belief enjoys sufficient evidence or whether it was formed impartially) and the belief's nonepistemic qualities (such as whether the belief is useful, affirming, respectful, safe, etc.).

Finally, a "metacognitive position" is a collection of beliefs about one's belief. The range and variety of metacognitive positions that this paper will explore will be restricted to the set that can be combined by considering the individual metacognitive beliefs of (a) evidence, (b) impartiality, and (c) justification. Figure 1 shows how different metacognitive beliefs can combine to yield the four different metacognitive positions that I will discuss.

This essay investigates a narrow set of metacognitive beliefs and metacognitive positions. There are many other ways of describing a belief as "biased," and likely many other metacognitive positions, that predict how people think, feel, and behave. My goal is not to provide a comprehensive review of every meaning of bias, every mechanism that gives rise to a biased belief, or every metacognitive judgment. Instead, my goal is to demonstrate that, even among this narrow framing of what it means for a belief to be "biased," there is metacognitive variation that is important to acknowledge and useful to measure.

Here is how the paper will proceed. First, I will review evidence that people believe that their biased beliefs are free of bias. This metacognitive position, commonly called the "illusion of objectivity," appears in accounts of "naïve realism" (Ross & Ward, 1996) and the "bias blind spot" (Pronin et al., 2004) as well as influential models of motivated reasoning and belief formation (e.g., Kunda, 1990; Pyszczynski & Greenberg, 1987). I will then discuss prominent views that argue that this metacognitive position is (effectively) universal. The rest of the paper then challenges these views by challenging common assumptions about people's belief regulation goals and capacities. Along the way, I will derive metacognitive positions that increasingly depart from an "illusion of objectivity" (Figure 1).

The Illusion of Objectivity

A common view of metacognition states that people tend to form and maintain biased beliefs while maintaining a conviction that their beliefs "follow from a relatively dispassionate, unbiased, and essentially 'unmediated' apprehension of the evidence or information at hand" (Ross & Ward, p. 110). In other words, people maintain an "illusion of objectivity" about their biased beliefs—they wrongly believe that their biased beliefs impartially reflect their evidence (Kunda, 1990; Pronin et al., 2004; Pyszczynski & Greenberg, 1987). Many studies demonstrate that people adopt this position for many biased beliefs. For instance, people regularly claim that their beliefs are based on evidence, logic, and common sense, while denying the influence of ubiquitous and well-known biases (Pronin et al., 2004). People commonly report that others, if they had the same information and thought about things rationally, would believe the same thing that they do (Marks & Miller, 1987; Robinson et al., 1995; Rogers et al., 2017). And, when people encounter others

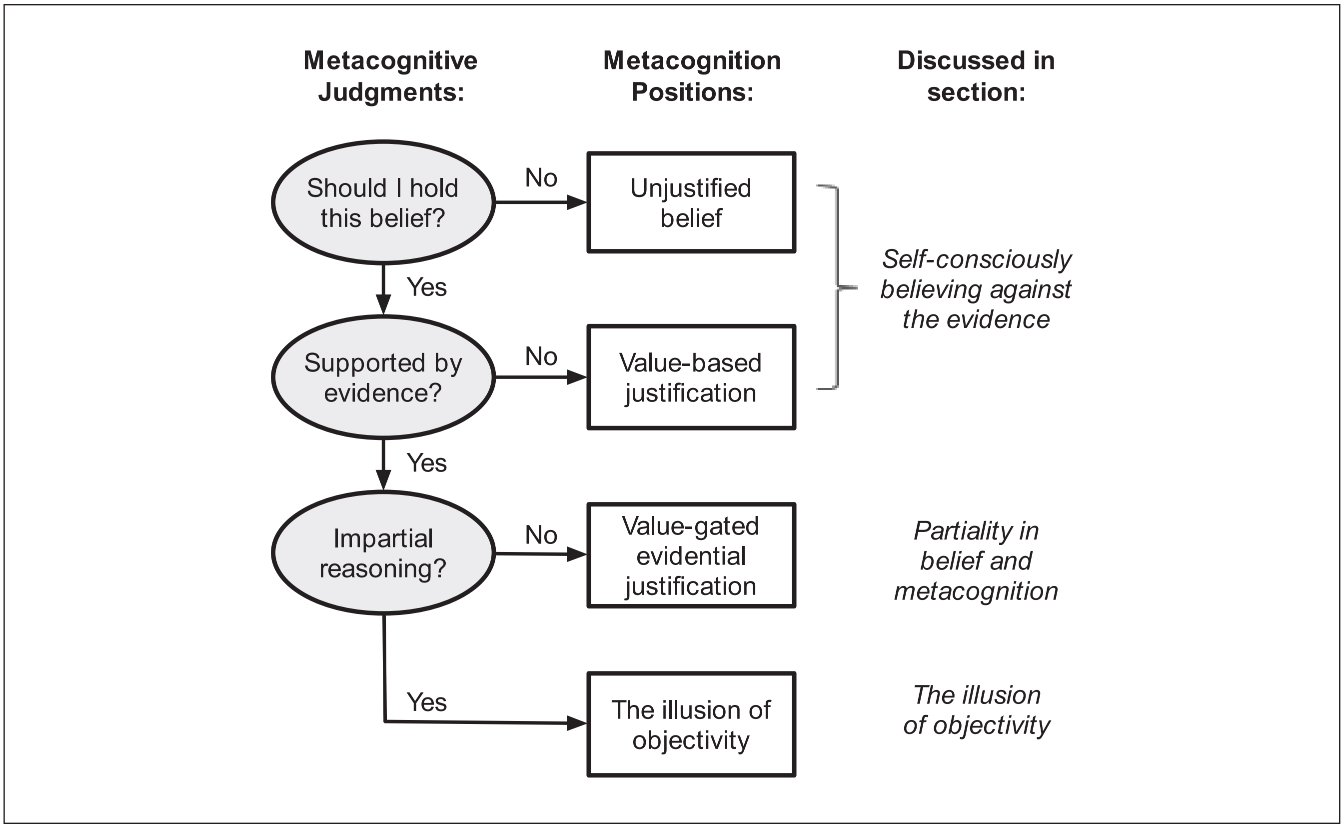

Figure 1. I derive four "metacognitive positions" (white boxes) toward a belief based on whether the belief (i) seems justified, (ii) is based on "sufficient evidence," and (iii) was formed impartially.

who disagree with them, they commonly explain away these disagreements by assuming that those others either lack the same information or must be biased (Kennedy & Pronin, 2008; Reeder et al., 2005).

There is no doubt that, on occasion, people are biased and fail to realize it. But how common is this position, and what guidance do we have for predicting when people are likely to be blind to their current biases? The findings above recommend the claim that people have at least a general tendency to be naïve about their biases. However, a common view in social and cognitive psychology makes a stronger claim, namely, that people *essentially always* possess an illusion of objectivity about their biased beliefs.

The illusion of objectivity is plausible as a near-universal description of metacognition because it is plausible that belief formation is constrained in a way that entails this position. Consider one challenge that all theories of belief must overcome: They all must account for the way in which belief formation is constrained. After all, although people can be biased, it is not as if people just believe *anything*. The most influential response to this challenge comes from Kunda (1990) who argues that people only form beliefs that pass a "reality constraint":

> people [who are] motivated to arrive at a particular conclusion attempt to be rational and to construct a justification of their desired conclusion that would persuade a dispassionate observer. They draw the desired conclusion *only if* they can muster up the evidence necessary to support it. (pp. 482–483, emphasis added).

This constraint is more aptly described as a metacognitive constraint. People are not constrained to believe what is true per se, only what they believe is unbiased to believe. And at least in the way it is articulated here,[1] this metacognitive constraint is a by-product of people's conscious belief regulation goals.

Like Kunda (1990) in the passage above, many theories derive a constraint on belief from the standards that people consciously hold themselves to while reasoning. Stated in other ways, for instance, people only adopt beliefs that they view as "legitimate" in the sense that they have evidence for thinking that they are true (Kruglanski, 1996, p. 503). Or, people make sure during reasoning that they do not form beliefs that threaten their self-concept of being a rational person (Pyszczynski & Greenberg, 1987, p. 302). Or, people think that brazen attempts to ignore the truth "make a mockery" of belief (Baumeister & Newman, 1994, p. 5), or that it is "perverse" to knowingly hold a biased belief (Pronin et al., 2004, p. 791). This characterization of reasoning accords with how people talk about the kind of believers they want to be. In questionnaires that measure people's beliefs about what constitutes good thinking, people strongly endorse open-mindedness, impartiality, logic, rationality, and evidence-based reasoning as important virtues (e.g., Baron, 2019; Pennycook et al., 2020; Stahl et al., 2016). And accordingly, people consciously avoid biases and correct for biases when they detect them. In other words, any time someone is about to believe that their belief is biased, their desire to be unbiased kicks in and they modify the belief to keep it unbiased. The only biases left over are the ones that escape this scrutiny—the unconscious ones. So, even if people in principle recognize that they are susceptible to biases, they "do not recognize that [they] are succumbing to them in any particular assessment [they] are currently making" (pp. 783–784, Pronin et al., 2004). In other words, at least when asked in general terms, people identify with Charles from the epigraph. And because people want to be impartial, evidence-based thinkers, that is what they think they are.

To explain why people might often be biased despite biases seeming undesirable, many scholars propose that biases operate "only tacitly and unconsciously" (p. 503, Kruglanski, 1996; see also Baumeister & Newman, 1994; Kunda, 1990). Indeed, it is common in psychology to assume that biases are subservient to metacognition such that they operate "in ways that enable one to maintain an illusion of objectivity" (p. 302, Pyszczynski & Greenberg, 1987). This claim is supported not only by the observation that people want to be rational and unbiased thinkers but also by early studies investigating when people seem to be biased and when they seem to recognize bias. In early studies of motivated reasoning, people did not form desirable beliefs that seemed to require too obvious a departure from impartial, evidence-based reasoning (Kunda, 1990). Other studies showed that, even when biases seemed like they ought to be obvious, people nevertheless routinely denied them. For instance, participants in many studies denied being biased even when they demonstrably were and even when they said that people just like them would be biased in the same situation (e.g., Ehrlinger et al., 2005; Frantz, 2006; Hansen et al., 2014; Pronin et al., 2002; West et al., 2012). These studies further reinforced a view of belief and metacognition wherein biases only operate unconsciously and people only think of themselves as unbiased.

In sum, one position that people can adopt toward their biased belief is to think that it is unbiased. An influential explanation for this error is that it is a by-product of how belief formation is constrained, and in particular, the way that belief formation is constrained by people's conscious goals to reason unbiasedly. If belief formation is constrained by this metacognitive position, then this position characterizes practically every biased belief that people hold. This characterization of belief and metacognition is common: It is found in theories of cognitive dissonance (Festinger, 1957), influential models of motivated reasoning (Baumeister & Newman, 1994; Epley & Gilovich, 2016; Kruglanski, 1996; Kunda, 1990; Pyszczynski & Greenberg, 1987; Sherman et al., 2009), the psychological immune system (Gilbert et al., 1998), and accounts of the bias blind spot (Pronin et al., 2004) (see Rosenzweig, 2016, for a review).

Theoretical and Practical Upshots of the Illusion of Objectivity as a Universal Description of Metacognition

If belief is constrained by metacognitive appraisals of objectivity, then explanations for how people adopt biased beliefs must accommodate how they do so while maintaining an illusion of objectivity. That is, a theoretical commitment to the claim that metacognition constrains belief itself constrains psychological explanations of belief. The case of religious belief provides a useful example. There is reason to think that religious beliefs are common in part because of the benefits that they provide to their adherents (Laurin & Kay, 2017). Accordingly, the ultimate reason why people believe in God is because this belief makes them feel close to their peers, makes them feel in control of their environment, and gives them a sense of purpose. However, appeals to the usefulness or desirability of belief cannot be consistent with an illusion of objectivity without also explaining how people tacitly trick themselves into thinking that their belief is justified by evidence. In this case, proximate explanations for religious belief may appeal to experiences that provide putative evidence for God, disembodied minds, and other supernatural forces (Atran & Norenzayan, 2004; White & Norenzayan, 2019) or appeal to testimony from committed and apparently knowledgeable people in the environment (Henrich, 2009). In other words, if people never self-consciously believe in God "on faith," then psychological explanations for religion must explain how people think that they believe in God as a result of impartial, evidence-based reasoning.

This constraint on psychological theorizing about belief generalizes to all functional explanations for belief. To see how, consider the claim by Tetlock (2002) that beliefs can be explained by appealing to metaphors of people variably thinking like *scientists, politicians, prosecutors*, and *theologians.* When someone is thinking like a scientist, they reason in an open-minded way, questioning their beliefs and scrutinizing their evidence. But when someone thinks like a prosecutor, they think in a one-sided way by collecting evidence with the goal of buttressing a particular conclusion and otherwise going easy on their preferred views. The mindset that someone adopts can flexibly change from moment to moment. For instance, someone may become more prosecutorial when their job suddenly relies on convincing their boss that they add value to the company or when an outsider questions a belief central to their identity. But according to the claims just discussed, this moment-by-moment heterogeneity in cognition coincides with moment-by-moment homogeneity in metacognition: Even when people suddenly start thinking like a prosecutor, they still believe they are thinking like a scientist.

There are two practical upshots of this view about how metacognition constrains belief. First, introspection should be a poor guide for understanding the influence of desires, values, and needs on someone's belief (Nisbett & Wilson, 1977; Pronin, 2009). It must be if the ultimate reasons why we believe something—reasons pertaining to the usefulness or desirability of belief—are always laundered into perceived evidence. Why trust someone when they say that they are thinking like a scientist when they would say that no matter how they were thinking? The second practical upshot of this view constitutes advice for the best way to debias people, change belief, and resolve disagreement. If people essentially always hold biased beliefs under an illusion of objectivity, then debiasing them essentially always requires educating them about their biases or otherwise explaining to them why some target belief is better recommended by impartial or evidence-based thinking.

Challenging the Illusion of Objectivity as a Constraint on Belief

The rest of this essay will challenge the common idea that people near-universally hold a metacognitive position akin to an illusion of objectivity. In doing so, this essay will also challenge the common theoretical and practical upshots of this view. However, it is worth remembering why the objectivity illusion is appealing as an (effectively) universal description of metacognition. First, it provides a ready explanation for why people are biased although they generally say they do not want to be (answer: they do not realize it). Second, this view seems to do a good job describing how biased and motivated reasoning is constrained (answer: people only believe what they can metacognitively believe they have "objective" reasons to believe). To challenge the claim that beliefs are constrained in this way, it is insufficient to just identify a handful of apparent counterexamples. Challenging this view requires explaining how people accommodate thinking that their belief is biased.

The rest of the article is organized around two observations. First, one of the main reasons to think that people always hold biased beliefs under an illusion of objectivity is that people believe that biased beliefs are unjustified. Pronin et al. (2004) appeal to this assumption when they write, "If one were aware that a given influence was compromising the accuracy of one's present judgments. . . one presumably would modify the judgment in question" because "to do otherwise would be perverse" (pp. 790–791). But, as reviewed below, people sometimes think that biased reasoning is legitimate (Armor et al., 2008; Cusimano & Lombrozo, 2021a; Tenney et al., 2015). If people do not always think that biases are perverse, then they may not always try to correct them. The second observation is that people sometimes hold obviously irrational beliefs even though they do not want to. This observation provides the grounds for a new way of thinking about how belief is constrained. And this observation provides the means to understanding how people can hold a range of metacognitive positions that involve self-attributing a lack of evidence.

Partiality in Belief and Metacognition

Although people endorse impartial, evidence-based reasoning as valuable in the abstract, they do not always do so when evaluating specific beliefs embedded in specific situations (Cusimano & Lombrozo, 2011a; Stahl & Cusimano, 2023). Indeed, when people evaluate their own and others' beliefs on specific issues, they tend to do so by considering both the belief's epistemic qualities (such as whether the belief is unbiased) and the belief's nonepistemic qualities (such as whether it is practically or morally valuable, e.g., Cusimano & Lombrozo, 2021a; Tenney et al., 2015). Here I review one way that people think their values can justifiably encroach into their beliefs, namely, by moderating the strength of evidence required before accepting or rejecting an idea. This metacognitive position is called "value-gated evidential justification."

Partiality in Belief Formation

People may be aware of how their values have affected their beliefs by being aware of how these values have affected how they have weighed evidence en route to belief. To appreciate how this may happen, consider the way that evidence gathering typically unfolds and leads to belief: When considering whether to believe some hypothesis, people seek out evidence that speaks for or against it, interrogate that evidence by generating alternative explanations, and repeat this process until they settle on a belief, or stop without feeling settled on the matter (Pyszczynski & Greenberg, 1987; Trope & Liberman, 1996). For instance, someone considering the impact of coffee consumption on their health might consider the possibility that coffee is bad for them, attend to evidence speaking for and against this idea, consider alternative interpretations of that evidence, and then repeat and refine this procedure.[2] This process can last a while: There is little limit to someone's ability to search for evidence and to generate alternative explanations. But in practice, people stop

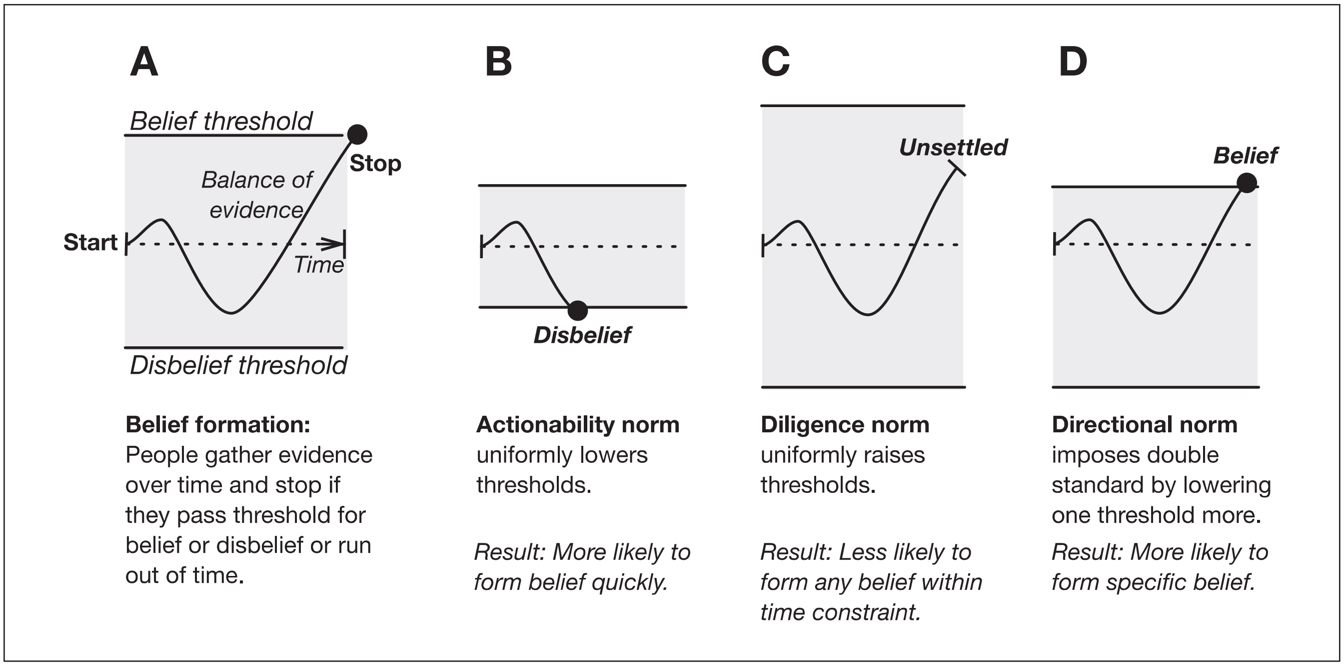

Figure 2. This figure demonstrates three different ways one's values can influence assessments of sufficient evidence. (A) In general, people consider a proposition and accumulate evidence for and against it until they hit a threshold of "sufficient evidence" or run out of time. (B) Actionability norms uniformly lower evidential threshold for adopting belief, increasingly likelihood of quickly forming belief. (C) Diligence norms uniformly raise evidential threshold for belief, increasing likelihood of feeling unsettled when one runs out of time. (D) Some norms, like giving someone the benefit of the doubt or avoiding risk of error, operate by creating evidential double standards that make it easier to satisfy the evidential requirements for a particular conclusion over its opposite.

thinking and make an inference. They do so when they judge either that they have expended the appropriate amount of time on the question (and are left feeling undecided), or because they have gathered what they consider to be sufficiently strong evidence for a particular conclusion (and so believe or disbelieve some hypothesis). This process of evidence gathering relative to stopping thresholds of sufficient evidence and time is described in many models of belief formation (Ditto & Lopez, 1992; Kruglanski, 2004; Kunda, 1990; Pyszczynski & Greenberg, 1987; Trope & Liberman, 1996) and is represented in Figure 2A.

Flexible Threshold Setting. One way that people's motives and values affect their beliefs is by changing the threshold that determines whether they think that they have sufficient evidence for belief (Kruglanski, 1990, 2004; Pyszczynski & Greenberg, 1987; Trope & Liberman, 1996). The most common motives that affect belief formation in this way derive from people's need to optimally distribute the limited time and energy that they have to think about things. For instance, when it is costly to not form a belief, such as when one needs to act right away, it makes sense to lower one's standard of evidence (Figure 2B). And indeed, when people need to act quickly, they form beliefs on less evidence (e.g., Kruglanski & Freund, 1983). By contrast, when a false belief could be especially costly (as it may be when one is deciding whether they remembered to lock every door before leaving for vacation), it makes sense to demand stronger evidence before coming to a decision (Figure 2C). And indeed, people spend more time obtaining and evaluating evidence and take longer to form beliefs when a false belief could be especially costly (e.g., Mayseless & Kruglanski, 1987, Study 1; McAllister et al., 1979). We can codify the opposing pressures of time and error avoidance on reasoning in value-laden terms that are easy to recognize—the former as the norm of *actionability* and the latter as the norm of *due diligence.*[3]

Trope and Liberman (1996) propose a model of evidence gathering and threshold setting in which the threshold for belief and disbelief may be asymmetric (Figure 2D; see also Arkes, 1991; Friedrich, 1993; Lord & Taylor, 2009). Accordingly, people independently set thresholds for accepting or rejecting a belief based on the risks of false acceptance and false rejection. This practice may result in asymmetric thresholds that favor adopting a particular belief over its opposite. For instance, the costs of falsely believing that one's friend is guilty of a minor crime might be much greater than the costs of falsely believing that the friend is innocent. After all, thinking poorly of a friend may end the relationship while thinking overly positively about the friend will not. Thus, when the evidence is imperfect and points both ways, someone bearing these risks in mind will be more likely to believe that their friend is innocent, and much less likely to believe that their friend is guilty. In believing this way, this person would be reasoning in line with a norm to *minimize asymmetric risks of error*.

Evidence Thresholds and Motivated Reasoning. Shifting what counts as "sufficient evidence" is one mechanism of biased reasoning, and therefore, a common cause of biased belief (Ditto & Lopez, 1992; Gilovich, 1991; Lord & Taylor, 2009; Trope & Liberman, 1996). For instance, a dedicated coffee drinker will consider it especially risky to jump to the conclusion that coffee is unhealthy. After all, if they are wrong, then they would have wrongly given up a cherished habit. To mitigate this risk, they will demand especially strong evidence before concluding that coffee is unhealthy. Likewise, this person may be less worried about potentially wrongly believing that coffee is healthy, and, so, will require less evidence to believe it. As a result, as they sift through evidence that coffee is healthy and unhealthy, the dedicated coffee drinker is more likely to form the belief that coffee is healthy than they are to believe that it is unhealthy or remain in a position where they feel like they have no view on the matter (Figure 2D). Generalizing from this example, people apply low evidential thresholds for desired beliefs, and high evidential thresholds for undesired beliefs, because it tends to feel less risky to form beliefs that enable one's preferences than it is to form beliefs that rule them out (Trope & Liberman, 1996). Gilovich (1991) captures this kind of bias in the helpful form of a slogan: When people want to believe something, they ask themselves whether they "can" believe it based on the evidence, but when they do not want to believe something, they ask whether they "must" believe it. This kind of double-standard threshold-setting is one way that people biasedly adopt beliefs while maintaining a commitment to believing things based on evidence.

Partiality in Metacognition

Consistent with the theories discussed above, according to which the illusion of objectivity is a near-universal description of metacognition, it is common to assume that considerations of actionability, due diligence, and risk minimization affect reasoners only unconsciously. After all, even if considerations of actionability and due diligence entail a commitment to evidence, reasoning this way violates a norm of impartial, value-free thinking (Bolinger, 2020; Cusimano & Lombrozo, 2021b). A coffee drinker who refuses to believe that coffee is unhealthy is being partial when their disbelief reflects their judgment that it is too risky *for them* to believe that coffee is unhealthy. Someone who disbelieves that their friend is guilty of a crime because it feels risky *as their friend* to believe it is likewise failing to be impartial. And a scientist who withholds belief in a controversial theory after obtaining evidence for that theory, on account of the moral controversy that would ensue, is not doing science in a value-free way (Rudner, 1953). If people judge violations of impartiality as unjustified, then they would correct for these violations whenever they detected them. The only instances leftover in which these biases would affect people would be the instances in which people fail to detect them. And these evidential double-standards would just be one other way that people's motives bias their beliefs unconsciously.

However, people need not think that it is unjustified to moderate their belief formation in line with norms of actionability, due diligence, and risk minimization.[4] If people think that it is justified for these considerations to impact their beliefs then, in principle, people's self-assessments of their evidence may vary, holding their belief fixed, without them believing that they are any more or less justified in their belief. For instance, people may acknowledge that they have accepted a belief on weaker-than-usual evidence but believe that the belief is justified in light of pressure to make a quick decision. Likewise, people might recognize that they have obtained stronger-than-normal evidence but feel justified in withholding belief because they think that they must be especially careful about what to believe in this situation. In other words: People may be aware that they are applying a double standard to their evidence without subsequently feeling like they need to correct their belief. But this way of thinking constitutes a metacognitive position alien to the illusion of objectivity. It entails that people sometimes judge that another person, who lacks the momentary demands of actionability or due diligence, but who otherwise has the same information, may hold a belief that is different than their own.

People do sometimes endorse norms of belief formation that permit shifting one's evidence threshold. For instance, Cusimano and Lombrozo (2021a) report a study in which participants read about a young newlywed who learns from a reputable source that 70% of couples in his demographic get divorced within 5 years of marriage. In light of this (and other) information about the newlywed's relationship, participants reported that, from the newlywed's point of view, the total evidence he has suggests there is a 59% chance that he divorces in the next 5 years. Participants then read either that the newlywed believes he has a 0%, or 70%, chance of divorce. Participants reported a stronger sentiment that he "should have collected more evidence before believing" in reaction to the 70% chance belief compared to the 0% chance belief. Other studies reported in Cusimano and Lombrozo (2021a) show people prescribing evidential double-standards in other common situations too, suggesting a general principle whereby people hold lower standards of evidence for morally preferable beliefs compared to morally bad beliefs.

If people are not motivated to correct violations of impartiality, then they might be aware of this bias in their reasoning. After all, when an error feels especially important to avoid, or pressure to make a quick judgment feels especially strong, these feelings are both more likely to affect belief and to be salient during introspection. One way to test whether people may be aware of this kind of bias is to put people in a situation where they are likely to withhold belief based on concerns related to risk or due diligence and then ask them whether their concerns about accepting certain conclusions affected their judgment. If people always think that their beliefs follow from an unmediated apprehension of their evidence, then people should deny that such concerns explain their belief. This finding would be consistent with an illusion of objectivity and consistent with the view that a conscious commitment to objectivity characterizes and constrains belief. But if people acknowledge this form of partiality, then it would follow that belief is not constrained by a conscious commitment to objectivity.

Cusimano and Lombrozo (2023) conducted several studies testing whether people detect this form of bias in their beliefs. In one study (Study 4), participants read about a published study that tracked the outcomes of gender dysphoric teenagers who took puberty suppressants. Gender-affirming care in teens is controversial in part because people differ with respect to which error they think is worse: Some people are really worried about a *false positive*—believing that gender affirming care is helpful when really it isn't. Others are more worried about the (corresponding) *false negative*—believing that gender affirming care is not helpful when really it is. Based on the model of belief reviewed above, the former group should be less accepting of new evidence that gender affirming care is helpful, while the latter group should be less accepting of new evidence that gender affirming care is unhelpful. This pattern is exactly what Cusimano and Lombrozo (2023) found. When participants read about a medical study reporting that gender dysphoric teens showed improved psychological adjustment after puberty suppressants, those who were concerned about falsely concluding that gender affirming care is helpful accepted the study's conclusion less frequently compared to others. Results flipped when participants read that, post intervention, feelings of gender dysphoria did not improve. Now, participants who were concerned about falsely concluding that gender affirming care does not help were less accepting. This preferential acceptance and rejection of new scientific evidence was genuinely biased: The strength of evidence for these two findings was the same, and participants' concerns about the relative risk of error predicted their beliefs even after accounting for differences in their prior beliefs about gender affirming care.

The key test in this study was whether participants were aware that their value judgments affected their acceptance of the study. Consistent with people being aware of the influence of their values on their beliefs, participants explained their acceptance (and nonacceptance) by citing their concerns about how risky it felt to them to accept the evidence. For instance, participants who disbelieved that teens were better adjusted after two years of puberty suppressants reported that one reason they disbelieved was "a concern about how believing the wrong thing could hurt teens who might get gender affirming care but regret it." Participants also by and large thought that their resulting belief was justified. When asked whether they ought to have incorporated this kind of concern in their belief, this same group almost universally reported that they weighed this concern "the right amount" in their reasoning about the study. In other words, participants not only knew that they were holding a risky belief to stricter standards, they thought they were reasoning just as they ought to. Insofar as participants judged their reasoning to be ideal, it was because they thought that they properly weighed motivational concerns in their evaluation of new evidence, not because they thought that they ignored these concerns altogether.

Metacognitive Awareness of Partiality in Belief—Summary

This section has argued for one plausible alternative to the illusion of objectivity. In this alternative metacognitive position, people retain a commitment to forming beliefs based on evidence and retain the sense that their beliefs enjoy sufficient evidence; however, they self-consciously deny a commitment to impartiality. Violations of impartiality derive from norms of reasoning that moderate the strength of evidence required prior to settling on a belief. These norms include norms of actionability and due diligence. In mundane situations, people might talk about these considerations by saying that they are "giving their friend the benefit of the doubt" or that they are avoiding "rushing into judgment" about a high-stakes issue. People endorse norms that moderate the strength of evidence required prior to belief (Cusimano & Lombrozo, 2021a) and demonstrate awareness (and acceptance) of biased reasoning in line with these norms (Cusimano & Lombrozo, 2023). I have labeled this position as one in which people believe they possess a "value-gated evidential justification" for their belief.

Self-Consciously Believing Against the Evidence

The illusion of objectivity and value-gated evidential justification share the metacognitive judgment that one's belief enjoys sufficient evidence. These positions differ only with respect to whether people believe that they have weighed and evaluated that evidence impartially. However, people are often biased such that their beliefs are unsupported by their total evidence. Is it possible to hold a metacognitive position that evaluates one's belief this way?

Addressing this question requires acknowledging another shared feature of the illusion of objectivity and value-gated evidential justification, namely, the cognitive model of belief change that these positions both assume. Both positions derive from a model of belief formation wherein belief is the output of gathering and appraising evidence. In this model, beliefs form when someone appraises some observation or argument as evidence for some idea. However, this model of belief formation cannot account for beliefs that people appraise as *un*supported by evidence. These latter kinds of beliefs—if they exist—must exist by virtue of cognitive processes that supersede or operate independently from consciously appraising evidence. In other words, whether people can adopt metacognitive positions that self-attribute a lack of evidence depends on whether appraising evidence is a necessary proximate cause for belief or whether beliefs can come and go as a result of altogether different processes.

The most straightforward alternative model of belief formation that would enable metacognitive judgments of insufficient evidence comes from *economic models of belief*. In these models, belief change is the output of judging that a belief is valuable (Abelson, 1986; Loewenstein & Molnar, 2018; Sharot et al., 2023; Van Bavel & Peirrera, 2018). Accordingly, people still attend to how evidentially supported a belief is, but this appraisal is just one of many that they consider when deciding what to believe. They also think about how good it would feel, how well it fits with their identity, and other features that speak to its overall, all-things-considered value. Whatever belief people then adopt is the output of a multiattribute choice that calculates the relative value of different beliefs based on their properties and weighted by the believer's preferences. If beliefs changed this way, then people could easily, and with some regularity, adopt a metacognitive position that involved appraising their belief as evidentially poor. For instance, if someone identified a proposition as one that is evidentially poor, but nevertheless the one that they most wanted to believe, then they would thereby believe it, and would hold that belief alongside their (meta)belief that it is evidentially poor.

However, it is very unlikely that beliefs work like preference-based multiattribute choices. Economic models of belief are misguided because they fail to account for the way in which belief formation seems to be constrained by forces that ignore people's preferences. And in particular, these models do not account for the common observation that people often fail to believe what they most prefer because beliefs seem to change spontaneously and automatically in response to evidence (Elster, 1979; Festinger, 1957; James, 1937). This observation about the way that belief formation appears to be constrained is very similar to the one reviewed in the previous section on the illusion of objectivity. As noted there, people do not simply believe whatever is most desirable, but instead tend to believe in ways that hew closely to what they can rationalize (Kunda, 1990). In many studies, for instance, motivated reasoning and self-deception all-but-disappeared when the desired conclusion was patently false (e.g., Bar-Hillel & Budescu, 1995; Sloman et al., 2010). Indeed, some of the evidence that supported the idea that people *want* to think that their beliefs are backed by evidence seems better explained by the idea that people *have* to think that their beliefs are backed by evidence. Accordingly, biases lose their potency when people become aware of them because reappraising one's evidence automatically triggers correction whether the believer wanted their belief corrected or not (see, e.g., discussions in studies by Balcetis & Dunning, 2006; Baumeister & Newman, 1994; Gilbert et al., 1998; Ross, 2018; Sherman et al., 2009). And finally, when people think about the evidence and arguments that support their beliefs, their beliefs feel unchangeable and outside of their control (Cusimano & Goodwin, 2020).[5] None of these individual sources of evidence is definitive—for instance, people might have poor insight into their capacity to voluntarily change what they believe—but taken together they strongly support the conclusion that beliefs are involuntarily tethered to conscious appraisals of evidence.

In many circumstances, what people believe seems constrained by what they think they have evidence to believe. This observation rules out many theories that propose mechanisms of belief change that supplant or supersede belief formation as a process of gathering and appraising evidence. For instance, it rules out theories according to which belief change is a product of appraising a belief's overall value. This line of reasoning has provided another argument for the claim that people near-universally self-attribute evidence for their beliefs (e.g., Baumeister & Newman, 1994; Epley & Gilovich, 2016; Kunda, 1990; Ross, 2018).

How People Sometimes Believe Beyond the Perceived Evidence

It is common to describe different beliefs as reflecting a difference between "intuition" and "reflection" (or "intuition" and "reason"; e.g., Denes-Raj & Epstein, 1994; Kahneman, 2011; Risen, 2016; Walco & Risen, 2017). But a conflict between intuition and reflection is sometimes better characterized as a conflict between *belief* (what "feels real") and a corresponding *metacognitive belief*, formed during reflection, about "what would be evidentially rational to believe." In these circumstances, intuitions can be thought of as beliefs that form via processes that do not give rise to, or seem to depend on, metacognitive beliefs about one's evidence.[6] Below I review evidence that beliefs sometimes float free of metacognitive judgments of evidence in this way. This happens when beliefs act like *habits* and when beliefs act like *reflexes*.

Habitual Belief Formation. Some beliefs are like habits: They are reinforced through repetition, triggered by environmental cues, and as a result, are resistant to change. People can act out of habit despite knowing that they are not acting rationally. So too can they believe out of habit despite knowing they are not believing rationally.

Research in psychopathology documents many such habituated beliefs. Consider, for instance, cases of clinical depression. Depression often reflects some obstinate irrational and maladaptive belief that must be trained away (Beck, 1979, 2008; Ellis, 1962). One challenge when treating clinical depression is that patients sometimes have an illusion of objectivity about their belief. In these circumstances, the therapist must find ways to challenge the patient's rationale for their belief (Baron et al., 1990). But succeeding at this task is rarely sufficient to extinguish the maladaptive belief and cure the patient's depression. Depressed people maintain their maladaptive beliefs even after they acknowledge that they do not make sense (e.g., Beck et al., 1979; Ellis, 1962).[7] And so, another challenge in therapeutic contexts is to help someone train their beliefs to fall in line with their (meta) attitudes about how they ought and want to think.

Depressive beliefs are stubborn because they often have been reinforced through repeated recall and elaboration (Brewin, 2006). As a result of this habituation, depressive beliefs are prone to intrusively popping back into one's mind, reflecting a feature of memory wherein the thoughts that most readily come to mind are the thoughts that one has most often had (Baddeley, 1990). As a result, treatment for depression often requires increasing the competitiveness of different thoughts through adversarial repetition, perseverance, and elaboration. This procedure often involves teaching a patient how to induce more realistic and constructive beliefs when the negative ones initially come to mind. Over time, these more constructive beliefs compete with (and eventually inhibit) the original maladaptive beliefs (see Barber & DeRubeis, 1989; Brewin, 2006; see Lane et al., 2015, for a slightly different mnemonic theory of therapeutic change).

Some beliefs outside psychopathology have a similar habituated quality. Many people form unscientific beliefs early in life and appear to keep them despite years of subsequent education (Shtulman & Harrington, 2016; Shtulman & Lombrozo, 2016; Shtulman & Valcarcel, 2012). Indeed, the habituated quality of certain unscientific beliefs helps explain people's apparent resistance to scientific education. In one study, Shtulman and Harrington (2016) recruited college students, adults from the Los Angeles community, and science faculty, to make speeded judgments about claims that were either scientifically supported ("being sneezed on can make a person sick") or unsupported ("being cold can make a person sick"). They also manipulated whether the statements were intuitive ("being cold can make a person sick") or held no intuitive appeal ("being happy can make a person sick"). All three groups correctly identified the scientifically supported claims and rejected the intuitive but wrong claims. However, every group experienced some difficulty rejecting the intuitive but wrong claims. Even science professors who had a lifetime of experience with the correct information were slower to reject intuitive (but wrong) claims compared with unintuitive (and wrong) claims. Education sometimes does not eliminate and replace memories that become beliefs, but instead piles on new memories that compete to become beliefs (Shtulman & Lombrozo, 2016).

Reflexive Belief Formation. Some beliefs are like reflexes—spontaneous acts triggered by environmental cues and ignorant to one's preferences and rational capacities. People can sometimes react reflexively despite awareness that it is not a rational way to act. So too can they believe reflexively despite awareness that it is not a rational way to believe.

Associative processes are frequent causes of these kinds of reflexive beliefs (Risen, 2016; Sloman, 1996). Ideas will pop into one's mind, and feel real, merely because they share some superficial connection to some other idea. These associations operate spontaneously and autonomously from other cognitive processes that analyze ideas along different dimensions (Risen, 2016; Sloman, 1996). For instance, Sloman (1996) reports that, although he knows that logic dictates the probability of Linda "being a bank teller" is higher than Linda "being a bank teller and active in the feminist movement," it nevertheless seems to him that the latter idea, which is representative of Linda, *feels* right.

Many superstitious and magical beliefs can be explained by reflexive belief-generating processes (Risen, 2016). For instance, people will refuse to eat chocolate shaped into feces, despite being fully aware that such thoughts are irrational, because they strongly associate the shape of feces with the properties of feces (Rozin et al., 1986). Attention also appears to have a reflexive effect on people's belief. Attending to certain outcomes spontaneously increases the perceived likelihood of those outcomes occurring (Block & Kramer, 2009; Risen & Gilovich, 2008). This feature of belief formation causes people to overestimate the likelihood of the salient outcome even in the face of strong reasons to rule the outcome out. For instance, many people who stand in the observation tower at the Grand Canyon report an uncontrollable fear that the glass will break because that outcome is especially salient when they stand on the glass and look down (Gendler, 2008). Risen and Gilovich (2008) show how this tendency leads people to think they are magically "tempting fate." For instance, students commonly believe that skipping the reading for class makes it more likely that the professor will call on them. This superstitious belief forms reflects the thought of being called out popping into mind at the moment when the student considers skipping the reading. The salience of the unwanted outcome, makes the outcome seem more likely.

Finally, people reflexively form beliefs when their environment educes innate, ready-made concepts (Carey, 2009; Carey & Spelke, 1996; Gelman & Legare, 2011). People are endowed with domain-specific concepts about physical objects, forces, people, and animals (among others). These concepts explain how people make quick, intuitive judgments that one object pushed another, that something is alive, or that something has a mind and acts intentionally. Very young children, for instance, balk at physically impossible events, affiliate with objects that minimally look and act like friendly creatures, and promiscuously attribute minds and purposes to the objects around them (Kelemen, 2004; Kelemen & DiYanni, 2005; Rochat et al., 1997). These intuitions are not a product of experience but of evolution. And for this reason, new experiences do not extinguish them (Carey, 2009; Kelemen & Rosset, 2009). Even to adults, fire seems alive because it dances while spores seem lifeless because they do not (Shtulman & Legare, 2020). Even to adults, simple cardboard shapes in stop-motion movies give the impression that they have minds of their own (Heider & Simmel, 1944; see also Barrett & Lanman, 2008, for discussion). And as with beliefs that people habituate, these innate intuitions interfere with people's ability to internalize new scientific ideas. Lessons learned from life experience and education do not supplant these beliefs but instead occupy the mind alongside them (Legare & Gelman, 2008; Legare et al., 2012; Shtulman & Lombrozo, 2016).

Summary and Application to Constraints of Belief

Habitual and reflexive beliefs demonstrate that beliefs are not the output of a single process, but instead, are the output of multiple processes. One important process involves gathering and appraising evidence. But beliefs also arise from mnemonic and associative processes as well as from innate concepts that comprise core knowledge. These competing causes of belief enable people to self-consciously believe beyond the evidence. We recognize this division of thought when we colloquially comment that we "know" what to believe but are struggling to "internalize" it.

The observation that beliefs have multiple causes provides a new way of thinking about constraints on belief formation. Contrary to some of the theories discussed above (e.g., Kunda, 1990), belief is not constrained to conscious appraisals of evidence. After all, beliefs can come about in the absence of, and in contradiction to, thinking about evidence. However, and contrary to economic theories of belief (e.g., Abelson, 1986), just because people can self-consciously believe beyond the evidence does not mean that people can believe whatever they want. Instead, in general terms, the best way to characterize how belief is constrained is to say that belief requires the presence of effective evidential or nonevidential cues to trigger it. Accordingly, whether someone is free to adopt a particular belief in a particular situation depends on myriad factors relating to the myriad mechanisms they might be able to leverage: *What does the evidence say?; Are the right associative cues readily available?; Has the belief been habituated?; and How good are they at suppressing thoughts or controlling their attention?* Appraising a belief as valuable or desirable does not appear to be an effective trigger. And because judging a belief as valuable is not effective, but thinking about evidence often is, people will feel forced to believe the evidence against their wishes. But people feel forced to believe things for other reasons too. People also feel forced to believe things habitually and reflexively despite knowing there are rational reasons not to and despite wishing they could believe otherwise.

Metacognitive Position: Unjustified Belief

Many of the "habitual" and "reflexive" beliefs reviewed above appear to lack any valuable, justifying qualities to their believers. Not only are these beliefs easily identified as irrational or inconsistent with a scientific worldview, but they are also devoid of any moral, emotional, or practical benefit. There is no emotional benefit to thinking that chocolate has the same properties as feces or that one is magically tempting fate by avoiding homework. Indeed, examples from psychopathology demonstrate that these recalcitrant beliefs are often acknowledged by their believers to be both irrational *and dysfunctional*. Clinical anxiety, obsessive-compulsive disorder (OCD), and phobias are often characterized by their "egodystonic" nature in that people readily acknowledge their own irrationality (Beck, 1979; see also Ellis, 1962; Kozak & Foa, 1994; Robbins et al., 2019).[8] For instance, most people diagnosed with OCD are at least somewhat aware that their fears are unreasonable (Kozak & Foa, 1993). Acknowledging that these beliefs are biased or irrational is not sufficient to extinguish them because these beliefs are not constrained by metacognitive judgments of objectivity. But they feel real, otherwise people would not acquiesce to them, and they would not seek out therapy, medication, and other tools to redress them.

Judging a belief as unjustified represents another counterexample to the apparent universality of the illusion of objectivity. Indeed, one popular description of people's default metacognitive position is that they "see things the way that they really are" (Pronin et al., 2004; Ross, 2018). This visual analogy can be turned on its head: People know that some of the things they see are visual illusions. When we see a stick bend in water, we recognize that what we see does not make sense even while we cannot get ourselves to see straight. Some beliefs are no different (Sloman, 1996). In such cases, people are unable to change their beliefs just by thinking about what would be rational (or even what would be most desirable) to believe. In these instances, belief change may either be impossible (like many visual illusions) or may require incremental change over a long time.

Metacognitive Position: Value-Based Justification for Belief

If people can believe something that they think they lack evidence to believe and prefer *not* to believe it, then people can believe something that they think they lack evidence to believe but *prefer* to believe it. This latter metacognitive position is one in which someone has a "value-based justification" for their belief. People can adopt this metacognitive position through the confluence of (a) nonevidential triggers of belief and (b) a lay ethics of belief that treats beliefs as justified even in the absence of sufficient evidence.

People sometimes judge beliefs to be justified because they confer potent practical, emotional, or moral benefits. For instance, people commonly believe that overoptimism helps people stay motivated toward a goal. So, when someone needs to stay motivated, people will prescribe overoptimism to that person (Tenney et al., 2015). In other circumstances, people think others ought to be overly pessimistic instead (Miller et al., 2021). For instance, people sometimes think others ought to be unrealistically pessimistic because they think pessimism leads people to take greater care to prevent bad outcomes (Miller et al., 2023). Cusimano and Lombrozo (2021a) found that people will prescribe beliefs to others that violate the evidence if those beliefs confer moral value. For instance, participants reported that others ought to hold overly optimistic beliefs about marriage, and someone's chance of recovering from cancer, because doing so is morally laudatory. In these cases, overoptimism either signaled loyalty or helped bring about good outcomes. People may not think it is always perverse to hold evidentially unsupported beliefs—instead, in some cases, it might seem like the best thing to do.

Consistent with the analysis presented so far, people sometimes hold beliefs that they think are valuable despite thinking that they lack evidence for them. Siepmann and colleagues (2004) asked participants to estimate the likelihood of different life events like getting married, breaking a bone, or getting treated for alcoholism. They also asked participants to estimate how desirable each of these outcomes would be and to what extent remaining overly optimistic about each of these outcomes would help them achieve their goals. Participants' judgments of how desirable it was to be optimistic correlated with their beliefs about how likely each outcome was. And, when asked about what factors influenced their beliefs, participants readily, and correctly, acknowledged that the desirability of being optimistic had done so. Cusimano and Lombrozo (2023) found that people sometimes judge their beliefs as evidentially poor but nevertheless justified because of how morally good they are (Figure 3). They asked participants to evaluate a variety of their beliefs along several dimensions, including how much evidence they had and how morally beneficial the belief was. Unsurprisingly, self-attributed evidence strongly predicted both belief in the proposition and judgments that the belief was justified. When participants thought that their evidence favored a proposition, they reported believing the proposition 85% of the time; and when their evidence did not favor the proposition, they only reported belief in the proposition 17% of the time. However, judgments of moral quality played a role, too: Participants were much more likely to hold a belief, despite their evidence failing to favor the belief, when the belief was morally valuable (Figure 3). This metacognitive position was most common for propositions like "God exists," and "animals experience suffering the same way that humans do."

The kinds of beliefs that people feel justified to hold without evidence seem to overlap with the kinds of beliefs that people can reflexively and habitually trigger. Religious beliefs, again, provide illustrative examples. People derive value from their religious beliefs because these beliefs make them feel in control of their lives and because these beliefs enable a sense of connection with their religious peers. And indeed, in Cusimano and Lombrozo (2023), religious beliefs were one of the most common in which people reported holding and valuing a belief despite thinking they could not defend the belief with evidence. Religious beliefs are also plausibly buttressed by nonevidential triggers. People reflexively experience many natural phenomena as mindful and goal oriented. Viewing the natural world this way is part and parcel of many supernatural beliefs (Barrett, 2000; Barrett & Lanman, 2008; Boyer, 2003). However, the evidence that these experiences provide for the supernatural is extremely limited—these experiences are incompatible with a scientific worldview and can be explained away. It makes more sense to think of people's intuitive beliefs of the natural world as contributing triggers to religious belief (Barrett & Lanman, 2008). Many religious beliefs also have a habitual quality to them. For instance, religious beliefs are often instilled at a young age. Through daily or weekly rituals, they are repeatedly recalled, rehearsed, and elaborated (see, e.g., examples in Luhrmann, 2020). Rituals may help religious beliefs become habituated, in turn sustaining them later in life when people confront arguments against God's existence and when their reasons for belief are explained away. And indeed, people who were raised religious still partially endorse religious beliefs even after, as adults, they have abandoned religion and their religious identity (Van Tongeren et al., 2021). Religious belief is just one example, but it illustrates how someone, with the help of nonevidential triggers of belief, can consciously sustain a valued belief "on faith".

Finally, people appear to understand that they can regulate their beliefs through both evidential and nonevidential means. For instance, people who want to disbelieve something despite having evidence for it exercise mental control strategies designed to trigger beliefs nonevidentially. They avoid information that they think will force them to feel certain ways and try to focus on only the cues that will trigger beliefs that they want to hold (see Maio & Thomas, 2007, for a review). People may be well-aware of their tendency to manipulate their beliefs in these ways. That is, people may be well-aware that they believe something desirable without evidence because they knowingly sought out habitual and

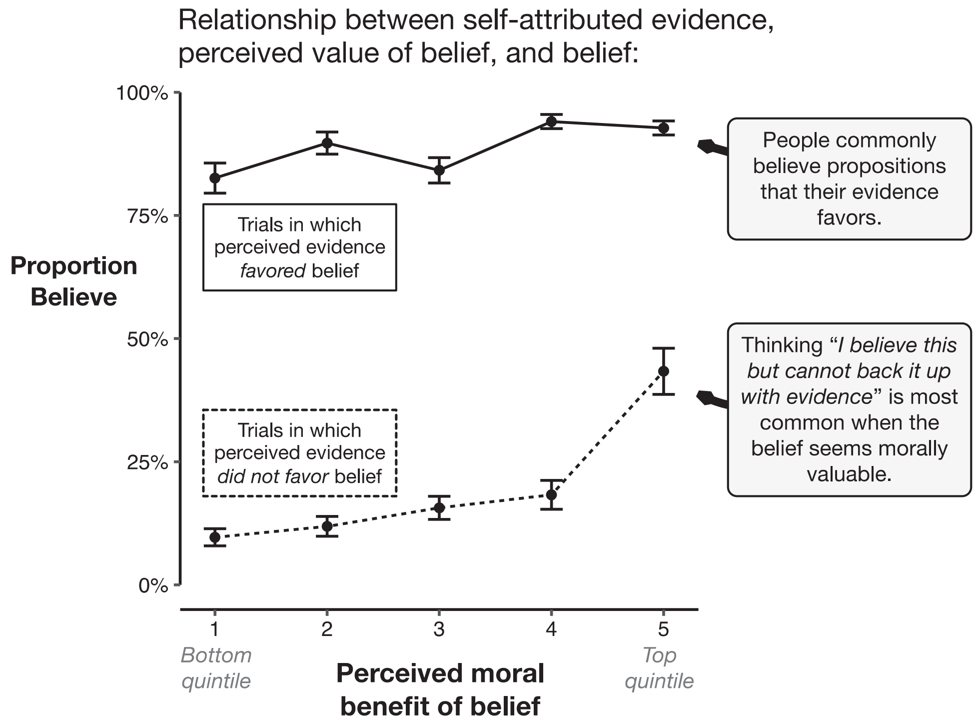

Figure 3. Data from Cusimano and Lombrozo (2023) documenting the relationship between self-attributed evidence, perceived moral value of belief, and belief.
Source. Adapted from Cusimano and Lombrozo (2023).

reflexive triggers to potentiate the belief. Each of these steps in belief formation—the evaluation of a desired belief as justified, the search for belief triggers, and their execution—would be transparent to the believer. And this transparency would not automatically extinguish belief because beliefs can be triggered and sustained even in the presence of metacognitive appraisals of poor evidence.

Evidence-Belief Dissociation: Summary

Another argument for the claim that people nearly universally believe that their beliefs are supported by evidence was the observation that beliefs seem to be involuntarily constrained by conscious appraisals of evidence. However, beliefs are not strictly governed by perceived evidence in this way. People sometimes believe things despite thinking that those beliefs are evidentially irrational. Examples include intuitions that people cannot explain or justify as well as some of the beliefs implicated in depression and anxiety. For these kinds of beliefs, people often act like the belief, in addition to seeming irrational, is undesirable and empty of other justifying qualities. In other words, people who hold these beliefs adopt a metacognitive position of "unjustified belief." These beliefs exist and persist because beliefs are sometimes the product of triggers that cause beliefs to behave like habits or reflexes. These triggers provide a means for understanding how people can also sometimes hold beliefs that they think are irrational but valuable—a metacognitive position called "value-based justification." And, consistent with this line of reasoning, recent studies document people who hold beliefs that they self-consciously cannot defend with evidence but that they think are good and valuable anyway.

Summary: Heterogeneity in Metacognitive Appraisals of Biased Beliefs

How do people evaluate their biased beliefs? A common and influential answer in cognitive and social psychology is that people do not distinguish their biased beliefs from their unbiased ones. Accordingly, when people reflect on their biased beliefs, they adopt an "illusion of objectivity" such that they believe those beliefs reflect an impartial weighing of evidence. This view of belief and metacognition is present in influential models of belief formation, motivated reasoning, bias attribution, and the psychological immune system. This view is derived from the observation that metacognitive appraisals of objectivity seem to constrain belief formation. On some accounts, this constraint is self-imposed such that the illusion of objectivity is a by-product of ubiquitous and successful efforts to be rational. On other accounts, this constraint is involuntary such that the illusion of objectivity is a by-product of beliefs spontaneously and automatically changing whenever conscious appraisals of evidence do. Recent findings demonstrate that metacognition does not constrain belief in either of these ways.

People do not view all biases, all the time, as bad. For instance, when people consider someone partially weighing evidence, such as by demanding stronger evidence for undesirable or risky propositions, people judge that person's reasoning as justified (Cusimano & Lombrozo, 2021a). Indeed, people think that others are reasoning poorly when they seem to "jump to conclusions" by accepting high-stakes conclusions without seeking out additional evidence. People's normative evaluations of others' beliefs extend to people's own beliefs (Cusimano & Lombrozo, 2023). People sometimes acknowledge and condone holding high-stakes beliefs to stricter standards of evidence. This metacognitive position, called "value-gated evidential justification," represents the first deviation from the illusion of objectivity.

People are also capable of sustaining beliefs despite thinking that they lack sufficient evidence for them. Beliefs sometimes have a habitual or reflexive quality in that they jump to mind even when people acknowledge that they cannot defend them with evidence. Because beliefs can sometimes form independently of appraisals of evidence, people can adopt two other metacognitive positions toward their beliefs. They can think that the belief altogether lacks meritorious qualities ("Unjustified belief") or that the belief makes up for its lack of evidence by being valuable in other ways ("Value-based justification").

In sum, people evaluate their biased beliefs in all sorts of ways. This variation in metacognition reflects the confluence of (a) variation in what qualities people want their beliefs to have, and (b) dissociable causes of belief. Metacognition neither universally imposes a standard of objectivity on belief nor does it universally constrain belief. Much more could be said. This review investigated only two notions of "bias" and reviewed only a small set of the known mechanisms and patterns underlying belief change. I suspect that people adopt an even wider variety of positions and that the positions discussed herein could be described with much more nuance. Indeed, the positions reviewed should be seen as oversimplified prototypes—real-life positions probably blend aspects of the positions above. But although the analysis above is coarse, it is nevertheless sufficient to yield important theoretical and practical upshots for people interested in studying, explaining, or changing belief. First, different kinds of belief may often associate with different metacognitive positions. Second, people who hold the same belief may occupy different metacognitive positions toward that belief. Third, different metacognition positions signal different belief formation processes. And finally, the specific position that someone takes toward their belief recommends specific interventions to change belief.

Variation in Metacognition Across Content, People, and Time

Metacognitive Variation Across Belief Content

Throughout this review, evidence for different metacognitive positions invoked different kinds of beliefs (Table 1). One direction for future research is to identify what kinds of beliefs tend to co-occur with different metacognitive positions. There are certain patterns we should expect. For instance, habitual and reflexive cognitive processes give rise to a limited range of beliefs thereby constraining what beliefs people can form without evidence. For instance, people might be able to maintain irrational beliefs about themselves, science, or God that they formed during childhood, but they may not be able to form those same beliefs (for the first time) as adults. Likewise, only a limited range of beliefs benefit believers, thereby limiting what beliefs people may judge as justified on moral or practical grounds. For instance, people may be likely to adopt a value-based justification position toward superstitious beliefs about relationships (e.g., "I have a soul mate") but not superstitious beliefs like tempting fate. Future work investigating people's belief in a particular domain would benefit from identifying what (if any) features of belief people tend to regard as valuable, and what norms of evidence people apply to that domain.

Metacognitive Variation Across Believers

Variation in metacognition may also be present among people who hold the same belief. Table 2 demonstrates this point by describing someone adopting different metacognitive positions for the same belief in God. Each version of Sebastian in Table 2 believes in God. However, each version thinks that he has different amounts of evidence for his belief, thinks that he needs different amounts of evidence, and thinks that different kinds of reasons speak in favor (or against) believing in God. One version of Sebastian believes that he is rational and impartial (consistent with the illusion of objectivity), while another believes that his belief is a justified act of faith. Despite their differences, each version of Sebastian is plausible. People within the same community, even the same church, might share the same belief but differ in the position they take toward it. Consistent with this idea, Cusimano and Lombrozo (2023) observed variation among people who held the same beliefs. Among people who

Table 1. Four Metacognitive Positions Organized by Their Characteristic Metacognitive Judgments.

	Metacognitive judgment			
Position	Justified?	Sufficient evidence?	Impartial reasoning?	Examples:
Unjustified Belief	-	-	-	- Magical contagion. - Maladaptive beliefs related to depression, phobia, anxiety.
Value-based Justification	✓	-	-	- Valued religious and karmic beliefs. - Morally desirable overconfidence.
Value-gated evidential justification	✓	✓	-	- Holding risky scientific findings to higher standards of evidence. - Giving friends the benefit of the doubt.
Objectivity Illusion	✓	✓	✓	- Classic examples of biased and motivated reasoning (e.g., self-serving biases).

Examples of the beliefs discussed in this review are provided.

believed in God, Karma, and ghosts, for instance, some people thought their belief was supported by evidence, others thought the belief was unsupported by evidence and overall unjustified, while still others thought their belief was justified despite lacking evidence.

Metacognitive Variation Across Time

Finally, we might observe variation in metacognitive positions within the same person, for the same belief, across time. Variation in metacognition over time might look something like this:

Sebastian grows up in a religious community and learns from people that he trusts that God exists (unbiased belief). As he grows older, he learns that this belief is controversial, and in the face of this controversy, unconsciously bolsters the apparent evidence for God (illusion of objectivity). At school, he learns arguments against God's existence. But he holds those arguments to extremely high scrutiny because it feels too risky to give up an important belief (value-gated evidential justification). He reaches a point where the evidence starts to point very strongly against his belief. So, to sustain his belief, he avoids his atheist friends, attends church more often, and prays more (value-based justification). Later, he becomes disillusioned with the church, leaves, and renounces his religious identity. But despite leaving that life behind, the feeling that God exists still lingers in his mind (unjustified belief). Only a long time later he happens to notice within himself a complete absence of feeling in God's existence (no belief).

This story is speculative. But without strict and simple constraints on belief and metacognition, it is also plausible. It is of course still possible that people have a general tendency to hold an illusion of objectivity about their biased beliefs. But when it is important to know why a particular person, in a particular place and time, believes what they do, we should regard it as an open question about how they relate to their belief. As I argue next, the position they take signals different belief-generating processes and recommends different persuasive strategies.

Metacognitive Position and Psychological Explanation

Many of the puzzles about belief are in fact puzzles about the metacognitive judgments that gird belief. For instance, one upshot of theories like the bias blind spot, and influential accounts of motivated reasoning reviewed in the sections above, has been that the primary puzzle to solve about many beliefs is how people can adopt them while maintaining the conviction that they are unbiased. In other words, one upshot of these theories was that the question "how did someone come to believe this idea?" should be substituted with the question "how did someone rationalize this idea?." But if people do not always judge themselves as unbiased, then this substitution is not always required or appropriate. Consider again Sebastian from Table 2. If Sebastian believes that his belief in God is unjustified, then we no longer must seek out explanations that conjure apparent evidential support for God.

Generalizing from this example, psychologists can diagnose the proximate cause of belief by identifying the metacognitive position that someone takes toward that belief (Table 3). When people adopt positions akin to the illusion of objectivity or value-gated evidential justification, their belief is most likely the product of gathering and appraising evidence. This quasi-scientific process is well-described in many places (e.g., Pyszczynski & Greenberg, 1987; Trope & Liberman, 1996). During this process, the proximate cause for belief is attending to information that seems diagnostic of some focal hypothesis. By contrast, metacognitive positions wherein the belief is judged to be unjustified, or justified based on its pragmatic or moral value, signal reflexive and habitual belief formation processes.

Table 2. Examples of How the Same Belief Can Manifest in Different Positions.

Belief in God:

Sebastian believes in God. He goes to church. He prays when no one is looking. When people ask him if he believes in God he says "yes." When he looks inward and evaluates his belief in God. . .

Position	Example of someone occupying that position
Unjustified belief	He knows good arguments against God's existence that he cannot refute. Indeed, he tried to leave the church himself. And, he does not bother trying to persuade anyone that God exists. He isn't so sure that he himself should believe. But sometimes he cannot shake the feeling that there is something out there bigger than himself. So, he goes along with it.
Value-based justification	He knows good arguments against God's existence that he cannot refute. However, believing in God is good for him. Believing helps him weather hard times, gives meaning to his life, and motivates him to act ethically. This is enough, in his mind, to protect and reinforce his gut belief in God.
Value-gated evidential justification	He does not have the same evidence for God that he has for his other beliefs, but he has all the evidence he needs to feel like he knows that God exists. He has heard some arguments against God, but they would have to be a lot stronger if he is going to change his whole life and risk an eternity in hell.
Illusion of objectivity	He has as much support for this belief as he has for any other belief. People he trusts believe in God; Isaac Newton believed in God; Barack Obama believes in God. Also, God explains why the universe exists, how complex intelligent life evolved, and how his friend was able to turn her life around. Anyone rational would come to the same conclusion.

The metacognitive position that someone adopts also may be a reliable signal about the role that their value splayed in biasing their belief. Consider again views wherein beliefs are constrained by metacognitive appraisals of objectivity. According to these views, this position was diagnostic of the immediate, proximate cause of belief (i.e., appraising the evidence called to mind). However, this position was highly undiagnostic of any functional or contributing factors that might have also explained the belief. It was therefore hard to say whether people who thought they were unbiased actually were. And, even if we had some reason to think that a person's values had affected their belief, that person was ill-placed to say precisely how. The analysis presented in this paper raises the possibility that metacognition can provide insight into the role of these forces.

This insight is possible because someone's values may often be both a common cause of their biased belief and their metacognition judgment. For instance, someone who knows that they want to feel optimistic because of the emotional benefits that optimism brings may seek out nonevidential triggers of belief. This person might suppress cues that trigger an undesired belief and seek out cues to trigger the desired one (Maio & Thomas, 2007). Their resulting metacognitive position will accurately account for bias because their belief regulation goal, and the actions they took to achieve that goal, were transparent to them. Likewise, a strong concern about a high-stakes belief is likely to affect whether someone accepts evidence for that belief. But this concern, by virtue of its strength, is also likely to be available to the believer during introspection. Consistent with this analysis, Cusimano and Lombrozo (2023) found that self-attributions of bias were sensitive to the actual presence of bias in people's reasoning. In their studies, people were much more likely to self-attribute biases—such as feeling concern around forming a belief that is risky or disrespectful—when they had in fact just formed (or withheld) a belief for those reasons.

Table 3 notes what each metacognitive position likely entails about the role (if any) of people's values and preferences in biasing belief. For instance, when people hold a value-based justification position toward their belief, they likely possess a motive to adopt and maintain a specific belief (like "God is real") that overrides a motivation to be accurate or rational. By contrast, when people adopt a value-gated evidential justification position, they need not have adopted this kind of motive. Instead of overriding a motive to be accurate, their values bias their reasoning by affecting the parameter of accuracy with which they are most concerned—either adopting a true belief or avoiding a false belief. For instance, when people feel pressure to be diligent, and not jump to conclusions about risky topics, they enter a state wherein they are especially concerned about avoiding believing something untrue. Relatively subtle differences in people's metacognitive position signal very different value-infected cognitive processes.

Of course, people may still be inaccurate about their biases. People might still operate under an illusion of objectivity in many situations. People might reason with a conscious goal to be impartial and to form rational beliefs, but fail to live up to this goal. Variation in metacognitive position does not guarantee accuracy. Moreover, people can be inaccurate about any of the metacognitive judgments I have reviewed. They may sometimes incorrectly self-attribute bias or self-attribute the wrong bias. This might happen, for

Table 3. Examples of Psychological Properties of Belief Associated With Each Metacognitive Position.

	Proximate belief formation mechanisms:		Role of motives or values in creating belief (if any):
Position	Label	Examples	Examples
Unjustified belief	*Habitual and reflexive triggers*	- Mnemonic access - Association - Innate concepts	-
Value-based justification	*Habitual and reflexive triggers*	- Same as above. - Automatic hyper-active agency detection.	- Motivates rituals that reinforce belief. - Information avoidance. - Motivates exposure to belief triggers.
Value-gated evidential justification	*Evidence gathering*	- Social hypothesis testing[a]	- Attend to risks of false acceptance/rejection during evidence gathering. - Appeal to evidence gathering norms like actionability, due diligence, and so on.
Illusion of objectivity	*Evidence gathering*	- Social hypothesis testing[a]	- Unconscious recall of desire-congruent evidence. - Unwitting rationalization of undesired information.

[a]I adopt the label "social hypothesis testing" from Trope and Liberman (1996). This label captures models of belief formation as a process of hypothesis generation and evidence gathering relative to evidential thresholds (e.g., Kruglanski, 1990; Pyszczynski & Greenberg, 1987; Trope & Liberman, 1996; see Figure 2).

instance, when their beliefs reflect their total evidence very well, but during introspection, they lack conscious access to that evidence. The analysis in this paper reveals not just the potential for accuracy in metacognition, but the potential for novel kinds of inaccuracy too. Nevertheless, on balance, psychologists should be more optimistic about the potential insight that metacognition provides into bias than they historically have been.

Metacognitive Positions and Persuasive Appeals

Because metacognitive positions signal what processes sustain belief, it should be helpful to diagnose someone's metacognitive position prior to attempting to change their belief. Here I discuss a few potential debiasing strategies and note how they might work, or not, given someone's metacognitive position.[9]

Exposure to New or Unappreciated Evidence

People often lack information or, as a result of narrow-minded or one-sided thinking, systematically neglect certain kinds of information available to them (Hoch, 1985; Koriat et al., 1980). For this reason, getting people to consider evidence or arguments for alternative views is a reliable way of changing someone's belief (Koriat et al., 1980; Lord et al., 1984). This strategy works best when people hold a metacognitive position, like the illusion of objectivity or value-gated evidential justification, that involves them believing that they have sufficient evidence for their belief. In these cases, getting someone to attend to new evidence works because it gives them reasons that, by their own lights, should motivate belief change. Of course, these strategies are not guaranteed to work (e.g., Lord et al., 1979; Lord & Taylor, 2009). Among many other reasons, people can often easily rationalize new information so that it no longer challenges their beliefs (Gershman, 2019). But this is not the only reason that these interventions might fail. We should expect these interventions to fail also when someone already believes that their beliefs are not justified by evidence. In this case, people think that the evidence is largely irrelevant to what they should believe (Value-based justification) or they already agree that they should change their mind but cannot do so (Unjustified belief).

Bias Education

When people have adopted an illusion of objectivity, dispelling their illusion should be a useful way to change their mind. And indeed, scientists have tried improving reasoning by teaching people about their unconscious tendency to think in self-serving ways, to neglect certain sources of information, to make certain kinds of miscalculations when collecting and appraising evidence, and to hold undesirable conclusions to stricter standards than desirable ones. But if people are not naïve to their bias, and they think that their beliefs are justified anyway, then bias education interventions have no reason to work. Telling a religious person that they believe in God because their belief gives their life meaning is not going to change their mind if they already know that or if they think that is a perfectly acceptable reason to believe something. When people are not blind to their biases, different tactics are required.

Appeal to the Value of Impartial, Evidence-Based Reasoning

Scholars sometimes appeal to the value of impartial, evidence-based reasoning as a reason why others should change what they believe. For instance, recent general audience books by Don Moore (2020) and Steven Pinker (2021) spend some time trying to convince their readers that they *ought* to try to be impartial and to form evidence-backed beliefs. For instance, one argument in favor of impartial, evidence-based reasoning is that it leads to better decisions. People generally want to make good decisions, so this observation makes a pretty good argument! But even if this observation makes a good argument, it could never really impact someone if everyone already thinks that they have reasoned objectively. Indeed, if reasoning is constrained by a need to think of oneself as unbiased, then arguments about the virtues of impartial, evidence-based reasoning do not matter (see also discussion in Alston, 1988). And in particular, these arguments would be irrelevant to the project of improving thinking.

But people can self-attribute partiality or evidential irrationality. And they can manipulate themselves into and out of these metacognitive positions by seeking out or neutering the right belief triggers. We therefore have some reason to think that people can, in principle, be incentivized to maintain beliefs that are impartial or based on evidence. In other words, the sorts of appeals that Moore (2020) and Pinker (2021) make may in fact be effective in changing how people think. And indeed, there is some suggestive evidence that these appeals may be able to change people's beliefs: People who endorse stricter norms of evidence-based reasoning tend to hold more scientific (and less religious and superstitious) beliefs (e.g., Pennycook et al., 2020; see reviews in Cusimano & Lombrozo, 2021b; Stahl & Cusimano, 2023). However, it is a valuable project for future research to test when changing someone's standards for belief affects their belief.

Appeal to the Cost of Error

One effective strategy for changing how people reason, and generally improving accuracy, is to raise the cost of error. When the stakes are high, people engage in more elaborate and open-minded reasoning, and their beliefs tend to be more accurate (Kruglanski, 2004; Trope & Liberman, 1996). Holding someone accountable for their judgment, such that they must defend their judgment to an intelligent and critical judge, is a decent way to get others to form more sophisticated and accurate beliefs (Lerner & Tetlock, 1999). As we reviewed earlier, people sometimes acknowledge that their beliefs have been influenced by these kinds of situational pressures. This observation is important to debiasing in two ways. First, appeals to "what the evidence objectively says" will fail if the believer knows that their stubbornness reflects an error management strategy. Consider the task of convincing a coffee drinker to believe that coffee is unhealthy. This person might acknowledge that there is evidence that coffee is unhealthy but also believe that they should withhold judgment until they get more information because the cost (to them) of getting it wrong feels high. It would be folly to try and change this person's belief by pointing out to them that they are holding the belief to a higher-than-normal evidential standard due to their idiosyncratic cares—they already know. At the same time, citing the costs of error may convince someone to loosen or tighten their standards of evidence, and in so doing, cause them to feel more, or less, certain in their belief. It might be helpful to point out to the coffee addict the potentially neglected risks of failing to accurately identify coffee as unhealthy.

Habituation

Finally, people attempting various persuasive and debiasing appeals would benefit from knowing that there is sometimes a nonrational, nonmotivational lag between metacognitive judgments and belief change. People may not change their minds even after a persuasive appeal successfully convinces them that their belief is unjustified. For instance, a person with a severe phobia does not lose the phobia merely upon encountering persuasive arguments that their phobia is irrational. This lesson is most apparent in clinical psychology, where cognitive change requires repeated exposure to counterarguments and the gradual build-up up of new thoughts that compete with old ones. This lesson also ought to apply to the beliefs that social and cognitive psychologists typically care about because these beliefs sometimes have the same character. For instance, a notable proportion of people who believe in Karma or ghosts already judge their beliefs as irrational and unjustified.

If people already judge their beliefs to be unjustified, then some debiasing strategies will be more effective than others. For instance, "bias education" will fail to be useful—again—because it does not provide the target with new information or motivation. A more promising strategy is to change the behaviors that reinforce belief, such as by eliminating behaviors that promote the intuition (like taking part in rituals) and replacing them with behaviors that promote competing intuitions (like practicing calling to mind arguments that point out the irrationality of the belief).

Summary

Different metacognitive positions recommend different strategies for changing belief. Bias education may be useful when people want to maintain unbiased beliefs and are suffering from a "bias blind spot," but this strategy is unlikely to be effective when people relate to their beliefs in other ways. The debiasing strategy that works best depends on which position people take toward their belief. Our discussion of what strategies likely work best is speculative as little work has tested alternative debiasing strategies. Thus, each of the alternatives above represent promising avenues for future work.

Application to Conflict and Disagreement

The reasons why different persuasive strategies may fail based on the believer's metacognitive position also illuminate why disagreements may fail to resolve. Although the illusion of objectivity is often conceived as a constraint on belief, it also has obvious application to ideological conflict (Pronin et al., 2004; Ross, 2018; Ross & Ward, 1996). Accordingly, people do not accede to their opponent's view because they assume that they themselves are informed and unbiased while their opponent is uninformed or biased.

However, given people's capacity to adopt a variety of metacognitive positions, disagreements may be difficult to resolve for other reasons, too. When two people have adopted different value-gated evidential justification positions, for instance, they might agree about how much evidence there is, but disagree about how much evidence there needs to be. For instance, one person might look at the current evidence about adolescent gender-affirming care and think that accepting it would be tantamount to jumping to conclusions, while another person might look at the same evidence and think that, given the urgency of the problem, the evidence is more than strong enough (Cusimano & Lombrozo, 2023). Neither person denies that their values are affecting their judgment; instead, both believe that their value-based reason for accepting or rejecting the evidence is morally superior. A similar line of reasoning applies when two people disagree and one of them holds a value-based justification for belief. An atheist and religious fundamentalist might agree that there is no good evidence for God but still disagree about whether believing in God is justified. In this case, the fundamentalist, who believes in faith, would not cede their belief although they would be attributing bias to themselves rather than to their opponent. People trapped in disagreements with others should not assume that the other person is operating under an illusion of objectivity.

Statement on Citations, Generalizability, and Author Positionality

This work argues for variation in people's everyday metacognitive evaluation of belief. Most of the work that this review engages with was conducted by a relatively homogeneous group: academic psychologists working in primarily western contexts. These scholars place high value on impartial, evidence-based reasoning. It was within this community that the theories I discuss—which posits a near-universal metacognitive position dominated by a concern for unbiased, evidence-based reasoning—were developed and became influential. Likewise, the studies that demonstrated support for these theories, by and large, were conducted on college students at competitive U.S. schools. As a result, one limitation of this review is that both the work documenting the illusion of objectivity and the work documenting its alternatives relied on participants from the usual western samples. Another limitation stems from constraints on my own thinking (which has been similarly shaped by WEIRD culture). In my attempt to delineate different metacognitive positions, I have relied on notions of "belief" "evidence" "rationality" and "impartiality" that make the most sense to me. But these concepts—their meaning and implicit standards—surely vary across communities. The analysis will not apply well to people who interpret these terms differently. It is unclear to me, at this moment, how well the analysis here generalizes to non-Western cultures.

Conclusion

How do people evaluate and relate to their biased beliefs? It depends. The same belief can be judged as sensible, lovely, or neither, and can be held to standards that demand objectivity or demand loyalty, confidence, or care. This variation in metacognition is the confluence of several features of belief. First, people hold their beliefs to varying standards of evidence and impartiality. Second, beliefs have multiple, competing proximate causes. Third, the type of thinking that gives rise to different kinds of biases and different kinds of beliefs may often give rise to corresponding metacognitive beliefs. As a result, people may sometimes have insight into their biases. Scientists can leverage metacognition to figure out why people believe what they do and what it would take to change their minds. Metacognition is useful to the study of belief because metacognitive judgments vary in ways that reflect, and so reveal, the complexity of belief.

Acknowledgments

I thank Sara Aronowitz, Geoffrey Goodwin, Emily Foster-Hanson, Jack Keefe, Ike Silver, and Kerem Oktar for providing comments on a prior draft, and Tania Lombrozo and Cindy Frantz for providing comments on several drafts. I thank Stefan Schubert for pointing me toward the dialogue in *Brideshead Revisited* that formed the epigraph for this essay. And special thanks are owed to several anonymous reviewers who repeatedly provided extremely helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Corey Cusimano https://orcid.org/0000-0002-4980-7839

Notes

1. Because this constraint (as described) is a by-product of people's goals, it can be thought of as kind of a voluntary, self-imposed constraint. Later, in the section "Self-consciously believing against the evidence," I review a closely related line of reasoning according to which a belief is constrained this way involuntarily. Scholars do not always distinguish between these two kinds of constraint. Indeed, Kunda (1990), as well as many of the others cited in this section, sometimes describe this constraint as a voluntary one and sometimes as an involuntary one.
2. I am ignoring here the ways that this process can be biased in the sense that it results in beliefs that do not reflect someone's total evidence. People frame their deliberation in ways that favor certain hypotheses over others, neglect prior probabilities when evaluating the diagnostic value of some piece of information, and often seek out evidence in one-sided or close-minded ways.
3. Kruglanski and colleagues refer these motives as "need for closure" and "fear of invalidity," respectively. I have opted for the terms related to actionability and due diligence because these terms are more intuitive to readers unfamiliar with this literature.
4. Some normative views in epistemology and the philosophy of science state that people should require stricter standards of evidence for risky propositions compared with less risky ones (Douglas, 2000; Rudner, 1953). For instance, if it would be particularly risky to believe that IQ correlates with race, then scientists hold investigations into that question to an especially high standard before accepting such a statement as true (Bolinger, 2020; Cusimano & Lombrozo, 2021b; Douglas, 2021). This example is much more serious than the coffee example—the risks here are to society rather than to one's morning routine—but the principle of shifting the demand for evidence is the same.
5. The introspective feeling that beliefs are uncontrollable in light of evidence is shared by many scholars (e.g., Alston, 1989; Elster, 1979; Epley & Gilovich, 2016; Festinger, 1957; James, 1937; see Cusimano & Goodwin, 2019, 2020). These introspective reports have played in influential role in theories of belief: Many scholars have argued that beliefs are involuntarily constrained by evidence because that is how they feel. Ordinary people also hold tend to think that beliefs are involuntarily constrained by evidence (Cusimano et al., 2024).
6. People may sometimes infer evidence for their belief by drawing on their naïve theories about their intuitions. For instance, people might (reasonably) assume that the ease with which information comes to mind is a reliable signal of its truth (Alter & Oppenheimer, 2009), or they might have background beliefs about how reliable their spontaneous flashes of insight are (Inbar et al., 2010). To wit: If something feels sufficiently real, it must do so on account of unknown-but-strong reasons. The evidence reviewed below suggests that intuitions sometimes persist despite people failing to rationalize them in this way.
7. Ellis (1962) provides an anecdote of this dynamic with one of his depressed patients (p. 24):

   '"So you still think," I said to the patient (for perhaps the hundredth time), "that you're no damned good and that no one could possibly fully accept you and be on your side?"

   "Yes, I have to be honest and admit that I do. I know it's silly, as you keep showing me that it is, to believe this. But I still believe it; and nothing seems to shake my belief."

   "Not even the fact that you've been doing so much better, for over a year now, with your husband, your associates at the office, and some of your friends?"

   "No, not even that. I know I'm doing better, of course, and I'm sure it's because of what's gone on here in these sessions . . . But I still feel basically the same way—that there's something really rotten about me."
8. The International Classification of Diseases, 10th Edition (ICD-10) states that one diagnostic criterion for phobia is that the individual recognizes their fear as disproportionate; the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5*; American Psychiatric Association, 2013) states the same criterion for social anxiety disorder. For OCD, the *DSM*-5 distinguishes between people who have "poor" "fair" or "good" insight.
9. It may not always be a good idea to try to debias reasoning. The biases discussed in this paper are plausibly normative (Cusimano & Lombrozo, 2021b).

References

Abelson, R. P. (1986). Beliefs are like possessions. *Journal for the Theory of Social Behaviour*, *16*(3), 223–250. https://doi.org/10.1111/j.1468-5914.1986.tb00078.x

Alston, W. P. (1988). The Deontological Conception of Epistemic Justification. *Philosophical Perspectives*, 2, 257–299.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, *110*(3), 486–498. https://doi.org/10.1037/0033-2909.110.3.486

Armor, D. A., Massey, C., & Sackett, A. M. (2008). Prescribed optimism: Is it right to be wrong about the future. *Psychological Science*, *19*(4), 329–331. https://doi.org/10.1111/j.1467-9280.2008.02089.x

Atran, S., & Norenzayan, A. (2004). Religion's evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioral and Brain Sciences*, *27*(6), 713–730. https://doi.org/10.1017/s0140525x04000172

Baddeley, A. D. (1990). The development of the concept of working memory: Implications and contributions of neuropsychology. In G. Vallar & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory* (pp. 54–73). Cambridge University Press.

Baker, L. A., & Emery, R. E. (1993). When every relationship is above average: Perceptions and expectations of divorce at the time of marriage. *Law and Human Behavior*, *17*(4), 439–450. https://doi.org/10.1007/bf01044377

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of*

Personality and Social Psychology, *91*(4), 612–625. https://doi.org/10.1037/0022-3514.91.4.612

Barber, J. P., & DeRubeis, R. J. (1989). On second thought: Where the action is in cognitive therapy for depression. *Cognitive Therapy and Research*, *13*, 441–457. https://doi.org/10.1007/BF01173905

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

Baron, J. (2019). Actively open-minded thinking and politics. *Cognition*, *188*, 8–18.

Baron, J., Baron, J. H., Barber, J. P., & Nolen-Hoekseman, S. (1990). Rational thinking as a goal of therapy. *Journal of Cognitive Psychotherapy*, *4*(3), 293–302. https://doi.org/10.1891/0889-8391.4.3.293

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends Cognitive Sciences*, *4*(1), 29–34. https://doi.org/10.1016/s1364-6613(99)01419-9

Barrett, J. L., & Lanman, J. A. (2008). The science of religious beliefs. *Religion*, *38*(2), 109–124.

Baumeister, R. F., & Newman, L. S. (1994). Self-regulation of cognitive inference and decision processes. *Personality and Social Psychology Bulletin*, *20*(1), 3–19. https://doi.org/10.1177/0146167294201001

Beck, A. T. (1979). *Cognitive therapy and the emotional disorders*. Penguin.

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry*, *165*(8), 969–977. https://doi.org/10.1176/appi.ajp.2008.08050721

Beck, A. T., Rush, A. J., Shaw, B. T., & Emergy, G. (1979). *Cognitive therapy of depression*. Guilford Press.

Bhatia, S. (2017). Conflict and bias in heuristic judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(2), 319–325. https://psycnet.apa.org/doi/10.1037/xlm0000307

Block, L., & Kramer, T. (2009). The effect of superstitious beliefs on performance expectations. *Journal of the Academy of Marketing Science*, *37*(2), 161–169. https://doi.org/10.1007/s11747-008-0116-y

Bolinger, R. J. (2020). Varieties of moral encroachment. *Philosophical Perspectives*, *34*(1), 5–26.

Boyer, P. (2003). Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Sciences*, *7*(3), 119–124. https://doi.org/10.1016/S1364-6613(03)00031-7

Brewin, C. R. (2006). Understanding cognitive behaviour therapy: A retrieval competition account. *Behaviour Research and Therapy*, *44*(6), 765–784. https://doi.org/10.1016/j.brat.2006.02.005

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, *63*(4), 515–533.

Cusimano, C., & Goodwin, G. P. (2019). Lay beliefs about the controllability of everyday mental states. *Journal of Experimental Psychology: General*, *148*, 1701–1732.

Cusimano, C., & Goodwin, G. P. (2020). People judge others to have more voluntary control over beliefs than they themselves do. *Journal of Personality and Social Psychology*, *119*, 999–1029. https://doi.org/10.1037/pspa0000198

Cusimano, C., & Lombrozo, T. (2021a). Morality justifies motivated reasoning in the folk ethics of belief. *Cognition*, *209*, 104513.

Cusimano, C., & Lombrozo, T. (2021b). Reconciling scientific and commonsense values to improve reasoning. *Trends in Cognitive Sciences*, *25*, 937–949.

Cusimano, C., & Lombrozo, T. (2023). People acknowledge and condone their own morally motivated reasoning. *Cognition*, *234*, 105379. https://doi.org/10.1016/j.cognition.2023.105379

Cusimano, C., Zorrilla, N., Danks, D., & Lombrozo, T. (2024). Psychological freedom, rationality, and the naïve theory of reasoning. *Journal of Experimental Psychology: General*, *153*(3), 837–863.

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, *66*, 819–829.

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568–584. https://doi.org/10.1037/0022-3514.63.4.568

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*(4), 559–579. https://doi.org/10.1086/392855

Douglas, H. (2021). *The rightful place of science: Science, values, and democracy*. Consortium for Science, Policy & Outcomes.

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin*, *31*(5), 680–692. https://doi.org/10.1177/0146167204271570

Ellis, A. (1962). *Reason and emotion in psychotherapy*. Lyle Stuart.

Elster, J. (1979). *Ulysses and the sirens: Studies in rationality and irrationality*. Cambridge University Press.

Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, *30*(3), 133–140. https://doi.org/10.1257/jep.30.3.133

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

Foley, R. (1987) *The theory of epistemic rationality*. Cambridge, MA: Harvard University Press.

Frantz, M. C. (2006). I am being fair: The bias blind spot as a stumbling block to seeing both sides. *Basic and Applied Social Psychology*, *28*(2), 157–167. https://doi.org/10.1207/s15324834basp2802_5

Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review*, *100*(2), 298–319. https://doi.org/10.1037/0033-295X.100.2.298

Gelman, S. A., & Legare, C. H. (2011). Concepts and folk theories. *Annual Review of Anthropology*, *40*, 379–398.

Gendler, T. S. (2008). Alief and belief. *The Journal of philosophy*, 105(10), 634–663.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. https://doi.org/10.3758/s13423-018-1488-8

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, *75*(3), 617–638. https://doi.org/10.1037/0022-3514.75.3.617

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. Free Press.

Hansen, K., Gerbasi, M., Todorov, A., Kruse, E., & Pronin, E. (2014). People claim objectivity after knowingly using biased strategies. *Personality and Social Psychology Bulletin*, *40*(6), 691–699. https://doi.org/10.1177/0146167214523476

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Henrich, J. (2009). The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, *30*(4), 244–260.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 719–731. https://doi.org/10.1037/0278-7393.11.1-4.719

Hume, D. (2017). *Enquiry concerning human understanding* (J. Bennett, Trans.). https://www.earlymoderntexts.com/assets/pdfs/hume1748_1.pdf (Original work published 1793)

Inbar, Y., Cone, J., & Gilovich, T. (2010). People's intuitions about intuitive insight and intuitive choice. *Journal of Personality and Social Psychology*, *99*(2), 232–247. https://doi.org/10.1037/a0020215

James, W. (1937). The will to believe. In *The will to believe, and other essays in popular philosophy* (pp. 1–31). Longmans, Green and Co.

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social metacognition: An expansionist review. *Personality and Social Psychology Review*, *2*(2), 137–154.

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

Kelemen, D. (2004). Are children "intuitive theists"? Reasoning about purpose and design in nature. *Psychological Science*, *15*(5), 295–301.

Kelemen, D., & DiYanni, C. (2005). Intuitions about origins: Purpose and intelligent design in children's reasoning about nature. *Journal of Cognition and Development*, *6*(1), 3–31.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143.

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, *34*(6), 833–848. https://doi.org/10.1177/0146167208315158

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Kozak, M. J., & Foa, E. B. (1994). Obsessions, overvalued ideas, and delusions in obsessive-compulsive disorder. *Behaviour Research and Therapy*, *32*(3), 343–353. https://doi.org/10.1016/0005-7967(94)90132-5

Kruglanski, A. W. (1990). Lay epistemic theory in social-cognitive psychology. *Psychological Inquiry*, *1*(3), 181–197. https://doi.org/10.1207/s15327965pli0103_1

Kruglanski, A. W. (1996). Motivated social cognition: Principles of the interface. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 493–520). The Guilford Press.

Kruglanski, A. W. (2004). *The psychology of closed mindedness*. Psychology Press.

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology*, *19*(5), 448–468. https://doi.org/10.1016/0022-1031(83)90022-7

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lane, R. D., Ryan, L., Nadel, L., & Greenberg, L. (2015). Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*, *38*, e1. https://doi.org/10.1017/S0140525X14000041

Laurin, K., & Kay, A. C. (2017). The motivational underpinnings of belief in God. In J. M. Olson (Ed.), *Advances in experimental social psychology* (pp. 201–256). Elsevier Academic Press.

Legare, C. H., Evans, E. M., Rosengren, K. S., & Harris, P. L. (2012). The coexistence of natural and supernatural explanations across cultures and development. *Child Development*, *83*(3), 779–793.

Legare, C. H., & Gelman, S. A. (2008). Bewitchment, biology, or both: The co-existence of natural and supernatural explanatory frameworks across development. *Cognitive Science*, *32*(4), 607–642.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, *2*(3), 166–167. https://doi.org/10.1038/s41562-018-0301-z

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243. https://doi.org/10.1037/0022-3514.47.6.1231

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

Lord, C. G., & Taylor, C. A. (2009). Biased assimilation: Effects of assumptions and expectations on the interpretation of new evidence. *Social and Personality Psychology Compass*, *3*(5), 827–841. https://doi.org/10.1111/j.1751-9004.2009.00203.x

Luhrmann, T. M. (2020). *How God becomes real: Kindling the presence of invisible others*. Princeton University Press.

Maio, G. R., & Thomas, G. (2007). The epistemic-teleologic model of deliberate self-persuasion. *Personality and Social Psychology Review*, *11*(1), 46–67. https://doi.org/10.1177%2F1088868306294589

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, *102*(1), 72–90. https://doi.org/10.1037/0033-2909.102.1.72

Mayseless, O., & Kruglanski, A. W. (1987). What makes you so sure? Effects of epistemic motivations on judgmental confidence.

Organizational Behavior and Human Decision Processes, *39*(2), 162–183. https://doi.org/10.1016/0749-5978(87)90036-7

McAllister, D. W., Mitchell, T. R., & Beach, L. R. (1979). The contingency model for the selection of decision strategies: An empirical test of the effects of significance, accountability, and reversibility. *Organizational Behavior and Human Performance*, *24*(2), 228–244. https://doi.org/10.1016/0030-5073(79)90027-8

Miller, J. E., Park, I., Smith, A. R., & Windschitl, P. D. (2021). Do people prescribe optimism, overoptimism, or neither? *Psychological Science*, *32*(10), 1605–1616.

Miller, J. E., Strueder, J. D., Park, I., & Windschitl, P. D. (2023). Do people desire optimism from others during a novel global crisis? *Journal of Behavioral Decision Making*, Article e2362.

Moore, D. A. (2020). *Perfectly confident: How to calibrate your decisions wisely*. HarperCollins.

Murray, S. L., & Holmes, J. G. (1997). A leap of faith? Positive illusions in romantic relationships. *Personality and Social Psychology Bulletin*, *23*(6), 586–604. https://doi.org/10.1177/0146167297236003

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition. *Metacognition: Knowing about Knowing*, *13*, 1–25.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295x.84.3.231

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgment and Decision Making*, *15*(4), 476–498.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335–346. https://doi.org/10.1016/j.cognition.2012.03.003

Pinker, S. (2021). *Rationality: What it is, why it seems scarce, why it matters*. Viking.

Porot, N., & Mandelbaum, E. (2021). The science of belief: A progress report. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(2), e1539.

Pronin, E. (2009). The introspection illusion. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, Vol. 41, pp. 1–68). Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(08)00401-2

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781–799. https://doi.org/10.1037/0033-295X.111.3.781

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (*Vol. 20*, pp. 297–340). Academic Press. https://doi.org/10.1016/S0065-2601(08)60417-7

Reeder, G. D., Pryor, J. B., Wohl, M. J. A., & Griswell, M. L. (2005). On attributing negative motives to others who disagree with our opinions. *Personality and Social Psychology Bulletin*, *31*(11), 1498–1510. https://doi.org/10.1177/0146167205277093

Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review*, *123*(2), 182–207. https://doi.org/10.1037/rev0000017

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, *95*(2), 293–307. https://doi.org/10.1037/0022-3514.95.2.293

Robbins, T. W., Vaghi, M. M., & Banca, P. (2019). Obsessive-compulsive disorder: Puzzles and prospects. *Neuron*, *102*(1), 27–47. https://doi.org/10.1016/j.neuron.2019.01.046

Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naïve realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, *68*(3), 404–417. https://doi.org/10.1037/0022-3514.68.3.404

Rochat, P., Morgan, R., & Carpenter, M. (1997). Young infants' sensitivity to movement information specifying social causality. *Cognitive Development*, *12*(4), 537–561.

Rogers, T., Moore, D. A., & Norton, M. I. (2017). The belief in a favorable future. *Psychological Science*, *28*(9), 1290–1301. https://doi.org/10.1177/0956797617706706

Rosenzweig, E. (2016). With eyes wide open: How and why awareness of the psychological immune system is compatible with its efficacy. *Perspectives on Psychological Science*, *11*(2), 222–238. https://doi.org/10.1177/1745691615621280

Ross, L. (2018). From the fundamental attribution error to the truly fundamental attribution error and beyond: My research journey. *Perspectives on Psychological Science*, *13*(6), 750–769. https://doi.org/10.1177%2F1745691618769855

Ross, L., & Ward, A. (1996). Naïve realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *The Jean Piaget symposium series. Values and knowledge* (pp. 103–135). Lawrence Erlbaum.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology*, *50*(4), 703–712. https://doi.org/10.1037/0022-3514.50.4.703

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, *20*(1), 1–6. https://doi.org/10.2307/185617

Sharot, T., Rollwage, M., Sunstein, C. R., & Fleming, S. M. (2023). Why and when beliefs change. *Perspectives on Psychological Science*, *18*(1), 142–151. https://doi.org/10.1177/17456916221082967

Sherman, D. K., Cohen, G. L., Nelson, L. D., Nussbaum, A. D., Bunyan, D. P., & Garcia, J. (2009). Affirmed yet unaware: Exploring the role of awareness in the process of self-affirmation. *Journal of Personality and Social Psychology*, *97*(5), 745–764. https://doi.org/10.1037/a0015451

Shtulman, A., & Harrington, K. (2016). Tensions between science and intuition across the lifespan. *Topics in Cognitive Science*, *8*(1), 118–137.

Shtulman, A., & Legare, C. H. (2020). Competing explanations of competing explanations: Accounting for conflict between scientific and folk explanations. *Topics in Cognitive Science*, *12*(4), 1337–1362.

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. In D. Barner & A. S. Baron (Eds.), *Core knowledge and conceptual change* (pp. 53–71). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190467630.003.0004

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*, 209–215.

Siepmann, M., Baron, J., Steinberg, K., & Sabini, J. (2004). *Disbelieved beliefs: Subjective estimates of bias in probabilistic beliefs and their relationships to desire* [Unpublished manuscript].

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Sloman, S. A., Fernbach, P. M., & Hagmayer, Y. (2010). Self-deception requires vagueness. *Cognition*, *115*(2), 268–281. https://doi.org/10.1016/j.cognition.2009.12.017

Sprecher, S., & Metts, S. (1989). Development of the "romantic beliefs scale" and examination of the effects of gender and gender-role orientation. *Journal of Social and Personal Relationships*, *6*(4), 387–411. https://doi.org/10.1177/0265407589064001

Stahl, T., & Cusimano, C. (2023). Lay standards for reasoning predict people's acceptance of suspect claims. *Current Opinions in Psychology*, *55*, 101727.

Ståhl, T., Zaal, M. P., & Skitka, L. J. (2016). Moralized rationality: Relying on logic and evidence in the formation and evaluation of belief can be seen as a moral issue. *PloS one*, *11*(11), e0166332.

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin*, *116*(1), 21–27. https://doi.org/10.1037/0033-2909.116.1.21

Tenney, E. R., Logg, J. M., & Moore, D. A. (2015). (Too) optimistic about optimism: The belief that optimism improves performance. *Journal of Personality and Social Psychology*, *108*(3), 377–399. https://doi.org/10.1037/pspa0000018

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451–471. https://doi.org/10.1037/0033-295X.109.3.451

Trémolière, B., & Djeriouat, H. (2019). Love is not exactly blind, at least for some people: Analytic cognitive style predicts romantic beliefs. *Personality and Individual Differences*, *145*, 119–131. https://doi.org/10.1016/j.paid.2019.03.025

Trope, Y., & Liberman, A. (1996). Social hypothesis testing: Cognitive and motivational mechanisms. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 239–270). The Guilford Press.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004

Van Tongeren, D. R., DeWall, C. N., Chen, Z., Sibley, C. G., & Bulbulia, J. (2021). Religious residue: Cross-cultural evidence that religious psychology and behavior persist following deidentification. *Journal of Personality and Social Psychology*, *120*(2), 484–503. https://doi.org/10.1037/pspp0000288

Walco, D. K., & Risen, J. L. (2017). The empirical case for acquiescing to intuition. *Psychological Science*, *28*(12), 1807–1820. https://doi.org/10.1177/0956797617723377

Waugh, E. (1945). *Brideshead revisited*. Little, Brown and Company.

Wegener, D. T., Silva, P. P., Petty, R. E., & Garcia-Marques, T. (2012). The metacognition of bias regulation. In P. Briñol & K. DeMarree (Eds.), *Frontiers of social psychology. Social metacognition* (pp. 81–99). Psychology Press.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, *103*(3), 506–519. https://doi.org/10.1037/a0028857

White, C. J. M., & Norenzayan, A. (2019). Belief in karma: How cultural evolution, cognition, and motivations shape belief in supernatural justice. In J. M. Olson (Ed.), *Advances in Experimental Social Psychology* (pp. 1–63). Elsevier Academic Press.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142. https://doi.org/10.1037/0033-2909.116.1.117