Synthese (2021) 199:7575–7594
https://doi.org/10.1007/s11229-021-03129-0

Beliefs and biases

Shannon Spaulding[1]

Received: 3 December 2020 / Accepted: 17 March 2021 / Published online: 27 March 2021

Abstract
Philosophers are divided over whether implicit biases are beliefs. Critics of the belief model of implicit bias argue that empirical data show that implicit biases are habitual but unstable and not sensitive to evidence. They are not rational or consistently action-guiding like beliefs are supposed to be. In contrast, proponents of the belief model of implicit bias argue that they are stable *enough*, sensitive to *some* evidence, and do guide our actions, albeit haphazardly sometimes. With the help of revisionary notions of belief, such as fragmented, Spinozan, and dispositional belief, these theorists argue that implicit biases are beliefs. I argue that both the critiques and defenses of belief models of implicit bias are problematic. This methodological critique suggests that debates about nature of the implicit bias ought to shift away from the belief question and toward more fundamental questions about stability and evidential sensitivity of implicit biases. I chart the path forward for this prescribed shift in the debate.

1 Introduction

Suppose that I assert that men and women are equally intelligent and hardworking and that individuals' differences in intelligence and diligence are not due to their sex. Nevertheless, I subtly behave as if I think men are more knowledgeable and competent than women. When I listen to two experts, even if I acknowledge that they are both well-educated and intelligent, the male expert typically seems *more* authoritative than the female expert. When I look at two otherwise equivalent resumés, even if I judge both resumés to be stellar, the male resumé just seems *more* impressive to me than the female resumé. These are cases in which my explicit, consciously endorsed beliefs are egalitarian, but my behavior indicates that I have an implicit sexist bias.

A pressing issue in discussions of implicit bias is whether implicit biases are stable and sensitive to evidence. If I have an implicit sexist bias, is it likely to manifest

✉ Shannon Spaulding
Shannon.spaulding@okstate.edu

[1] Philosophy Department, Oklahoma State University, Stillwater, Oklahoma, USA

consistently across time and across contexts, or does it frequently fluctuate, sometimes exhibiting a strong force and other times not? In other words, are implicit biases more like cholesterol levels or heart rates? Another important question is whether implicit biases are responsive to evidence. Can I change my implicit sexist bias by learning more about sex, gender, and intelligence, or is my implicit bias unresponsive to rational intervention? These two features—stability and evidential sensitivity—are at the core of a debate about the nature of implicit bias.[1]

I address this debate in this paper. Some philosophers argue that implicit biases are unstable and are not sensitive to evidence that bears on their truth in the way we expect beliefs to be, and thus implicit biases are more like mere associations or character traits than beliefs. Proponents of the belief model of implicit bias argue that implicit biases are sensitive to some evidence, and lots of representational states we commonly identify as beliefs display similar kinds of instability and insensitivity to evidence. I argue that the critics and proponents of the belief model of implicit bias face a troubling dilemma, which is mirrored in other similar debates, e.g., belief models of cognitive delusions, self-deception, and religious avowals. The way out of this dilemma, I suggest, is to get a more fine-grained understanding of the content of the representation in question. Without that, we cannot say whether a bias is a stable or sensitive to evidence. Thus, the debate about whether implicit biases are beliefs puts the cart before the horse. We ought to instead focus our efforts on developing a better understanding of what makes biases unstable (when they are), context sensitive (when they are), and insensitive to evidence (when they are).

2 Implicit biases

To understand implicit bias, it is important to understand social categorization, which is our tendency to sort people, behaviors, and events into social categories. Social categorization is reflexive and rapid. We cannot help but see a person in terms of their social category, and with some highly salient categories, social categorization is very fast. Within 100 ms of seeing a face, we can sort people by age, gender, and race (Ito et al., 2004; Liu et al., 2002). Although age, race, and gender often are the most salient social categories, we rapidly sort people into numerous social categories, e.g., nationality, sexual orientation, religious affiliation, etc. Our situational context, cognitive load, and goals in a social interaction influence what is salient to us, and what is salient determines which social categories we employ when we reflexively sort people into social groups (Wheeler & Fiske, 2005; Gilbert & Hixon, 1991).

[1] Throughout this paper, I use *responsiveness* and *sensitivity* to evidence interchangeably. As an anonymous reviewer notes, some epistemologists distinguish these terms. However, in the ordinary language usage I adopt here, a representation that is unresponsive to evidence is insensitive to evidence and vice versa, and a representation that is responsive to evidence is also sensitive to evidence and vice versa.

The categories we use to sort people, behavior, and events are associated both explicitly and implicitly with various features.[2] These features may be affectively neutral "semantic" features or affectively laden features. For example, some people tend to associate BLACK and + ATHLETIC or FEMALE and + WARMTH.[3] Explicit associations are relatively easy to test because usually subjects can reflect on and report their explicit associations. You can simply ask people what characteristics they associate with social categories and try to control for social desirability self-censorship. This may require a bit of creativity especially for social categories associated with negatively valenced features, but it is at least clear that subjects have access to their explicit associations. Implicit associations are more difficult to test because subjects typically are unable to directly report their implicit associations.[4] For this reason, experimenters construct tasks that are designed to elicit behavior that is sensitive to implicit associations, and from the elicited behavior they estimate subjects' implicit bias.[5]

We can divide these tasks into two broad categories: those that detect differential patterns in *deliberative* behavior and those that detect differential patterns of *spontaneous, rapid* behavior. With respect to deliberative behavior, some experiments examine how subjects evaluate resumés that are identical except for the name on the resumé (Bertrand & Mullainathan, 2004; Isaac et al., 2009; Moss-Racusin et al., 2012). If subjects give lower ratings to resumés with stereotypical female names than stereotypical male names, this is evidence that they have an implicit sexist bias. Some priming experiments detect implicit bias from patterns in deliberative, explicit judgments. For instance, Graham and Lowery (2004) primed juvenile probation and detention officers with words related to the racial category Black and then asked them to evaluate a hypothetical offender, whose race is unspecified. Subjects primed with words related to the racial category Black rated the hypothetical offender as having a worse personality, being more blameworthy, more likely to reoffend, and

[2] Although I describe explicit and implicit biases in terms of *associating* a category with a feature or valence, I do not presuppose a side in the associationism vs. propositionalism debate. I use the term *association* neutrally to describe correlations between mental representations. Whether these correlations are ultimately cashed out in terms of associationist networks or propositional attitudes is an open question that I address toward the end of this paper.

[3] In practice, there is likely to be significant overlap between neutral semantic associations and affective associations (Holroyd and Sweetman, 2016). See also fn. 5.

[4] Subjects may have indirect awareness of the content of their own implicit biases. If you tell subjects that experimenters can detect bogus self-reports (Nier 2005) or ask subjects to predict their implicit biases (Hahn et al. 2014), subjects' scores on explicit and implicit measures more closely align. These studies do not show that subjects have direct awareness of their implicit biases—that they can simply read off what their implicit biases are—but it does add nuance to the claim that we are unaware of our implicit biases. In some circumstances, we at least have indirect awareness of our own biases.

[5] Other ways of estimating implicit bias focus more on the bias of groups and less on measuring the bias of individuals. For example, Payne et al. (2017) recent work conceives of implicit bias as a group-level phenomenon that passes through individual minds, much like the wave at sporting events. Payne, et al. argue that we should think of implicit bias in terms of concept accessibility and thus focus on the situational contexts that evoke implicit bias rather than on tests of individual levels of bias. On this view, tests of individuals' implicit bias measure how situations evoke concepts for certain populations.

they recommended harsher punishments. This is an indication of racial implicit bias, detected in patterns of deliberative reasoning.

Other implicit bias tests examine patterns of spontaneous behavior in rapid decision making. For instance, shooter bias studies show participants a visual scene for less than a second, and then instruct participants to decide whether the person in the scene was holding a gun or a neutral object, e.g., a cell phone (Payne, 2001). If subjects are more likely to misidentify the object as a gun when it is held by a Black person than a White person, they have an implicit racist bias. The Implicit Association Test (IAT) measures how quickly and accurately subjects categorize stereotypic and counter-stereotypic associations (Greenwald et al., 1998, 2009). In one version of the IAT, subjects are instructed to categorize as quickly and accurately as possible pictures of old people with pleasant words (e.g., beautiful, fantastic, happy) and pictures of young people with unpleasant words (e.g., pain, hate, annoy). Subjects are then instructed to categorize the stimuli according to the opposite rule: old faces with unpleasant words and young faces with pleasant words. If subjects categorize faster and more accurately according to one of these rules, they are said to have an ageist implicit bias.[6] Some priming tasks fall into this category, as well. For instance, Affective Priming tasks (Klauer & Musch, 2003) present subjects with a prime stimulus that may have a positive or negative valence and then they are presented with a target. Participants are typically instructed to ignore the prime and categorize the target as positive or negative. It is easier to categorize the targets if the prime stimulus matches the valence of the target. Categorizing the negatively-valenced targets faster when exposed to a certain priming stimulus indicates a negative implicit bias with respect to that stimulus category.

Psychologists use a variety of tests of deliberative and spontaneous behavior to examine implicit biases. Though there are many important and relevant distinctions between these tests, the overarching goal of these experimental approaches is to see how associations between social categories (e.g., WHITE, MUSLIM, or ELDERLY), valences (i.e., positive or negative), and features (e.g., + DANGEROUS, + PEACEFUL, or + INCOMPETENT) influence our slow, careful reasoning and our split-second reactions in ways that we may not be able to detect or directly control.[7]

[6] The Go/No-Go Association Task is similar to the IAT but with a simpler task design (Nosek and Banaji, 2001). In this kind of task, subjects are asked to categorize one type of target with a feature. For instance, they may be asked to categorize instances of fruit (e.g. apples, bananas, etc.) as good and ignore non-fruit targets. And then subjects are asked to categorize according to the opposite rule, e.g., fruit as bad. Like the IAT, if subjects are faster and more accurate at categorizing certain target stimuli and features, they have a stronger association between those features and categories.

[7] In principle, we can distinguish associations between categories and valences ("affective associations") from associations between categories and features ("semantic associations"), however in practice this distinction falters as many semantic associations are valenced. See footnote 1.

3 Beliefs

Now that we have a relatively brief overview of implicit bias, we can examine beliefs. What are beliefs and how can we tell whether some phenomenon is a belief? At a minimum, beliefs are attitudes that are supposed to accurately represent how the world is. We can think of them in terms of a vehicle (a propositional attitude) that has a certain content, and when that content accurately represents the world the belief is true. This thin characterization of belief, while relatively uncontroversial, is not particularly helpful in this context. What we need is a more comprehensive conception of belief. Though, in order to avoid begging any questions, we need an account that is not significantly revisionary to ground our discussion of the metaphysics of implicit bias. I will use Peter Railton's (2014, 2018) conception of belief. Here are the features of belief Railton articulates:

1 Representational
2 True or false
3 Mind-to-world direction of fit
4 Degrees of strength
5 Spontaneously, non-inferentially action-guiding
6 Spontaneously, non-inferentially thought-guiding
7 Spontaneously, non-inferentially feeling-guiding
8 Implicit/unconscious as well as conscious
9 Phenomenologically thin
10 Spontaneously projective and evidence-responsive
11 Absent changes in evidence, experience, beliefs inertial and context insensitive
12 Relational and intensional
13 Non-volitional
14 Spontaneously resistant to instrumentalization

Although some of these features probably are essential to belief (e.g., 1–3), I will not make that case here. I present this simply as a list of characteristic features of belief to help us navigate the arguments for and against conceiving of implicit biases as beliefs. As we shall see, the list helps us see what proponents and opponents of the belief model of implicit bias really disagree about.

When I sincerely claim that men and women are equally intelligent but nevertheless behave as if I think men are more intelligent than women, I have a sexist implicit bias. One may wonder: Do I really *believe* that men are more intelligent than women? Many say no (Gendler, 2008a, 2008b; Levy, 2014; Machery, 2016; Sullivan-Bissett, 2019). The arguments vary, but we can identify two major reasons why critics claim that implicit biases are not beliefs.[8]

[8] Here I set aside skepticism about the concept of belief itself. Such skeptical views hold that the folk psychological concept of belief does not track anything useful or interesting in psychology, and we should not analyze psychological phenomena in terms of belief. My argument targets views that hold that belief is a legitimate construct but question whether implicit biases should count as beliefs.

First, implicit biases do not seem to be sensitive to evidence like beliefs are.[9] Providing evidence that men and women are equally intelligent will not change an individual's sexist implicit bias.[10] In fact, many interventions on implicit biases—e.g., vividly imagining a counterstereotype exemplar or shifting group boundaries through competition—curtail the bias only for a few hours or days (Lai et al., 2016). Tamar Gendler argues, "Beliefs change in response to changes in evidence; aliefs [implicit biases are a kind of alief] change in response to changes in habit. If new evidence won't cause you to change your behavior in response to an apparent stimulus, then your reaction is due to alief rather than belief" (2008b, p. 566). Others make similar arguments, e.g., Grace Helton (forthcoming), Alex Madva (2016), Neil Levy (2014), and Ema Sullivan-Bissett (2019). These arguments conclude that implicit biases are unlike beliefs, which are sensitive to relevant changes in evidence. In other words, implicit biases lack feature 10.[11]

The second reason critics doubt that implicit biases are beliefs is that implicit biases seem to be too unstable to count as beliefs (Machery, 2016).[12] Suppose you take an IAT multiple times, and one time it says your association between elderly and negative features is moderate, another time it is strong, and another time it is slight. This kind of fluctuation is common, which suggests that whatever implicit bias tests are measuring is highly variable (Gawronski et al., 2017). Relatedly, experimental measures that purport to tap into the same implicit bias only weakly correlate with each other (Cameron et al., 2012; Greenwald et al., 2009). As Edouard Machery argues, "if different indirect measures really tapped into the same implicit attitude, we would expect much larger correlations than those found" (2016, p. 116). The low-to-medium test–retest reliability rates and weak-to-moderate correlations amongst measures of implicit bias suggest that implicit biases are too unstable to count as beliefs. In other words, implicit biases lack feature 11.[13]

The evidential invulnerability and instability of implicit biases, these theorists argue, disqualify implicit biases from counting as beliefs. The main arguments against conceiving of implicit biases as beliefs concern features 10 and 11. However, if implicit biases lack features 10 and 11, this will have downstream consequences

[9] Those who advocate for fragmented, Spinozisitic, or dispositional beliefs can explain the findings on insensitivity to evidence described in this paragraph. I address these revisionary notions of belief below.

[10] In another version of this kind of argument, critics point out that implicit biases are not sensitive to the *logical form* of the evidence (Madva, 2016). For instance, subjects form equivalent implicit attitudes on the basis of information and the negation of that information (Gawronski et al., 2008). Relatedly, Levy (2014) argues that sensitivity to evidence and inferential promiscuity are two sides of the same coin and that implicit biases lack both of these features.

[11] Critics of the belief model do not maintain that genuine beliefs are always perfectly sensitive to evidence and that implicit biases never respond to evidence. Rather, they argue beliefs typically are responsive to (some) epistemically relevant evidence, and implicit biases do not seem to be similarly responsive. One could present this debate as a question about the threshold for evidential sensitivity. I will not do that.

[12] Indeed, Machery argues that implicit biases are too unstable to even be *attitudes*. Machery argues that implicit biases are traits, not mental states.

[13] Again, one could have this debate turn on establishing a threshold for "stability," but I will not do that. Instead I will argue below that we must understand what information a bias encodes in order to properly evaluate its stability.

for other features. For instance, if implicit biases are unstable, they will not consistently, spontaneously guide feelings, thoughts, and actions (5–7). Indeed, if implicit biases are completely unresponsive to evidence one might question whether they even have a mind-to-world direction of fit (3). Philosophers who have argued that implicit biases are not beliefs defend different sorts of positive accounts of implicit bias: aliefs (Gendler, 2008a, b), patchy endorsements (Levy, 2014), character traits (Machery, 2016), unconscious imaginings (Sullivan-Bissett, 2019), etc.

In contrast, several philosophers have argued that implicit biases are beliefs (Egan, 2008, 2011; Mandelbaum, 2015; Quilty-Dunn & Mandelbaum, 2018; Schwitzgebel, 2013). Implicit biases are sensitive to *some* kinds of evidence (Dessel et al., 2018; Houwer, 2014; Hu et al., 2017), even if they are not perfectly sensitive to all relevant evidence. But this is also true of representations that we have no trouble calling beliefs. Ordinary beliefs (such as my belief that the people who raised me are my biological parents), aesthetic beliefs (like my belief that my children are the sweetest, cleverest, cutest children around), value-laden beliefs (e.g., that separating children from their parents when they are seeking asylum is cruel and immoral), conspiratorial beliefs (e.g., that Bill Gates engineered the COVID-19 pandemic) or superstitious beliefs (e.g., that walking under ladders is bad luck) are all selectively resistant to evidence and thorough rational deliberation, but we still call them all beliefs.[14] Proponents of the belief model of implicit bias argue that implicit biases have *enough* of the relevant features of belief to count as beliefs.

Let's step back for a bit to consider the dynamics of this debate. The question at the center of the debate is whether some psychological phenomenon counts as a belief. It is belief-like in some ways but not belief-like in other ways. This kind of question arises in structurally similar ways in several other debates, and we can learn something about why the belief question matters and what is really at stake in this debate by considering these structurally similar debates. Consider cognitive delusions. Subjects suffering from delusions in some ways act as if they believe the content of their delusions, e.g., Capgras patients will sincerely report that their loved ones have been replaced by impostors and feel uneasy or anxious around the "impostors." But in other ways, delusional subjects do not act as if they *really* believe the content of their delusions, e.g., most Capgras patients do not file missing persons reports, try to find their "real" loved ones, or follow through the logic of their delusion. This divergent behavior has led to a debate about whether delusional subjects really believe the content of their delusion (Bayne & Pacherie, 2005; Bortolotti, 2010) or not (Currie, 2000; Dub, 2017; Stephens & Graham, 2004).

Similarly, a central debate about self-deception centers on whether self-deceived subjects really believe their own deception. In a standard case of self-deception, a self-deceiver acts as if she believes P (the unwarranted but wanted proposition) in some circumstances, but she acts as if she believes ~P (the warranted but unwanted proposition) in other circumstances. Thus, the question arises whether she truly

[14] For an interesting take on superstitious and confabulatory attitudes, see Ichino (2020) who argues that many such attitudes are *not* in fact beliefs, though they are commonly regarded as such. For a similar argument about conspiracy theories, see (Ichino and Räikkä, 2020).

believes both P and ~P (Lynch, 2012; Mele, 2001), or has some other type of attitude toward these propositions (Egan, 2009; Funkhouser, 2005; Gendler, 2007).

To canvas just one more example, consider the debate about whether religious convictions are beliefs. On the one hand, religious individuals sincerely assert that they believe the content of their religious convictions and in some ways act in line with those convictions, e.g., they may attend religious services, answer affirmatively when asked whether they believe in God, and—in some contexts at least—think, feel, and act just as we would expect believers to act. However, in other ways it seems as if (at least some) religious individuals do not treat their religious convictions as factual beliefs. They may act as if they believe their religious convictions only at religious services. Their answers to factual questions may vary depending on whether they are in a secular or religious context. They are motivated to behave piously to avoid God's punishment only in certain settings; in other settings this motivation does not arise. They do not reject their religious convictions when there is strong evidence that they are mistaken. This discrepant behavior leads to the familiar debate over whether this psychological phenomenon really is a belief (Levy, 2017) or not (Leeuwen 2014, 2017a, b).

In each of these areas of research—implicit bias, delusions, self-deception, and religious convictions—philosophers debate whether the psychological phenomenon in question is a belief or something else.[15] In terms of Railton's list of characteristic features of belief, philosophers in each of these debates primarily disagree about whether the psychological phenomena in question are responsive to evidence (feature 10) and stable (feature 11).

4 An apparent dilemma

As the previous section illustrates, the question of whether some psychological phenomenon is a belief arises in many debates. Why in general do philosophers care whether some psychological phenomenon is a belief? The allure of a belief model is that *if* the psychological phenomenon in question is a belief, this gives us some traction on rational evaluation. On most accounts, the functional role of belief is to represent the world accurately. This is captured in features 1–3 of Railton's list. In light of this functional role, theorists propose several different rationality constraints on beliefs: one's beliefs ought to be consistent with one's other closely related beliefs, they ought to be sensitive to evidence but stable absent evidential challenges, inferentially coherent, and we ought to regard them as at least epistemically possible. Because beliefs have a mind-to-world direction of fit, they are subject to certain

[15] With delusions and religious convictions, the mental state in question is explicit. With implicit bias, and sometimes with self-deception, the mental state in question is implicit. This difference does not change the dynamics of the belief debates that I describe in the next section.

kinds of rational scrutiny. These rationality constraints are represented in Railton's features 1–3, 5–7, and 10–11.[16]

Furthermore, if the psychological phenomenon in question is a belief, this should allow us to effectively explain, predict, and intervene on the believer's actions. Though the connection between belief and action is rarely direct (in part because our actions are guided and motivated by other mental states, as well), knowing what a target believes *does* shed light on how she has or will act. (At least that's the presupposition of the enormous interdisciplinary literature on folk psychology.) And knowing that a target has a belief, and that beliefs typically are sensitive to evidence, should allow us to construct more targeted interventions. These aspects of belief are reflected in features 5–7 and 10–11 of Railton's list.

Thus, the allure of a belief model of implicit bias is that it gives us some traction on questions about rational evaluation of our implicit biases and a basis to explain, predict, and intervene on behavior. If the psychological phenomena described above are beliefs, then we have a straightforward way to answer questions about their rationality, figure out why people act as they do, and perhaps construct appropriate interventions.

The difficulty with a belief model is that the ordinary conception of belief (exemplified by Railton's features 1–3) does not seem to map neatly onto messy psychological phenomena. As we saw in the discussion of implicit bias, delusions, self-deception, and religious convictions, these psychological phenomena can exhibit *some* important features of belief and apparently lack others. As a result, some philosophers have proposed revisionary notions of belief. Implicit biases—and delusions, self-deception, religious convictions—they argue are fragmented beliefs (Egan, 2008, 2011), or in-between beliefs (Schwitzgebel, 2013), or Spinozistic beliefs (Mandelbaum, 2014, 2015).

It appears as if there is a dilemma lurking here. Some behavior indicates that subjects' implicit biases are belief-like, and other behavior indicates that subjects' implicit biases are not belief-like. It seems there are two general options: we can refine the concept of the psychological phenomenon in question to answer the belief question *or* refine the concept of belief in order to answer the belief question. Both of these moves have been attempted in all of the debates canvased above, and both of these moves generate serious problems.

Consider how these moves have played out in the implicit bias literature. In pursuit of the revise-the-concept-of-the-psychological-phenomenon strategy, some argue that implicit biases are too unstable and insensitive to evidence to count as beliefs, so they are something akin to mere associations (Gendler, 2008a, 2008b), patchy endorsements (Levy, 2014), character traits (Machery, 2016), etc. Beliefs are supposed to be insensitive to irrelevant evidence and sensitive to relevant evidence.

[16] Belief is not the only rationally evaluable mental state. Some argue for rationality constraints on imagination (Currie, 2010; Doggett & Egan, 2012), desire (Audi, 2001; Smith, 1994; Wrenn, 2010), and emotions (Na'aman, 2020; Prinz, 2004). Rational evaluation for these mental states will hinge on different features than rational evaluation of belief. For instance, Audi grounds rational evaluation of desires in terms of desiring the good, and Smith evaluates rationality of desire in terms of what an ideal agent would desire.

That much is clear. But, how can we tell whether evidence is relevant? Implicit biases typically are characterized at a superficial level, associating a social category (e.g., ELDERLY) with a feature (e.g., + INCOMPETENT). However, when you do not know the precise informational content of a representation, it is difficult to tell whether it is sensitive to epistemically relevant evidence. Tests of implicit bias can give us some of the elements of the representation—that a feature is associated with social category—but this underdetermines the actual content of the representation.[17]

An example will help to illustrate the problem. Suppose that a person asserts that the 2020 US Presidential election was fraudulent. If you show her a court case that examined and dismissed these allegations, she will not reject the claim that the election was fraudulent. In fact, even if you presented her with information that all the court cases and recounts vindicated the overall tally of votes, she probably would not feel compelled to reject her claim. But that does not mean that her representation "The 2020 US Presidential election was fraudulent" is not a belief. It is sensitive to relevant evidence. After all, if former President Donald Trump, Fox News, NewsMax, One America News Network, and other *trusted* sources testified that the election was legitimate, she *would* revise the claim. In this case, what *counts* as relevant evidence is in dispute. Part of the content of the representation that the election was fraudulent is the idea that there was a massive cover up evidence of electoral malfeasance. That is why it is regarded as fraudulent and not just an innocent mistake. Thus, the appearance of insensitivity to evidence is misleading in this case. When one has a more fine-grained understanding of the content of the representation, one can see what would count as relevant evidence.

Consider one more example. Suppose someone asserts that pit bulls are dangerous. At this level of description, it will sometimes seem as if the subject believes this (perhaps we can find evidence of this with an adapted priming task or IAT). But other times, it will seem as if the subject does not believe this. If you present the subject with pit bull puppies, he will not think or act as if they are dangerous. If you tell him about "nanny dogs"—gentle, loving pit bulls that look after children—he will not think or act as if they are dangerous. The subject is not representing *all* pit bulls as dangerous. Rather, he represents a certain subset of pit bulls—those that are neglected, raised to fight, and abused by their owners—as dangerous because of their upbringing and natural strength. This stereotype of a pit bull may be prominent for the subject, so he may simply assert that "pit bulls are dangerous" when what he really means is that this property applies to a salient subset of pit bulls. Thus, the mixed behavior may look like instability when it actually indicates a more nuanced content.

These two examples demonstrate that a representation may seem unresponsive to evidence or unstable across contexts when we do not know the precise informational content of the representation. Knowing the informational content is necessary to understand what evidence is relevant. Without that, we cannot say whether a representation is unstable or insensitive to evidence. In other words,

[17] Del Pinal and Spaulding (2018) make this case very clearly. I will address this account in the next section.

without knowing the content, we cannot make a judgment on the vehicle. Critics of the belief model consider associations between BLACK-MEN and + DANGEROUS or ELDERLY and + INCOMPETENT, inferred from implicit bias tests, to be unresponsive to relevant evidence and unstable. But given the complex relation between beliefs and evidence, especially when we do not know the exact informational content of the biases, it seems inappropriate to conclude that the representation is unstable and insensitive to evidence and therefore not a belief. In short, these arguments underestimate the complex relations between evidence and belief. Thus, the arguments for rejecting the belief model in favor of a new model of implicit bias seem faulty.

In the refine-the-concept-of-belief camp, some argue that the concept of belief employed by philosophers does not capture the nuances of typical human psychology, so we should adopt a revisionary conception of belief. On this view, the fact that a subject does not behave a particular way is not a good enough reason to deny that she believes X. After all, human psychology and behavior are messy.

In order to make this move plausible, one needs to supply an alternative model of belief. There are many options here: different versions of dispositionalism (Quine, 1960; Davidson, 2001; Stalnaker, 1984), Spinozistic beliefs (Mandelbaum, 2014), in-between beliefs (Schwitzgebel, 2013), and fragmented beliefs (Egan, 2008; Stalnaker, 1984) The downside of *this* kind of move is that it mitigates the motivation for the belief model: appealing to revisionary notions of belief to make sense of messy psychological phenomena puts separation between belief, rational evaluation, and action explanation, prediction, and intervention (Quilty-Dunn & Mandelbaum, 2018).

Philosophers have in some cases developed accounts of rationality for these revisionary notions of beliefs. However, rational evaluation becomes quite nuanced and difficult for Spinozistic, fragmented, or in-between, or dispositional beliefs, and the connection to action is even more indirect. In terms of Railton's features of beliefs, Spinozistic, fragmented, in-between, and dispositional beliefs less clearly exhibit features 5–7 because they only sometimes spontaneously guide actions, thoughts, and feelings. Moreover, they do not clearly exhibit feature 10 because only sometimes will they be spontaneously projective and evidence responsive. Thus, while a revisionary notion of belief may be able to capture the messy psychological phenomena—indeed that is what these accounts are constructed to explain—they also relinquish straightforward connections to rationality and action. My argument is not that these revisionary accounts of belief are wrong. That kind of argument would look really different than this. Rather, I am arguing that the motivation to posit a revisionary notion of belief in the case of implicit bias cuts against the reason for wanting a belief model in the first place.

Thus, to put the dilemma more directly, in answering the belief question there are two unappealing options: (1) No, X is not a belief because it does not behave like a typical belief; instead it is Y; and (2) Yes, X is a belief, but on some refined conception of belief. Option (1) as it is typically advanced seems to underestimate the complex relation between evidence and belief. Option (2) mitigates the initial appeal of the belief model because it seriously complicates the connection to rational evaluation and action explanation and prediction.

This dilemma suggests that the debate about whether implicit biases are beliefs is premature. What we need prior to answering that question is an account of the informational content of implicit biases. In the next section, I will sketch one account of the informational content of implicit biases. Once we have such an account, we can see the way forward for both the proponents and critics of the belief model of implicit bias.

5 Beliefs and biases

Critics of the belief model of implicit bias argue that implicit biases are unresponsive to evidence and unstable. Most proponents of the belief model of implicit bias rebut these charges by adopting a revisionary conception of beliefs (e.g., implicit biases are fragmented, or in-between, or Spinozistic beliefs). In this section, I will illustrate how understanding the informational structure of a social bias illuminates the kind of evidence that would be relevant to a social bias. First, I will present one account of the informational structure of implicit biases. Then, I will demonstrate how different informational structures yield different implications for stability and sensitivity to evidence.

In a previous paper, with Guillermo Del Pinal and I offer a useful framework for understanding the different kinds of content a bias may have (Del Pinal & Spaulding, 2018).[18] Consider the representation that women are family oriented. One's association between WOMAN and + FAMILY-ORIENTED may encode salience, e.g., family-orientation is *prominent* in one's conception of women, or the inference that a woman is family-oriented is readily *available*. For example, when you think of women, family just springs to mind. Let's call biases with this informational structure *salient biases.* Alternatively, one's association between WOMAN and + FAMILY-ORIENTED may encode statistical information. One's representation that women are family-oriented may encode cue validity (women represent a higher percentage of all the people who are family-oriented) or typicality (most women are family-oriented). Let's call biases that encode this kind of information *statistical biases*. Finally, one's representation that women are family-oriented may encode causal-explanatory relations. For instance, family-orientation may be more central to one's conception of women than other features associated with women insofar as being family-oriented causes or explains other features associated with women. Let's call biases with this informational structure *causal-explanatory biases.*[19]

[18] Our theoretical framework is based on robust empirical evidence on how information is encoded in concepts. I divide the ways of encoding information a bit differently than in the previous paper, but the general framework is the same.

[19] A common objection is that implicit biases cannot represent causal-explanatory relations because implicit biases are the product of System 1 cognition, and only explicit, controlled System 2 cognition can track causal and explanatory relations. This objection reflects an empirically outdated take on implicit cognition. See Sloman and Lagnado (2015) and Kurdi et al. (2020) for empirical evidence that implicit cognition encodes causal and explanatory information.

On this theoretical framework, there are three different kinds of informational content a representation can have. A representation can encode salient information, statistical information, or causal-explanatory information.[20] Existing experimental paradigms do not—and are not designed to—pull apart these different kinds of biases, however. This is unfortunate, Del Pinal and Spaulding argue, because these three kinds of biases have different behavioral profiles (Sloman et al., 1998). Both salient and statistical biases are highly context sensitive. Indeed, it is sometimes difficult to distinguish salient and statistical biases because they often go hand-in-hand. For example, +FEMININE is both a prominent and cue valid feature for the category WOMAN. That is, women are more likely to exhibit femininity than men, and femininity is a salient feature of women. However, it is important for this debate to be able to distinguish merely salient biases from biases that are statistical or both salient and statistical. Salient biases are extremely malleable and can change with even slight contextual shifts. These representations are tracking prominence and availability of a feature to a concept, and which features are prominent or available is a function of one's specific context. Thus, *merely* salient biases by their nature are unlikely to be stable and evidentially sensitive and, hence, unlikely candidates for belief.

Statistical biases are highly context sensitive for a different reason than salient biases. Statistical biases track typicality and cue validity, i.e., the probability of the feature given the category and the probability of the category given the feature. These probabilities will vary when you change the context or change the level of categorization. For instance, the feature +FAMILY-ORIENTED may be typical for the basic-level category WOMAN, but it may be less typical for FEMALE WALL STREET BANKER**.** In different contexts, different features are typical and cue valid. Implicit salient biases and implicit statistical biases breakdown in sub-categorization and contextual change for very different reasons. Statistical biases break down in predictable ways that are explicable in terms of the content and level of categorization of the bias. The breakdown of salient biases is more arbitrary. Thus, if we want to know whether biases are merely salient, we need to turn to empirical evidence on how biases behave in context change.

In contrast to both salient and statistical biases, casual-explanatory biases are more stable across contextual shifts. Suppose again that your context is a Wall Street banker. In this context, you might not as easily or quickly infer that women are family oriented as you do with the basic-level category WOMAN. However, you might still infer that women are nurturing even in the Wall Street banker context. Of course, this is simply a hypothetical example to illustrate the concepts here. But, if this were the case, +NURTURING would be more central than +FAMILY-ORIENTED for

[20] Though I have presented this three-way distinction in terms of different ways of *associating* a category with a feature, I intend the talk of "association" here to be neutral. That is, I am not presupposing an answer to the associationism vs. propositionalism debate. Whether the representations described by salient, statistical, and causal explanatory biases are *merely* associations or propositions is an open question. Indeed, below I present a way for belief models of implicit bias to use this framework to establish that statistical and causal explanatory biases are beliefs. If one finds the terminology in the framework distracting, one could substitute *correlation* for *association* without any loss of explanatory power.

the social category WOMAN, which means that we are more likely to infer that women in many different contexts are nurturing. Centrally encoded features are more likely to persist in sub-categorization and contextual change than merely salient features or statistical features. Thus, causal-explanatory biases are likely to be more resilient across contexts than other types of biases.

With this framework at our disposal, we can see that implicit biases with these three informational structures will behave very differently from each other. An example will help to illustrate these differences. Suppose that A associates women with empathy. In A's representation, empathy is a deep property of women because it has a relatively central causal-explanatory role. It explains many other the properties A associates with women (why many women take on mentoring roles in the workplace, why they are overrepresented in human service jobs, etc.). Because + EMPATHETIC is a central feature in A's representation of the category WOMAN, A will spontaneously invoke it to explain other properties of women. Because it is a central feature, it can take quite a lot of evidence to uproot that bias. Nevertheless, it should be responsive to the right kind of evidence. For example, suppose there is a set of conditions C that A regards as optimal enabling conditions to manifest empathetic behavior. Suppose A then finds out that a large subset of women under conditions C do not manifest empathetic behavior. That is, A encounters evidence of a group of neurotypical women, raised in typical, often patriarchal families, who take on mentoring roles in the workplace, become teachers and caregivers, are mothers, but are relatively impoverished in empathy. In these optimal enabling conditions, women lack empathy while exhibiting the features that empathy was supposed to cause or explain. If A accepts this evidence, A should then revise their bias. On the other hand, suppose A discovers that the majority of women are not empathetic, not more likely to be empathetic than men, not more likely to voluntarily take on mentoring roles at work, etc. This statistical information does not challenge the informational content of the belief because the content of the bias is not about the statistical distribution of empathetic behavior in women. In response to such statistical evidence we would expect A to retain the belief that "women are empathetic," perhaps adding a hypothesis about enabling conditions that explains away that new information. Thus, if the bias that associates WOMAN with + EMPATHETIC encodes causal-explanatory information, this is how we would expect it to behave.

Suppose, as a second example, that B has a bias that associates WOMAN with the feature + BAD-AT-MATH. Assume that, in B's representation of this bias, being bad at math is not a deep property of women, i.e., it is not a part of their biological essence, nor does it play a central role in the causal-explanatory networks that connect other properties in the representation of women. However, B tends to believe that (i) most women are bad at math and (ii) they are much more likely to be bad at math than men (and other relevant gender/social categories). In other words, being bad at math is both relatively typical and cue valid for women but does not have a deep causal-explanatory role. Suppose B is presented with information that women are not biologically more disposed to be bad at math than men. This evidence that women are not innately less skilled with math would not challenge the content of the bias. On the other hand, suppose B encounters evidence that women score higher on standardized math tests, that women's grades in math classes are better than men's

grades when teachers grade anonymously, etc. If B accepts this evidence, it would challenge his bias that women are bad at math in a way that information about innate biological traits would not. Thus, given the content of this bias, we would expect this kind of information to result in revising the bias, e.g., the property of being bad at math will be slightly less typical for women, slightly more for men, and accordingly less cue valid for women.

Putting these two examples together, we can see that the same kind of information, statistical information, is relevant to the content of the second bias that women are bad at math but not directly relevant to the first bias that women are empathetic. Similarly, information on innate, biological dispositions is relevant to the first bias that women are empathetic but not the second bias that women are bad at math. We can see how easy it would be to mistakenly conclude that these two biases are unstable and not responsive to evidence and therefore not beliefs if we are not cognizant of the different behaviors of statistical and causal-explanatory biases.

This account of the informational structure of implicit bias is, of course, just one possible account. However, the attractive feature of this particular account for our purpose here is that it explicitly delineates how differently encoded biases ought to behave in response to various sorts of evidence and shifts in context. And it does this in a way that does not presuppose an answer to the belief question about implicit bias. Given that the debate about whether implicit biases are beliefs is, at bottom, a question of whether beliefs are sensitive to changes in relevant evidence (feature 10) and resilient in the face of irrelevant evidence (feature 11), this framework is particularly helpful for answering that question.

Using this framework, we can see the work that proponents and critics of the belief model of implicit bias need to do in order to make their case. Proponents of the belief model need to establish that most implicit biases are causal-explanatory biases or statistical biases rather than salient biases. Causal-explanatory biases would have the kind of stability we expect of beliefs, and they ought to be sensitive to the right kind of evidence (in particular, evidence that targets causal-explanatory relations). Statistical biases would also be stable in the relevant context and sensitive to the right kind of evidence (in particular, evidence that targets statistical relations). Salient biases, in contrast, are relatively fragile and not responsive to relevant evidence, and thus do not exhibit the characteristic features of belief.

To establish that many implicit biases are causal-explanatory or statistical, proponents of the belief model will need to go beyond one-off tests of spontaneous behavior, like IATs or priming tasks. These tasks are not designed do not tap into causal-explanatory relations and do not clearly test sensitivity to evidence. Given the worries about test–retest reliability for some of these tests, they may indicate a level of instability more consistent with salient biases rather than statistical biases.[21] The proponent of the belief model needs to invoke empirical evidence that individuals' implicit biases are stable when we vary irrelevant details, and they are sensitive to relevant changes in evidence. There is some indication of contextual sensitivity *suggestive* of a kind of stability in some older work on implicit bias. For example, using

[21] Though see Brownstein et al. (2019) for more on the test–retest reliability of implicit bias tests.

both IAT and sequential priming, Wittenbrink et al. (2001) find that when a Black face is superimposed on a picture of a church interior, there is a significantly weaker association between BLACK and negative features than when a Black face is superimposed on a picture of a street corner. The bias associating BLACK with negative features survives at the generic level but not for the subcategorization CHURCH-GOING BLACK MAN. More work like this may indicate a statistical bias. Similarly, Govan and Williams (2004) find that performance on a standard IAT with generic names/faces differs from performance on an IAT with well-known and liked Black individuals (e.g., Michael Jordan) and well-known and disliked White individuals (e.g., Adolf Hitler). The racial implicit bias is eliminated but not reversed in these cases. The fact that the bias was not reversed in this case suggests that participants may have a positive centrally encoded bias with respect to the racial category WHITE. That is, WHITE + GOOD may be a deep causal-explanatory bias for the participants in the study. These studies suggest a contextualized stability of some biases (feature 11), but they do nothing to establish that the biases are responsive to evidential challenge (feature 10). Thus, proponents of the belief model have more work to do here.

Critics of the belief model, in contrast, need to establish that most implicit biases are merely salient biases. These biases are unstable and break down with even minor, evidentially irrelevant shifts in context. The data on the low test–retest reliability of IAT results and low correlations amongst different types of implicit bias tests are not sufficient to establish this, however. Proponents of the belief model could easily maintain that *some* implicit biases are salient while arguing that many or most are statistical or causal-explanatory. Pointing to correlations between IAT scores and regional levels of discrimination (Payne et al., 2017), one could argue that implicit biases are more robust and predictive of behavior than we would expect if they were simply salient biases. Thus, for critics to make their case, they would have to argue that most implicit biases do not encode statistical information or causal-explanatory relations. In part, this would involve addressing the appearance of stability that proponents of the belief model highlight. But they would also need to show that implicit biases are not sensitive to relevant evidence (feature 11). This is not a straightforward task as many of the interventions on implicit bias that have been tested are short-term, one-off interventions that are not actually evidentially relevant to the content of the biases. For example, imagining counterstereotype exemplars or forming implementation intentions are not interventions that target either statistical or causal-explanatory relations that an implicit bias may encode. Indeed, the fact that most of these interventions do *not* work (Lai et al., 2014), and the fact that even the most effective ones stop working after hours or days and individuals' implicit bias scores regress back to the mean (Lai et al., 2016) is suggestive of resilience to irrelevant evidence. Thus, demonstrating that implicit biases are not responsive to relevant evidence is more challenging than typically recognized by critics of the belief model of implicit bias.

Reformulating the debate in these terms helps us see what the essence of the debate between proponents and critics of the belief model of implicit bias really is—whether no, some, or all implicit biases are stable in the relevant contexts and responsive to epistemically relevant evidence. Reformulating the debate also reveals

how much work both proponents and critics of the belief model have left to do to resolve this debate.

6 Conclusion

The philosophical debates about whether implicit biases are beliefs is at bottom a debate about whether implicit biases are stable and sensitive to evidence like we expect beliefs to be. Critics of the belief model argue that empirical evidence suggests that implicit biases are not stable or sensitive to evidence and are therefore something like aliefs, patchy endorsements, character traits, unconscious imaginings, or mere associations. Proponents of the belief model argue that implicit biases are stable enough and sensitive enough to evidence to count as beliefs, particularly if we model them as dispositional, fragmented, or Spinozistic beliefs. I argued here that both arguments are problematic because they fail to attend to the informational structure of implicit biases. Without knowing the informational structure of a representation, we cannot decide whether or not it is stable and sensitive to evidence. I presented an account of the informational structure of implicit biases that both critics and proponents of the belief model of implicit bias could use to make their case and laid out what more each side needs to do to make their case.[22]

References

Audi, R. (2001). *The architecture of reason: The structure and substance of reality*. . Oxford University Press.

Bayne, T., & Pacherie, E. (2005). In defence of the doxastic conception of delusions. *Mind and Language, 20*(2), 163–188.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review, 94*(4), 991–1013.

Bortolotti, L. (2010). *Delusions and other irrational beliefs*. . Oxford University Press.

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure? *Wiley Interdisciplinary Reviews: Cognitive Science*. https://doi.org/10.1002/wcs.1501.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*(4), 330–350.

Currie, G. (2010). Tragedy. *Analysis, 70*(4), 632–638.

Currie, G. (2000). Imagination, delusion and hallucinations. *Mind and language, 15*(1), 168–183.

Donald, D. (2001). *Inquiries into truth and interpretation: Philosophical essays*. . Oxford University Press.

[22] This paper has benefited from feedback from many people. Most significantly, I'm grateful for the hours of conversation with Guillermo Del Pinal workshopping the ideas in this paper. Without these conversations, I doubt the paper would have come to fruition. I am also indebted to Neil Van Leeuwen for helpful comments on the structure of the paper and, more generally, for always being game to read a friend's works in progress. Thanks also to audiences at University of Maryland, Iowa State University, and University of Miami. Finally, thank you to the anonymous reviewers at *Synthese* for constructive, incisive feedback on the paper.

Del Pinal, G., & Spaulding, S. (2018). Conceptual centrality and implicit bias. *Mind and Language, 33*(1), 95–111.
Doggett, T., & Egan, A. (2012). How we feel about terrible, non-existent Mafiosi. *Philosophy and Phenomenological Research, 84*(2), 277–306.
Dub, R. (2017). Delusions, acceptances, and cognitive feelings. *Philosophy and Phenomenological Research, 94*(1), 27–60.
Egan, A. (2008). Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies, 140*(1), 47–63.
Egan, A. (2009). Imagination, delusion and self-deception. In T. Bayne & J. Fernandez (Eds.), *Delusions and self-deception: Affective and motivational influences on belief formation.* (pp. 263–280). Psychology Press.
Egan, A. (2011). Comments on Gendler's, "the epistemic costs of implicit bias." *Philosophical Studies, 156*(1), 65.
Funkhouser, E. (2005). Do the self-deceived get what they want? *Pacific Philosophical Quarterly, 86*(3), 295–312.
Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44*(2), 370–377.
Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43*(3), 300–312.
Gendler, T. S. (2008a). Alief and belief. *Journal of Philosophy, 105*(10), 634–663.
Gendler, T. S. (2008b). Alief in action (and reaction). *Mind and Language, 23*(5), 552–585.
Szabó, G. T. (2007). Self-deception as pretense. *Philosophical Perspectives, 21*(1), 231–258.
Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60*(4), 509.
Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology, 40*(3), 357–365.
Graham, S., & Lowery, B. S. (2004). Priming unconscious racial stereotypes about adolescent offenders. *Law and Human Behavior, 28*(5), 483.
Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464.
Helton, Grace. forthcoming. "If you can't change what you believe, you don't believe it." *Noûs*.
Holroyd, J., & Sweetman, J. (2016). "The heterogeneity of implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy.* Oxford University Press.
Houwer De, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass, 8*(7), 342–353. https://doi.org/10.1111/spc3.12111.
Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43*(1), 17–32.
Ichino, A. (2020). Superstitious confabulations. *Topoi, 39*(1), 203–217. https://doi.org/10.1007/s11245-018-9620-y.
Ichino, A., & Räikkä, J. (2020). Non-doxastic conspiracy theories. *Argumenta*. https://doi.org/10.14275/2465-2334/20200.ich.
Isaac, C., Lee, B., & Carnes, M. (2009). Interventions that affect gender bias in hiring: A systematic review. *Academic Medicine: Journal of the Association of American Medical Colleges, 84*(10), 1440.
Ito, T. A., Thompson, E., & Cacioppo, J. T. (2004). Tracking the timecourse of social perception: The effects of racial cues on event-related brain potentials. *Personality and Social Psychology Bulletin, 30*(10), 1267–1280.
Klauer, K. C., & and Musch, J. (eds.) (2003). Affective priming: Findings and theories. In *The psychology of evaluation: Affective processes in cognition and emotion* (vol. 7, p. 49).

Kurdi, B., Morris, A., & Cushman, F. A. (2020). The role of causal structure in implicit cognition. *PsyArXiv*. https://doi.org/10.31234/osf.io/r7cfa.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., & Marshburn, C. K. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*(8), 1001–1016.

Lai, C., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., & Koleva, S. P. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765–1785.

Levy, N. (2014). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Nous, 49*, 800–823.

Levy, N. (2017). Religious beliefs are factual beliefs: Content does not correlate with context sensitivity. *Cognition, 161*, 109–116.

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: An MEG study. *Nature Neuroscience, 5*(9), 910–916.

Lynch, K. (2012). On the "tension" inherent in self-deception. *Philosophical Psychology, 25*(3), 433–450.

Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy*. Oxford University Press.

Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese, 193*(8), 2659–2684.

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

Mandelbaum, E. (2015). Attitude, inference, association: On the propositional structure of implicit bias. *Nous, 50*(3), 629–658.

Mele, A. R. (2001). *Self-deception unmasked*. . Princeton University Press.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, Jo. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474–16479.

Na'aman, Oded. (2020). Emotions and process rationality. *Australasian Journal of Philosophy*. https://doi.org/10.1080/00048402.2020.1802764.

Neil, V. L. (2014). Religious credence is not factual belief. *Cognition, 133*(3), 698–715.

Neil, V. L. (2017a). Do religious "beliefs" respond to evidence? *Philosophical Explorations, 20*(sup1), 52–72.

Neil, V. L. (2017b). Two paradigms for religious representation: The physicist and the playground (a reply to Levy). *Cognition, 164*, 206–211. https://doi.org/10.1016/j.cognition.2017.03.021.

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations*, *8*(1), 39–52.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social cognition, 19*(6), 625–666.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of personality and social psychology, 81*(2), 181.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233–248.

Pieter, V. D., De Houwer, J., & Smith, C. T. (2018). Relational information moderates approach-avoidance instruction effects on implicit evaluation. *Acta Psychologica, 184*, 137–143.

Prinz, J. (2004). Emotions embodied. In R. Solomon (Ed.), *Thinking about feeling: Contemporary philosophers on emotions*. Oxford University Press.

Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies, 175*(9), 2353–2372.

Quine, W. V. O. (1960). *Word and object*. . MIT press.

Railton, P. (2014). Reliance, trust, and belief. *Inquiry, 57*(1), 122–150.

Railton, P. (2018). Learning and doing: Toward a unified account of rationality in belief, desire and action. *The John Locke Lectures*.

Schwitzgebel, Eric. 2013. A dispositional approach to attitudes: Thinking outside of the belief box. In: N. Nottelmann (Ed.). *New Essays on Belief: Constitution, Content and Structure*. London: Palgrave Macmillan. https://doi.org/10.1057/9781137026521_5.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology, 66*, 223–247.

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22*(2), 189–228.
Smith, M. (1994). *The moral problem*. . Blackwell.
Stalnaker, R. (1984). Inquiry. *Mind, 94*(376), 627–630.
Stephens, G. L., & Graham, G. (2004). Reconceiving delusion. *International Review of Psychiatry, 16*(3), 236–241.
Sullivan-Bissett, E. (2019). Biased by our imaginings. *Mind and language, 34*(5), 627–647.
Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science, 16*(1), 56–63.
Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*(5), 815.
Wrenn, C. (2010). A puzzle about desire. *Erkenntnis, 73*(2), 185–209.