

Understanding implicit bias: A case for regulative dispositionalism

Annemarie Kalis & Harmen Ghijsen

To cite this article: Annemarie Kalis & Harmen Ghijsen (2022) Understanding implicit bias: A case for regulative dispositionalism, *Philosophical Psychology*, 35:8, 1212-1233, DOI: 10.1080/09515089.2022.2046261

To link to this article: <https://doi.org/10.1080/09515089.2022.2046261>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 02 Mar 2022.



Submit your article to this journal [↗](#)



Article views: 1395



View related articles [↗](#)





View Crossmark data [↗](#)

ARTICLE

 OPEN ACCESS

 Check for updates

Understanding implicit bias: A case for regulative dispositionalism

Annemarie Kalis ^a and Harmen Ghijsen ^b

^aDepartment of Philosophy and Religious Studies, Utrecht University, Utrecht, The Netherlands;

^bFaculty of Philosophy, Theology and Religious Studies, Radboud University Nijmegen, Nijmegen, The Netherlands

ABSTRACT

What attitude does someone manifesting implicit bias *really* have? According to the default representationalist picture, implicit bias involves having conflicting attitudes (explicit versus implicit) with respect to the topic at hand. In opposition to this orthodoxy, dispositionalists argue that attitudes should be understood as higher-level dispositional features of the person as a whole. Following this metaphysical view, the discordance characteristic of implicit bias shows that someone's attitude regarding the topic at hand is not-fully-manifested or 'in-between'. However, so far few representationalists have been convinced by dispositionalist arguments, largely because dispositionalism cannot provide explanations in terms of underlying processes. We argue that if dispositionalism wants to be a genuine contender, it should make clear what it has to offer in terms of understanding of implicit bias. As a concrete proposal, we combine dispositionalist metaphysics with the idea that our normative practices of attitude ascription partly determine what it means to have an attitude. We show that such *regulative dispositionalism* can account for two prominent normative features of implicit bias. We conclude by suggesting that in order to engage in a meaningful debate with representationalism, dispositionalists might have to put the question 'what counts as a good explanation?' back on the table.

ARTICLE HISTORY

Received 11 November 2019
Accepted 18 February 2022

KEYWORDS

Implicit bias;
dispositionalism;
representationalism; implicit
attitudes; belief;
mindshaping

Introduction

Suppose you sincerely claim to believe that women are just as capable as men in any field of science, and yet after taking an *Implicit Association Test* (IAT, Greenwald et al., 1998), you turn out to be consistently slower and more error-prone when grouping together stimuli related to *female* and *science* than when you're grouping together stimuli related to *male* and *science*. In other words, you seem to suffer from implicit bias. The default understanding of the discordance characteristic of implicit bias is that

CONTACT Annemarie Kalis  A.Kalis@uu.nl  Department of Philosophy and Religious Studies, Utrecht University, Janskerkhof 13, Utrecht 3512BL, The Netherlands

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

subjects have conflicting attitudes with respect to the topic at hand.¹ While one such state, the *explicit* attitude, is causally responsible for what someone says they believe, another state (or set of states) forms the *implicit* attitude which is causally responsible for producing the kind of automatic behavior measured by tests like the IAT. Even though there are many different accounts of the nature of these conflicting attitudes, the default position is that one must appeal to conflicting attitudes to provide a satisfactory account of implicit bias (e.g., Gendler, 2008a, 2008b, 2011; Holroyd, 2016; Levy, 2015; Madva, 2016; Madva & Brownstein, 2018; Mandelbaum, 2016).

During the last two decades dispositionalist approaches have challenged this way of thinking about implicit bias, by rejecting its underlying metaphysical understanding of what attitudes are. Rather than viewing attitudes as relations to internal representations, as is the representationalist orthodoxy (and more or less the default in philosophy of cognitive science), dispositionalists argue that attitudes, such as how one evaluates gender equality, are higher-level dispositional states of *the entire person*. From this alternative metaphysical understanding, a different account of implicit bias follows. When attitudes are seen as states of the person as a whole, it no longer makes sense to explain implicit bias in terms of contradictory attitudes: that would require an individual to be in two contradictory states at the same time. Instead, dispositionalists explain the discordance typical of implicit bias as different strands of a person's cognition and behaviour pointing into conflicting directions, suggesting that the person's attitude toward the topic is not-fully-manifested or 'in-between' (Schwitzgebel, 2001, 2010).

The aim of this paper is to show what is needed in order to develop dispositionalist accounts of implicit bias into a genuine alternative to the default representationalist understanding. We believe this is necessary because so far, dispositionalist accounts of implicit bias have been quite unsuccessful in convincing representationalists (Brownstein, 2019; Holroyd, 2016; Johnson, 2020; Quilty-Dunn & Mandelbaum, 2018a). One important advantage of representationalist accounts is that they can easily connect their analysis of implicit bias as discordance between explicit and implicit attitudes to a large body of empirical work on underlying processes and mechanisms (Porot & Mandelbaum, 2021; Quilty-Dunn & Mandelbaum, 2018a). In contrast, dispositionalism deliberately presents itself as a 'superficial' account in the sense that it does not commit to specific claims on how dispositional patterns are realized at the level of cognitive processes and mechanisms (Schwitzgebel, 2013). According to representationalists, this makes the account superficial also in the more pejorative meaning of the term: it seems to show that dispositionalism just lacks the kind of explanatory power that representationalism brings to the table (Greely, 2014; Quilty-Dunn & Mandelbaum, 2018a).

In order for dispositionalism to be a genuine contender in the debate, it thus needs to provide a better story of the kind of understanding of attitudes and implicit bias it has to contribute. More specifically, it needs to show that its metaphysical approach comes with explanatory tools that are not part of the representationalist toolbox. Our proposal is that by combining a dispositionalist metaphysics with a regulative approach to attitude ascription, dispositionalists can show that certain normative aspects of our folk psychological practices partly determine what it means to have an attitude. We will argue that such a *regulative dispositionalism* has unique resources to explain two important normative features of implicit bias, which also have wider ramifications for our understanding of attitudes in general. Firstly, regulative dispositionalism can explain why the discordance characteristic of implicit bias bothers us in a specific way, and secondly, it can show how first-person attitude statements can be a tool for change.

The paper will be structured as follows. In section two we provide an overview of those existing forms of dispositionalism that feature most prominently in the debate, and show how they fall short in providing a convincing alternative to representationalism. We continue by developing our proposal for such an alternative (regulative dispositionalism) in section three. In section four we argue that regulative dispositionalism can account for two important normative features of implicit bias that cannot easily be grasped from a representationalist perspective. In the conclusion (section five) we discuss some possible representationalist objections, and argue that dispositionalists should reopen the debate with representationalism by putting the question ‘what counts as a good explanation?’ back on the table.

Dispositionalist approaches to implicit bias

As said, dispositionalist accounts of implicit bias provide a specific kind of metaphysical understanding of attitudes. Whereas representationalists claim that attitudes should be understood as relations to representations, dispositionalists claim that attitudes are multi-track dispositions or, in another version of the same idea, a pattern of different dispositions that we ascribe to individuals (Baker, 1993; Machery, 2016; Ryle, 1949; Schwitzgebel, 2002). Attitudes thus manifest in a multitude of ways: the attitude someone has shows in what that person thinks, feels, says and does, both consciously and non-consciously. Because attitudes are dispositional patterns, dispositionalists reject the idea that attitudes could be either implicit or explicit: the predicates ‘implicit’ and ‘explicit’ just cannot be meaningfully applied to states of an individual (Machery, 2016).

Not all forms of dispositionalism about implicit bias present themselves in opposition to representationalism. For example, Welpinghus (2020) defends a view according to which implicit bias should be understood as

the disposition to evaluate members of a different social group less (or more) favorably, without intending to do so. However, she does not defend the more general view that attitudes are dispositions, nor does she develop her position as a contender to representationalism. Another example of such a position is found in Johnson (2020): although she raises objections against traditional representationalist accounts (most notably that they cannot account for *truly* implicit bias), her functionalist account still grounds the dispositions relevant for implicit bias in specific types of representational states. Here we will focus on accounts that propose a dispositionalist understanding of implicit bias in full opposition to the idea that attitudes are relations to representations. The two prominent accounts defending such a view are Machery's trait view (Machery, 2016) and Schwitzgebel's liberal dispositionalism (Schwitzgebel, 2001, 2002, 2010, 2013). We will discuss them in turn, and show how they fail to convince as genuine alternatives to the representationalist approach.

Machery's trait view

Machery (2016) argues that attitudes are traits, and thus that we should understand them in the same way in which we understand character traits like courage. Just as we do not take traits to be reducible to a specific underlying representational state, we should not expect an attitude to be reducible in this way either. Machery takes attitudes to be dispositions to respond to stimuli in the environment in a certain way, dispositions that are brought about by various underlying states and processes. For instance, a negative attitude toward a certain social group is related to the set of one's moral beliefs, emotions, non-propositional associations between concepts, etc., which together determine whether one will be disposed to, say, see members of this group in a bad light.² On Machery's view, this set of underlying states and processes forms the *psychological basis* of the disposition.

Although Machery does not explicitly argue against representationalism, he does reject a core assumption of the representationalist framework: the idea that attitudes are mental states. Thus, by extrapolation his view necessarily also goes against the idea that attitudes would be *representational* mental states. By arguing that attitudes are not mental states, Machery (2016) reaches the conclusion that attitudes cannot be either implicit or explicit. Whereas mental states and processes might be labeled as implicit or explicit (depending for instance, on whether they're introspectable or conscious), attitudes describe broad behavioral tendencies, and thus this categorization simply does not apply to them.

This also leads Machery to a different account of what is expressed by explicit avowals, such as someone saying "Yes, I really believe the genders are equal!". Usually, such avowals are taken as expressions of an explicit attitude,

which might conflict with one's implicit attitude. But Machery rejects the idea that attitudes can be either explicit or implicit; his alternative suggestion is that explicit avowals can be understood in various different ways. They can be directives (ordering oneself to believe in gender equality), expressives (expressing one's positive feelings about gender equality), commitments (committing oneself to the norm of gender equality), or subjective reports of what one takes one's attitude to be (Machery, 2016, p. 114). With regard to the latter interpretation, Machery stresses that subjective reports of one's own attitudes will often be mistaken. Being dispositions, attitudes are not themselves subject to introspection. People's subjective reports of their attitudes are (mostly) formed by considering those components of the disposition's psychological basis that *are* introspectable³: their picture is thus necessarily incomplete. In cases of implicit bias, what might look like a conflict between a so-called 'explicit attitude' and 'implicit attitude' actually just reflects the fact that one's avowals only tell part of the story about what one's attitude really is: what is often overlooked are precisely those behaviors, thoughts and/or feelings that are outside the agent's awareness and control.

Machery's main argument for accepting this dispositional trait view of attitudes comes from psychological research on implicit bias. By now it is well known that people's scores on different psychological measures of implicit bias vary over time and between contexts, and are only modestly correlated with one another (see, e.g., Gawronski, 2019; Jost, 2019; Machery, 2021 for discussion). Machery argues that this lack of coherence in implicit measures undermines the idea that the construct measured by these psychological measures is one unitary underlying mental state. In contrast, he argues, by taking an attitude to be more like a trait, low correlation between different measures can be explained by the idea that each measure taps into a different element of the psychological basis that determines the attitude trait.⁴

Even though we share Machery's conclusion that the fragmentation shown by implicit measures raises serious doubt about the validity of a unitary understanding of 'implicit attitude', we are not convinced that a trait concept does much better here. After all, wouldn't one expect a trait like courage to be determined by a set of underlying states and processes that *all point in a similar direction*? If not, it seems difficult to see why these underlying states and processes together should be interpreted as determining one and the same trait. For example: if someone has a strong fear response to darkness, snakes, spiders and a whole list of other commonly feared entities, yet at the same time has such a strong sense of moral duty that they would run into a burning building to save a stranger, then why assume that this diverse collection of psychological factors belongs to the psychological basis of just one trait, namely, courage? Just as with attitudes, in ascribing a trait we expect sufficient coherence in one's cognitive and behavioral responses, and refrain from ascribing the trait if such coherence is absent.⁵

Secondly, we are not convinced Machery's trait view should be seen as a genuine alternative to representationalism: many representationalists clearly acknowledge that we shouldn't think of implicit attitudes as a unitary type of mental state. For instance, Holroyd and Sweetman (2016) recommend to be cautious with generalizations about implicit bias, because there is "functional heterogeneity in the way that different implicit associations operate" and because "there may be heterogeneity in the processes underpinning different implicit associations" (p. 88). Similarly, Byrd (2021) argues that, although strong debiasing experiments point in the direction of an associationist view of implicit bias, there is still room for an interactionist view according to which implicit bias can be related to *both* associative and non-associative processes – again making 'implicit attitude' into a heterogeneous kind. Such representationalists are perfectly happy to accept that implicit attitudes are not a unitary kind of mental state, even though they are still firmly wedded to the idea that we should understand them within a (broader) representationalist framework. This more nuanced understanding of implicit attitudes seems perfectly compatible with Machery's (2016) trait view, especially because Machery explicitly acknowledges (Machery, 2016, p. 107) that the trait view is not meant to exclude such a position. The main difference seems to be that Machery uses the label 'attitude' to refer to a disposition which is determined by a motley of relevant representational states and processes, whereas on the above nuanced representationalist views, each of the underlying representational states is labeled as a separate implicit attitude.

Machery's dispositionalism thus cannot really be seen as a fully-fledged alternative to representationalist ways of thinking. His argument against standard representationalist views backfires by also being applicable to his own account, and a nuanced version of his position seems to be quite compatible with nuanced forms of representationalism. It thus seems that Machery's view, like those of Johnson (2020) and Welpinghus (2020), is not the best candidate for showing how dispositionalism could be a genuine contender to representationalist accounts of implicit bias.

Schwitzgebel's liberal dispositionalism

In various papers, Eric Schwitzgebel has developed a more radical anti-representationalist account of implicit bias, which he grounds in his dispositionalist metaphysics of attitudes.⁶ According to his metaphysical position, attitudes are constituted by a cluster of dispositions which include not just behavioral, but also cognitive and phenomenal dispositions (Schwitzgebel, 2002). He labels his form of dispositionalism 'liberal', to emphasize that it allows for 'inner' manifestations of attitude dispositions, thus contrasting it to more traditional forms of dispositionalism that faced charges of behaviorism.

To give an example, holding the belief that the genders are equal involves, amongst other things, being disposed to say certain things (such as affirming that the genders are equal when asked), being disposed to respond to certain situations in certain ways (such as not eliminating potential job candidates based on their gender), and being disposed to feel certain things (such as feeling outrage when a woman is belittled). Which precise dispositions ‘belong’ to a certain attitude is difficult to specify, but according to Schwitzgebel we have a fuzzy set of folk psychological expectations of someone who believes in gender equality. This is referred to as the *dispositional stereotype* for that specific attitude; we have such stereotypical expectations for many common attitudes.

However, these expectations are sometimes violated. Some of our attitude patterns only partially manifest, resulting in in-between or fragmented attitudes (Schwitzgebel, 2001, 2010). On Schwitzgebel’s account, this shows that having an attitude is not a dichotomous matter: there is not always a clear answer to the question whether or not a person believes or desires something. Implicit bias is a core example of such fragmented believing. Although we explicitly profess to believe in the equality of the genders, thereby manifesting one of the behavioral dispositions for the belief that the genders are equal, we also display behavior that does not fit so well with that ascription of the belief (e.g., we are faster to associate *male* rather than *female* with *science*). This also provides an explanation of the observed low correlations between different implicit measures: following Schwitzgebel’s account, such low coherence just indicates the fragmentation of such attitude patterns.

For Schwitzgebel, the existence of in-between attitudes is an important argument against representationalist theories. Whereas representationalists are committed to there being a yes or no answer to the question whether someone has a certain attitude (given that this is determined by the presence of a specific kind of mental state), liberal dispositionalism has conceptual room for fuzzy cases. After all, on their view the question whether or not someone has a certain attitude becomes the question whether the person displays the relevant stereotypical pattern *to a sufficient extent*. However, this argument of in-between believing has not convinced most representationalists. As outlined in the previous section, by now there are several nuanced views on the table that allow for attitudes to be realized by a variety of representational states and processes, which can point in different directions. For example, Quilty-Dunn and Mandelbaum (2018a) claim Schwitzgebel’s (2013) description of representationalism as a ‘belief-box’ account is misleading: they argue that representationalism actually does not commit to an inflexible notion of belief storage. On their view, all representationalism commits to is the claim “that different mental states can share contents because they incorporate the same representations” (Quilty-Dunn and Mandelbaum, p. 2357). So, someone’s belief that tigers have stripes, and that

person's belief that tigers are dangerous, share certain constituents (namely: the concept TIGER). And importantly: "since constituents are repeatable in different contexts, the representational theory of mind can explain how we can freely recombine concepts in systematic and productive ways" (p. 2357). This architectural flexibility provides a different explanation of what Schwitzgebel calls 'fuzzy cases': various implicit measures pick out different (only partially overlapping) implicit attitudes, which incorporate the same representations only in so far as they share the same content. True, this account does not allow for *real* metaphysical fuzziness about attitudes: each separate implicit attitude is still either present or absent. However, representationalists do not seem convinced that allowing for such genuine metaphysical fuzziness would provide explanatory advantage. After all, don't we want an account of belief precisely to answer questions on when someone can and cannot be said to have a belief? Greely (2014) even argues that Schwitzgebel's position is not a genuine metaphysical account, precisely because it leaves the metaphysical question open for all difficult cases.

So what are the advantages of defending a liberal dispositionalist account over a representationalist one? A core advantage Schwitzgebel (2013) brings forward is that his liberal dispositionalism is a deliberately *superficial* account. Dispositionalist criteria for ascribing attitudes do not need to presuppose the existence of specific underlying mechanisms or processes. So in order to determine whether or not someone has a certain attitude, we don't need to 'look inside' to investigate someone's cognitive apparatus: all we need to do is answer the question whether someone manifests the relevant dispositional pattern to a sufficient extent.

However, it is not clear that liberal dispositionalism can actually answer this question, with in-between attitudes being so central to the account. Moreover, Quilty-Dunn and Mandelbaum (2018a) convincingly argue that the feature of superficiality makes dispositionalism also superficial in a second, potentially more harmful sense: dispositionalism has relatively little explanatory or predictive value in making sense of the cognitive science of belief.⁷ The cognitive science of belief addresses important explanatory questions: for instance, why does having a certain belief lead you to selectively avoid counterevidence to it, and why do people suffer from fragmented forms of believing such as implicit bias? Liberal dispositionalism has no answer to such questions.

Schwitzgebel's response has been that this is not what dispositionalism is supposed to contribute. What it *does* contribute is a metaphysical account that has pragmatic value, in that it "directs our attention to what we ought to care about most in thinking about belief" (Schwitzgebel, 2021, p. 351). As he argues, many "intellectualist" approaches attach too much weight to explicit attitude statements (such as "I strongly believe the genders are equal!") in determining what our attitudes are. In contrast, for a dispositionalist, 'what

one says' is just one of the nodes in our dispositional profile. This shows that we shouldn't take our own attitude statements too seriously: in order to genuinely count as having a certain attitude, we should 'not only talk the talk, but also walk the walk'. This can make us take the phenomenon of implicit bias more seriously, and helps to avoid taking a noxiously comfortable view of ourselves (Schwitzgebel, 2013, 2021). However, many representationalists will completely agree that explicit attitude claims should not be taken at face value: precisely because they identify attitudes with internal representational states, their approach has ample room for the fact that we often seem to give mistaken reports about our own attitudes. So this pragmatic argument will also not be of help for developing a dispositionalist alternative to representationalist approaches.⁸

To conclude, we have argued that whereas Machery's account is not in clear opposition to nuanced versions of representationalism, Schwitzgebel's more radical account fails to show how liberal dispositionalism has clear advantages over representationalism: specifically, it fails to show how it can understand features of implicit bias that representationalism cannot. In the next sections we want to take up this challenge. As we will show, a dispositionalist metaphysics of attitudes suggests that our folk psychological practices of attitude ascription might partly determine what it means to have an attitude. We will embrace this suggestion and show how it leads to a position which we label *regulative dispositionalism*. Crucially, we will argue that such regulative dispositionalism comes with explanatory tools that do not belong to the representationalist toolbox.

Regulative dispositionalism

In an early paper Schwitzgebel remarks that the "stereotypes [for belief and other folk psychological categories] capture more than mere statistical regularities [...] They capture something about how we think people *ought* to think, feel, and behave" (Schwitzgebel, 2002, p. 262).⁹ Tumulty (2011, 2014) also notes that Schwitzgebel's notion of dispositional stereotypes relies on the presence of a folk psychological practice somehow 'prescribing' these stereotypes. However, neither Schwitzgebel nor Tumulty have further developed the idea that dispositional stereotypes originate in our normative folk psychological practices, nor applied this idea to questions concerning the understanding of implicit bias. That is exactly what we will do in this section, by starting from recent work on the idea that folk psychology is a regulative practice.

For a long time, it was taken for granted that the practice of ascribing folk psychological states such as beliefs and desires was geared toward *description*: the idea was that in ascribing attitudes, we aim to report on the presence of psychological states in the other person (or, in case of self-ascription, in

oneself). This was assumed by views that had otherwise completely opposite views on how to best understand this descriptive practice, with theory-theory and simulation theory being the main contrast (Churchland, 1981; Davies & Stone, 1995; Gordon, 1986; Stich & Ravenscroft, 1994).

Recent critiques of the view of folk psychology as a descriptive practice have led to the idea that folk psychology is primarily a practice that *regulates* and *shapes* our minds and behavior (De Bruin, 2017; McGeer, 2007, 2015, 2021; Vierkant & Paraskevaides, 2012; Zawidzki, 2008, 2013). The main idea of these so-called mindshaping approaches is that in ascribing attitudes to each other, we impose normative expectations; we appeal to the agents in question to also manifest other, not (yet) observed features that belong to the ascribed attitude. Such regulative practices make us more predictable and readable to each other, thereby enabling successful coordination and collaboration.

Regulation is not the only role of folk psychological ascription: we also use such ascriptions to teach others what it means to have attitudes, and what the criteria for legitimate ascription are (which can be as mundane as telling your children that if they say they *want* a peanut butter sandwich, then we expect them to actually eat it). In addition, the success of regulative practices that shape our minds and behaviors also implies that attitude ascriptions will often be pretty helpful as *descriptions* of our psychology (for an overview of the three roles of regulation, pedagogy and description, see, McGeer, 2021). But the crucial idea of mindshaping approaches is that our folk psychological practice does not describe a fully preexisting psychological reality: that reality is, at least in part, created by the mindshaping effects of our folk psychological practices.

So when we ascribe attitudes, what kind of normative expectations are at stake? McGeer (2007) argues that some of these directly follow from norms of basic rationality: in ascribing to someone the belief that the earth is round we expect the person not to say things that blatantly contradict that idea, such as saying “the earth is flat”. However, McGeer also emphasizes that many of the norm-governed expectations are not grounded in norms of rationality. The norms governing our social interactions are wildly varied, often specific to particular cultures or groups within cultures, and most of the time only vaguely articulated (McGeer, 2021, pp. 1050-1).¹⁰ The examples McGeer mentions in this context have to do with other folk psychological practices than just ascribing attitudes, such as greeting politely or expressing anger in an appropriate way, but it’s not difficult to see that these practices have important connections. For instance, ascribing the belief to someone that they were insulted, will lead to different behavioral expectations depending on the prevailing norms for expressing anger (also, you would have to know a great deal about what could be insulting in this context to ascribe the belief in the first place).

Despite the impossibility of providing an exhausting list of normative expectations inherent in folk psychological ascriptions, McGeer does hold that they will be built around a set of general constraints that are written into our nature as rational beings (McGeer, 2021). Among these are *evidential constraints* that govern what is appropriate to believe, *evaluative constraints* that govern what is appropriate to desire, and *executive constraints* that govern what is appropriate to prefer and do in light of one's beliefs and desires. This bedrock to our folk psychological practice also underscores McGeer's realism with regard to our ascriptions: "3rd person attributions are, typically, not just a facon-de-parler. Typically, we do not treat others *as if* they had the requisite states; we assume our psychological attributions actually *describe* how things are with them, psychologically speaking" (McGeer, 2021, pp. 1054-1055).¹¹ However, McGeer does not elaborate on the details: as far as we know neither she or any of the other mindshaping theorists has so far provided an account explaining under what conditions we can legitimately say that an agent has a certain attitude.

Our core suggestion is that mindshaping accounts like McGeer's can be seen as a natural partner for a dispositionalist metaphysics of attitudes. Combining these perspectives leads to a position we propose to call *regulative dispositionalism*: a position which claims that because attitude ascription is a practice that regulates and shapes our minds according to norms, the normative aspects of these attitude practices form an intrinsic part of the answer to the question what attitudes are. To answer the question whether someone with implicit bias does or does not really believe that the genders are equal, we should investigate whether or not the relevant dispositional pattern is sufficiently manifested (this is the dispositionalist view). But in order to answer *that* question, we should examine how much coherence (and what kind) we expect from people who explicitly commit to gender equality. In other words: according to regulative dispositionalism, we can only answer metaphysical questions about attitudes by looking into the normative criteria for attitude ascription.

Note that, like Schwitzgebel's view, regulative dispositionalism acknowledges that there are cases in which there is no determinate answer to the question whether or not someone has a certain attitude. However, we believe regulative dispositionalism has an important advantage over Schwitzgebel's view. By combining a dispositionalist metaphysics with a mindshaping approach to folk psychology, the position becomes significantly stronger as an alternative to representationalist views. Firstly, it is a genuine alternative: it clashes with representationalist metaphysics. After all, according to regulative dispositionalism the content of someone's attitude and whether or not someone has that attitude, are determined by normative criteria that are defined and upheld within a folk psychological community. On the other hand, for a representationalist, the contents and

presence of attitudes are fully determined by the presence or absence of the relevant representational states. From a representationalist point of view, our folk psychological practices cannot have anything to bear on the metaphysics of attitudes, except as potential causal factors in bringing about the relevant states. This means that regulative dispositionalism and representationalism give mutually exclusive answers to the question what are the criteria for having attitudes.¹²

Secondly, next to it thus being a genuine contender, we also believe that regulative dispositionalism has advantages over representationalist accounts. As we will explain in the next section, by emphasizing that normativity is part of the metaphysics of attitudes, regulative dispositionalism becomes able to account for two important normative features of implicit bias. Firstly, it can show why the discordance characteristic of implicit bias is an uncomfortable resting point for agents. And secondly, it can show how it is possible that explicit attitude statements can be tools for change.

Two normative features of implicit bias

We started the paper with the observation that implicit bias is often characterized by *discordance*: whereas some of the things someone does or says suggest that he or she believes that the genders are equal, other behaviors, responses or feelings of that person suggest otherwise. Representationalism has a clear explanation of such discordance: “beliefs are representational states that are literally stored in the mind, just as episodic and semantic memories are. The idea that our beliefs are fragmented can therefore be explained by positing architectural divisions between belief stores. It is thus because two inconsistent sets of beliefs are stored separately that they persist despite inconsistency, and that they are accessed at different times to produce different behaviors” (Quilty-Dunn & Mandelbaum, 2018a, p. - 2358).¹³ Dispositionalists also provide an explanation: they understand discordance as having an attitude which is incompletely manifested or in-between. However, this metaphysical debate has so far failed to address an important aspect of discordance: the fact that the discordance characteristic of implicit bias is an uncomfortable resting point for agents. Indeed, if fragmented attitudes are the rule rather than the exception (as most participants in the debate seem to argue), then why do we feel that something is painfully amiss if our verbal reports and behaviors don’t line up? Finding out that one suffers from implicit bias is worrying in a specific way: one realizes that one responds, feels or makes decisions in ways that “conflict with our professed beliefs and values” (Holroyd et al., 2017). But why would this be worrying, to the extent that it often comes with a felt pressure to change? Seen from a representationalist point of view, such discomfort

cannot arise from the mere fact of attitude fragmentation itself. The cognitive system does not contain rules that ‘check’ for consistency between different attitude fragments: indeed, the fact that there are no such checks precisely explains how fragmentation is possible (Porot & Mandelbaum, 2021; Quilty-Dunn & Mandelbaum, 2018a). So in order to explain the normative pressure agents experience, representationalism needs to bring in ‘external’ normative structures and motivational processes that explain why inconsistencies bother us.

Quilty-Dunn and Mandelbaum (2018a) propose that dissonance theory is a good candidate for such a normative structure, and this indeed seems plausible for various forms of fragmented believing such as induced compliance. In studies investigating this phenomenon, subjects are manipulated into doing things that go against their standing attitudes. In order to reduce dissonance, participants in such studies often respond by changing their own narrative about what they believe, without being aware of doing so. However, precisely for implicit bias, this cannot be the full story: dissonance theory claims that people resolve dissonance by choosing the path of least resistance, in other words that they reduce inconsistencies via the easiest available route (Harmon-Jones & Harmon-Jones, 2007; Quilty-Dunn & Mandelbaum, 2018a). However, people who find out that they manifest implicit bias, generally draw the conclusion that they should bring their responses, feelings and so on in line with their explicit attitude statements – and certainly not the other way around, even if this would be the path of least resistance. But why would this be so?

This is where regulative dispositionalism has a story to offer. According to this perspective, implicit bias is a prime example of agents partially failing to meet the normative expectations that guide attitude ascription. Although someone meets some of the norms held in our community for believing in, for example, gender equality (such as how we expect people with this belief to *talk* about gender equality), he or she violates certain other relevant norms for the same belief (such as the norm not to favor men in hiring decisions). The relevant norms are, as explained above, of various kinds. They involve basic inference rules but also norms regarding the relative ‘weight’ of the various elements of an attitude. Crucially, they also involve norms prescribing which attitudes a society considers morally preferable. Taken together, all these constraints form a normative framework regulating what counts as having a ‘fully-fledged’ attitude, what counts as coherence or the lack of it, and how discordance should be resolved. The lack of coherence displayed by people suffering from implicit bias, makes them candidates for being regulated or corrected in specific directions (McGeer, 2021, p. 1054). This is what we do when we say to someone: “you say you believe in gender equality, but if so why do you keep hiring only men?” Importantly, social regulation is thus not geared toward coherence *per se*.

What we expect from each other is not that agents resolve inconsistencies in whatever way (by choosing the path of least resistance), but that they resolve them in a way that leads them to have attitudes that withstand social and moral scrutiny.¹⁴

Moreover, the idea that in ascribing attitudes we impose normative expectations, also applies to self-ascription. In saying things like “yes, I believe in gender equality!” we impose normative expectations on ourselves; we appeal to ourselves to actually meet the normative expectations that go with this self-ascription. As McGeer (2007) argues: when agents “have publicly attributed a belief to themselves, they feel some pressure not to let their companions down in the expectations those companions now form about what they will say or do, and they find themselves responding to that pressure by monitoring what they say or do a little more carefully” (p. 146). This sheds interesting light on why the discordance or in-between believing manifested in implicit bias is such an uncomfortable resting position for agents: it is precisely when they explicitly express their attitudes that agents open themselves up for regulation and correction.

Representationalists might argue that this normative background story is precisely that: a mere background story, which need not be part of our metaphysical or psychological account of attitudes. Instead, they prefer to separate normative questions from metaphysical and psychological ones (for a similar argument regarding human thinking, see, Elqayam & Evans, 2011). Our response to such an objection would be that this is precisely where regulative dispositionalism is in genuine and meaningful disagreement with the representationalist point of view. Whereas it is obvious that normative standards in themselves cannot explain why human attitudes manifest the way they do (given that deviation from such standards is so common), regulative dispositionalism argues that such normative standards are an intrinsic part of the explanation, needed in order to understand why the discordance characteristic of implicit bias bothers us in the way it does.

The second normative feature regulative dispositionalism can shed light on, directly follows from this analysis of discordance. As said, agents suffering from implicit bias experience normative pressure to resolve the discordance in a specific way: namely by bringing their behavior, feelings etcetera in line with their first-person attitude statements. However, how is it even possible to do something like this? We want to propose that regulative dispositionalism can clarify how it is possible that agents leverage their own statements regarding what their attitudes are, for developing more coherent and justifiable attitudes. As said, McGeer (2021) distinguishes three different roles for attitude ascription: regulation, pedagogy and description. She points out that these roles are differently distributed across first-personal, second personal and third-personal ascriptions of attitudes. Whereas we use third-personal ascriptions (‘John believes that women are

equally good at chess as men') most often for descriptive purposes, second-person ascriptions usually have either a pedagogical or regulative role. First-person ascriptions sometimes have a descriptive function (when we want to give others information about ourselves) but according to McGeer, their dominant role is regulative: "by attributing mental states to ourselves in an engaged 1st-personal mode, we give others ample opportunity to call us to account when our actions do not mesh with the claims we make about how we are minded" (McGeer, 2021, p. 1054). Thus, first-personal expressions of attitudes play the role of commitments (McGeer, 2015).

McGeer here builds on work by Richard Moran on the relation between self-knowledge and avowal (Moran, 2001). Moran developed the influential position that statements such as 'I believe women are just as capable as men' should not be understood as reports of one's own belief state, but as avowals that are backed by reasons. Such statements are transparent in the sense that the evidence for their truth is not found 'in our heads': we back up our belief statements not by pointing inwards (I really believe so!) but by directly pointing at the reasons for considering women as being equally capable as men. From this it follows that in asking ourselves what we feel, think, believe etcetera, we are not asking ourselves to report on our inner states, but to actively make up our minds. In other words, figuring out what one thinks is a matter of figuring out *what to think*.

Now compared to mindshaping accounts, Moran's understanding (having a strong Kantian foundation) of the role of normative constraints is very narrow: for him, the norms on the basis of which we figure out what to think, are basically just the norms of rationality. Moreover, in the same Kantian spirit Moran argues that the first-personal stance is the *only* stance from which agents can authentically shape their own minds: observing one's own mind from an outside perspective is, for Moran, always a form of alienation (Moran, 2001). As a critical response to this narrow analysis, recent mindshaping accounts have emphasized that self-regulation should be seen as a dynamic interplay between making first-personal commitments and observation of one's own thoughts, emotions and behavior (De Bruin, 2017; McGeer, 2015).

Combining these ideas on the role of first-personal mindshaping with a dispositional metaphysics suggests that explicit attitude statements such as 'I believe that the genders are equal' should be seen as avowals by which agents, in making themselves susceptible to normative regulation by others and by oneself, can work toward the development of fully-fledged attitudes in the dispositionalist sense. Thus, such avowals can help agents to 'not only talk the talk, but also walk the walk'. This picture stands in sharp contrast to that sketched by representationalists, and is also different (although less radically so) from the kinds of dispositionalism defended by Machery and Schwitzgebel. All of them tend to stress the descriptive understanding of first-personal attitude statements as self-reports of our

attitudes.¹⁵ On their own, such self-reports do not count as genuine manifestations of the relevant attitudes. And given that our capacities for self-knowledge are pretty limited, this leads them to the conclusion that ascribing attitudes to oneself amounts to just ‘saying things’: self-reports will, most of the time, not stand up to scrutiny. We believe that regulative dispositionalism has unique tools for showing in what sense explicit attitude claims are *more* than just reports on one’s internal state, by providing an account of the power first-person expressions can have. For many people, taking an explicit stance on where they stand with regard to important issues is a crucial step toward living out those statements in their everyday lives. True, this is not easy—and having ideally coherent attitudes might be just that: an ideal which human beings might strive for but will never completely reach. But nevertheless, we should not underestimate the role of ‘what we say we believe’ in our normative practices. By expressing such a commitment, agents open themselves up to regulation and correction by others and themselves, and thus open up a path for change.

By emphasizing the power of explicit avowals, our regulative dispositionalism is probably closer to intellectualism than someone like Schwitzgebel might like. However, in our view precisely this makes our dispositionalist position stronger vis-à-vis representationalism. Representationalists cannot account for the idea that avowals could have power as tools for change, and might not want to. As with the issue of discordance, they might prefer to explain any effectiveness of avowals externally by reconceptualizing such ‘power’ as a causal force brought about by social pressures. Our ambition in this paper is not to refute that way of thinking, but to clarify that it is here that we find the core point of disagreement between representationalism and regulative dispositionalism. Whereas representationalism proposes a metaphysics of attitudes that aims to be purely descriptive and to a substantial extent normatively neutral,¹⁶ regulative dispositionalism argues that certain normative features of attitudes and implicit bias can be understood much better by acknowledging that our normative practices of attitude ascription are part and parcel of the metaphysics of attitudes themselves.

Conclusions

In this paper we have proposed to reframe the opposition between representationalism and dispositionalism in terms of the following question: is it possible to understand what it means to have an attitude, or to suffer from implicit bias, without taking the normative features of practices of attitude ascription into account? Representationalism argues *yes*, while regulative dispositionalism answers *no*. We believe our position fares better than

existing dispositionalist accounts on two points. Firstly, by claiming that the normative features of attitude ascription should be part of the metaphysical account, regulative dispositionalism offers a genuine alternative to representationalism. Secondly, regulative dispositionalism has a clear answer to the question what it contributes to our understanding of the phenomena at stake: by integrating normativity in the metaphysics, it becomes possible to explain certain normative features of attitudes in general, and implicit bias in particular. For these reasons, we believe regulative dispositionalism offers a better starting point for engaging in philosophical debate with representationalism.

However, even if representationalists were to agree concerning the contributions just mentioned, these clearly do not provide a decisive argument against the representationalist understanding of implicit bias, or attitudes in general. Representationalists might object that their lack of tools for explaining normative features is a small price to pay, compared to the advantage of having conceptual tools that connect so easily with the cognitive science of attitudes. Moreover, they might consider it perfectly satisfactory to refer to 'external' causal social mechanisms for explaining normative features like the ones we have been discussing. This raises the question: doesn't the debate ultimately come down to a fundamental disagreement regarding the question what a metaphysical account of attitudes and implicit bias is supposed to do?

We think this might indeed be the case. However, this does not close off meaningful philosophical debate. To the contrary, we hold that our understanding of both attitudes in general, and implicit bias in particular, could benefit from addressing these fundamental disagreements directly, by putting the question 'what counts as a good explanation?' back on the table. Whereas representationalism has teamed up with cognitive science in providing explanations in terms of underlying processes and mechanisms, regulative dispositionalism can be seen as offering a different form of explanation: it provides insight in how our social-cultural settings and folk psychological practices shape our attitudes. Moreover, it shows how these practices both sustain, and can help reduce, implicit bias. Our society is structured such that we have different expectations of women than of men, of blacks than of whites—and such expectations play an important role in causing and maintaining practices of discrimination (Beeghly & Madva, 2020; Haslanger, 2019). This means that mindshaping "is, thus, a double-edged sword: it will certainly enrich our cognitive powers, but not always in a salubrious direction" (McGeer, 2021, p. 1052). Individuals manifesting implicit bias seem to be hit by precisely this double-edged sword: in so far as they explicitly express egalitarian attitudes, our society expects them to be consistent, and to adhere to the normative stereotype of the relevant attitude. However, the implicit normative expectations that saturate social

categories like gender and race, simultaneously nudge us toward ‘discordant strains’ of discriminatory thoughts, emotions and behavior. This dynamic makes it difficult for agents to attain genuine, coherent egalitarian beliefs.

So even if representationalists correctly point out that dispositionalism is a superficial account in that it neither provides nor supports specific cognitive scientific explanations, this does not mean that dispositionalism does not have explanatory potential. Like representationalists, dispositionalists want to understand why people suffer from phenomena like implicit bias, and more in general why they have the attitudes that they do. But while representationalists look for answers in underlying processes and mechanisms, dispositionalists claim that answers to these questions require an analysis of the social and cultural processes that shape our attitudes: processes that are inherently normative in nature. Our understanding of implicit bias could benefit greatly from a genuine debate between representationalism and dispositionalism; however, such a debate might require the participants to face their deeper philosophical disagreements.

Notes

1. To clarify our use of the relevant concepts: Throughout the paper we will use the term ‘*attitude*’ as referring to any kind of evaluative relation human beings take toward an object, thus remaining non-committal on for example, whether attitudes are necessarily propositional. We use the term ‘*belief*’ as referring to one of the more common types of attitudes human beings adopt. The term ‘*implicit bias*’ refers to the situation in which someone displays biases in behavior, thoughts and/or feelings toward specific social groups that are largely outside the agent’s awareness and control (Brownstein & Saul, 2016; Johnson, 2020). Note that we focus on cases of implicit bias that are characterized by discordance. According to e.g., Holroyd (2016), it is important to also acknowledge cases in which implicit and explicit bias are aligned (as for example, in ‘wholehearted sexism’). Whereas we agree that this is an important type of bias, we think describing such cases in terms of ‘alignment’ makes sense only from a representationalist point of view: on the kind of dispositionalist account we will defend in this paper, ‘wholehearted sexism’ just display a wholehearted (non-fragmented) sexist attitude.
2. Note that Machery explicitly distinguishes attitudes from beliefs. Where beliefs are introspectable mental states, attitudes are dispositional entities that might be partly determined by beliefs. We will later discuss Schwitzgebel’s view, which also takes *beliefs* to be dispositional entities.
3. Mostly, because one can also base one’s subjective report by, for instance, reflecting on one’s own past behavior.
4. Note that Machery (2021) is even more pessimistic, taking the fact that psychological research on implicit measures is largely stagnant with respect to these fundamental anomalies as a reason to think that they cannot even measure broad traits.
5. Machery (2021) seems more amenable to this idea: “If indirect measures are only predictive in narrow contexts, then it makes no sense to conclude that one is racist, sexist, and so on, when one receives an apparently damning IAT score since racism and other biases manifest themselves across contexts: Racists think, speak, and act

- racist in many different contexts, although of course not invariably” (p. 8). Our disagreement with Machery (2016) might thus mostly concern the question how much coherence between implicit measures can be expected.
6. Note that Schwitzgebel, in contrast to Machery, is explicitly concerned with attitudes in the sense of *propositional attitudes*, such as beliefs and desires.
 7. Note that Quilty-Dunn and Mandelbaum raise multiple objections against dispositionalism. It would go beyond the scope of this paper to respond to all of them, but we do think that the objection regarding cognitive science and belief is particularly relevant in the context of this paper.
 8. Schwitzgebel (2021) himself also acknowledges that this pragmatic argument does not in itself go against all forms of representationalism.
 9. See, Schwitzgebel (2010, p. 547; 2013, p. 95) for other places where he briefly refers to the relevance of normative features.
 10. Also see, Lavelle (2021) for an analysis of culture-specific folk psychological norms.
 11. On this point, McGeer’s mindshaping theory differs from some other versions. Whereas for example, Fernandez Castro (2020) argues that “propositional attitude ascriptions emerge in contexts where our normative expectations are violated” (p. 60), thus reserving so-called ‘mentalizing’ for situations in which we require a specific explanation for the other’s behavior, McGeer holds that “typical adult human beings are in fact inveterate and prolific mentalizers both in quotidian and non-quotidian contexts” (McGeer, 2021, p. 19).
 12. This is not to say that it is impossible to provide a representationalist re-interpretation of the role of folk psychological normative criteria in the ascription of attitudes. A representationalist could argue that through social learning, such criteria are ‘implanted’ into agent’s representational systems, and that it is the presence of such representations that ultimately determines whether or not the agent has the relevant attitude. Unsurprisingly, we do not take this to be a promising response (mostly because we think changes in our normative practices *directly* change the ascriptive criteria without these changes first having to be ‘implemented’ in individuals’ representational systems), but this is certainly an issue up for debate.
 13. The position of Quilty-Dunn and Mandelbaum (2018a) is one of the main representationalist accounts of fragmentation, but of course there are other options on the table. One might also appeal, for instance, to two (or more) different *types* of representational states to explain how conflicting behavior is possible without leading to a full-fledged contradiction between these representational states (e.g., Gendler, 2008a).
 14. Furthermore, in a sense measures of implicit bias like the IAT actually ‘track’ such normative expectations. We consider it important to find out whether people are faster in grouping together stimuli related to *female* and *science* than grouping together stimuli related to *male* and *science*, because we think such a difference is normatively relevant.
 15. Note that Machery *also* mentions that a first-person attitude statement can express a commitment to a moral norm (p. 114), and that Schwitzgebel, as mentioned, now and then hints at the potentially commissive nature of avowals (e.g., Schwitzgebel, 2010, p. 547). However, neither of them provide a substantial account of avowals as commitments.
 16. Although for example, Quilty-Dunn and Mandelbaum (2018b) argue that basic rules of logic such as modus ponens are built into our cognitive architecture, thus explaining how *within* belief fragments, agents can make implicit inferences.

Disclosure statement

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

Funding

This research was supported by the Dutch Research Council (for Annemarie Kalis, grant VI.VIDI.195.116; for Harmen Ghijsen, grant VI.VENI.275.20.056);

Notes on contributors

Annemarie Kalis is Associate Professor in Theoretical Philosophy at Utrecht University. Her areas of expertise are philosophy of psychology and philosophy of mind and cognition, focusing in particular on agency and normativity. She has published on various topics such as weakness of will, intention, reasoning, the nature of folk psychology and self-control.

Harmen Ghijsen is Assistant Professor in the Center for Cognition, Culture and Language at Radboud University. His research is focused on epistemology, philosophy of mind and philosophy of science and investigates the relations between belief, bias and perception.

ORCID

Annemarie Kalis  <http://orcid.org/0000-0001-8574-492X>

Harmen Ghijsen  <http://orcid.org/0000-0002-3005-972X>

References

- Baker, L. R. (1993). What beliefs are not. In S. J. Wagner & R. Warner (Eds.), *Naturalism: A critical appraisal* (pp. 321–337). University of Notre Dame Press.
- Beeghly, E., & Madva, A. (2020). *An introduction to implicit bias: Knowledge, justice, and the social mind*. Routledge.
- Brownstein, M. (2019). Implicit bias. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Fall 2019. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2019/entries/implicit-bias>
- Brownstein, M., & Saul, J. (Eds.). (2016). *Implicit bias and philosophy, volume 1: Metaphysics and epistemology*. Oxford University Press.
- Byrd, N. (2021). What we can and can't infer about implicit bias from debiasing experiments. *Synthese*, 198, 1427–1455. <https://doi.org/10.1007/s11229-019-02128-6>
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78 (2), 67–90. doi:2025900.
- Davies, M., & Stone, T. (Eds.). (1995). *Folk psychology: The theory of mind debate*. Wiley Blackwell.
- De Bruin, L. C. (2017). First-person folk psychology: Mindreading and mindshaping. *Studia Philosophica Estonica*, 9(1), 170–183. <https://doi.org/10.12697/spe.2016.9.1.07>
- Elqayam, S., & Evans, J. S. B. (2011). Subtracting “ought” from “is”: Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, 34(5), 233–248. <https://doi.org/10.1017/S0140525X1100001X>

- Fernandez Castro, V. (2020). Regulation, normativity and folk psychology. *Topoi*, 39(1), 57–67. <https://doi.org/10.1007/s11245-017-9511-7>
- Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, 14(4), 574–595. <https://doi.org/10.1177/1745691619826015>
- Gendler, T. S. (2008a). Alief and belief. *The Journal of Philosophy*, 105(10), 634–663. <https://doi.org/10.5840/jphil20081051025>
- Gendler, T. S. (2008b). Alief in action (and reaction). *Mind & Language*, 23(5), 552–585. <https://doi.org/10.1111/j.1468-0017.2008.00352.x>
- Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, 156(1), 33–63. <https://doi.org/10.1007/s11098-011-9801-7>
- Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language*, 1(2), 158–171. <https://doi.org/10.1111/j.1468-0017.1986.tb00324.x>
- Greely, N. (2014). Troubles for a new dispositional account of belief. *Philosophy in Practice*, 8, 1–20.
- 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), 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Harmon-Jones, E., & Harmon-Jones, C. (2007). Cognitive dissonance theory after 50 years of development. *Zeitschrift für Sozialpsychologie*, 38(1), 7–16. <https://doi.org/10.1024/0044-3514.38.1.7>
- Haslanger, S. (2019). Cognition as a social skill. *Australasian Philosophical Review*, 3(1), 5–25. <https://doi.org/10.1080/24740500.2019.1705229>
- Holroyd, J. (2016). What do we want from a model of implicit cognition? *Proceedings of the Aristotelian Society*, 116(2), 153–179. <https://doi.org/10.1093/arisoc/aow005>
- Holroyd, J., Scaife, R., & Stafford, T. (2017). Responsibility for implicit bias. *Philosophy Compass*, 12(3), e12410. <https://doi.org/10.1111/phc3.12410>
- Holroyd, J., & Sweetman, J. (2016). The heterogeneity of implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology* (pp. 80–103). Oxford University Press.
- Johnson, G. M. (2020). The structure of bias. *Mind*, 129(516), 1193–1236. <https://doi.org/10.1093/mind/fzao11>
- Jost, J. T. (2019). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science*, 28(1), 10–19. <https://doi.org/10.1177/0963721418797309>
- Lavelle, J. S. (2021). The impact of culture on mindreading. *Synthese*, 198(7), 6351–6374. <https://doi.org/10.1007/s11229-019-02466-5>
- Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Nous*, 49(4), 800–823. <https://doi.org/10.1111/nous.12074>
- Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology* (pp. 104–129). Oxford University Press.
- Machery, E. (2021). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews. Cognitive Science*, 13(1), e1569. <https://doi.org/10.1002/wcs.1569>
- Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese*, 193(8), 2659–2684. <https://doi.org/10.1007/s11229-015-0874-2>
- Madva, A., & Brownstein, M. (2018). Stereotypes, prejudice, and the taxonomy of the implicit social mind 1. *Noûs*, 52(3), 611–644. <https://doi.org/10.1111/nous.12182>
- Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, 50(3), 629–658. <https://doi.org/10.1111/nous.12089>

- McGeer, V. (2007). The regulative dimension of folk psychology. In D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137–156). Springer.
- McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, 18(2), 259–281. <https://doi.org/10.1080/13869795.2015.1032331>
- McGeer, V. (2021). Enculturating folk psychologists. *Synthese*, 199, 1039–1063. <https://doi.org/10.1007/s11229-020-02760-7>
- Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton University Press.
- Porot, N., & Mandelbaum, E. (2021). The science of belief: A progress report. *Wiley Interdisciplinary Reviews. Cognitive Science*, 12(2), e1539. <https://doi.org/10.1002/wcs.1539>
- Quilty-Dunn, J., & Mandelbaum, E. (2018a). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, 175(9), 2353–2372. <https://doi.org/10.1007/s11098-017-0962-x>
- Quilty-Dunn, J., & Mandelbaum, E. (2018b). Inferential transitions. *Australasian Journal of Philosophy*, 96(3), 532–547. <https://doi.org/10.1080/00048402.2017.1358754>
- Ryle, G. (1949). *The concept of mind*. University of Chicago Press.
- Schwitzgebel, E. (2001). In-between believing. *The Philosophical Quarterly*, 51(202), 76–82. <https://doi.org/10.1111/1467-9213.00215>
- Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, 36(2), 249–275. <https://doi.org/10.1111/1468-0068.00370>
- Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, 91(4), 531–553. <https://doi.org/10.1111/j.1468-0114.2010.01381.x>
- Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann (Ed.), *New essays on belief: Constitution, content and structure* (pp. 75–99). Palgrave Macmillan.
- Schwitzgebel, E. (2021). The pragmatic metaphysics of belief. In E. Borgoni, D. Kindermann, & A. Onofri (Eds.), *The fragmented mind* (pp. 350–375). Oxford University Press.
- Stich, S., & Ravenscroft, I. (1994). What is folk psychology? *Cognition*, 50(1–3), 447–468. [https://doi.org/10.1016/0010-0277\(94\)90040-X](https://doi.org/10.1016/0010-0277(94)90040-X)
- Tumulty, M. (2011). Delusions and dispositionalism about belief. *Mind & Language*, 26(5), 596–628. <https://doi.org/10.1111/j.1468-0017.2011.01432.x>
- Tumulty, M. (2014). Managing mismatch between belief and behavior. *Pacific Philosophical Quarterly*, 95(3), 261–292. <https://doi.org/10.1111/papq.12032>
- Vierkant, T., & Paraskevides, A. (2012). Mindshaping and the intentional control of the mind. F. Paglieri. *Consciousness in Interaction: The Role of the Natural and Social Context in Shaping Consciousness*, 86. Advances in Consciousness Research. John Benjamins Pub Co. 107–123.
- Welpinghus, A. (2020). The imagination model of implicit bias. *Philosophical Studies*, 177(6), 1611–1633. <https://doi.org/10.1007/s11098-019-01277-1>
- Zawidzki, T. W. (2008). The function of folk psychology: Mind reading or mind shaping? *Philosophical Explorations*, 11(3), 193–210. <https://doi.org/10.1080/13869790802239235>
- Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.