Synthese (2023) 202:129
https://doi.org/10.1007/s11229-023-04364-3

ORIGINAL RESEARCH

Implicit bias as unintentional discrimination

Lieke Joske Franci Asma[1]

Received: 12 June 2023 / Accepted: 20 September 2023 / Published online: 17 October 2023

Abstract
In this paper, I argue that instead of primarily paying attention to the nature of implicit attitudes that are taken to cause implicit discrimination, we should investigate how discrimination can be implicit in itself. I propose to characterize implicit discrimination as unintentional discrimination: the person responds to facts unintentionally and often unconsciously which are, given their end, irrelevant and imply unfair treatment. The result is a unified account of implicit bias that allows for the different ways in which it can display itself and can be explained. Furthermore, the view can account for the central characteristics of implicit bias: (1) that it is, for a variety of reasons, difficult to control, (2) that we are not necessarily unconscious of implicit bias but not properly conscious either, and (3) that we can unintentionally discriminate regardless of whether we claim to care about fairness.

Keywords Implicit bias · Implicit attitude · Acting under a description · Automaticity · Intentional action · Discrimination

1 Introduction

The fact that agents sometimes treat persons unfairly or unfavorably on the basis of their gender, skin color, or other characteristics that imply membership of a certain social group while at the same time maintaining they are objective has worried and fascinated many psychologists and philosophers (e.g., Beeghly & Madva, 2020; Brownstein & Saul, 2016; Gawronski & De Houwer, 2014; Greenwald & Banaji, 1995; Greenwald et al., 1998; Holroyd, 2012; Mandelbaum, 2016; Toribio, 2018). One main aim of many scholars has been to explain the occurrence of such cases of implicit bias. Apparently, we cannot make sense of what we do and decide merely in terms of explicit attitudes; we must have implicit attitudes as well. In this picture, the crucial question is how these implicit attitudes can best be characterized. Are they, to give

✉ Lieke Joske Franci Asma
liekeasma@gmail.com

[1] Munich School of Philosophy, Kaulbachstraße 31/33, 80539 Munich, Germany

some examples, associations (e.g., Holroyd, 2012), unconscious beliefs (Mandelbaum, 2016), or perhaps repressed attitudes (Krickel, 2018)?

As a result of this focus on psychological explanations, less attention has been paid to the phenomenon implicit attitudes are supposed to explain: implicit bias, or implicit *discrimination*—the things people do, decide, feel, and think that are discriminatory but, in some sense, not explicit. This is noteworthy, because psychological experiments and even implicit measures like the IAT do no directly target or measure implicit attitudes. These studies and measures are concerned with *behavior*: that participants tend to *choose* the male candidate for traditionally masculine jobs while maintaining they are objective (e.g., Uhlmann & Cohen, 2005), tend to *sit* further away from black compared to white people (e.g., Dovodio et al., 2002), *'shoot'* black men even when they are holding an innocuous object (see, e.g., Payne & Correll, 2020), or *respond* faster when white faces and positive words share a key (Greenwald et al., 1998). On the basis of the finding that participants respond to certain features of the people they are concerned with, for example their gender or skin color, implicit attitudes are *ascribed* to the participants (see also Gawronski, 2019, p. 575). But if that is the line of reasoning, we should first have a clear picture of the phenomenon we are trying to explain. On what ground do we ascribe these implicit attitudes to agents? Moreover, it should be clear that they cannot be that which *characterizes* the behavior as implicit and discriminatory; that would amount to circular reasoning.

What is more, should we even assume that one kind of psychological state can explain all these cases? Indeed, research with the IAT, but also the aforementioned discussion about what implicit attitudes exactly are, suggests that the behavior we are interested in does not have one unique kind of psychological explanation (e.g., Brownstein et al., 2020; Byrd, 2019; De Houwer, 2019, p. 836; Holroyd & Sweetman, 2016; Huebner, 2016; Ito et al., 2015; Klauer et al., 2010; McFarland & Crouch, 2002; Meissner & Rothermund, 2015; Nosek et al., 2007, p. 267; Payne et al., 2017, p. 235; Toribio, 2018). The upshot is, again, that we should be paying attention to the nature of the peculiar behavior we are trying to explain. In what sense is the kind of discrimination research on implicit bias targets different from explicit discrimination?

That brings me to the aims of this paper: (1) develop a notion of implicit bias as unintentional discrimination, (2) show its advantages over other proposals, and (3) argue that the account captures the central characteristics of implicit bias. The paper is structured as follows. In Sect. 2, I develop my account of implicit bias as unintentional discrimination. I argue that in these cases, the behavior, broadly understood, is discriminatory, but the agent does not act intentionally under that description. More concretely, it implies responding to facts unintentionally and often unconsciously which are, given the end, irrelevant and imply unfair treatment. I support my proposal by pointing out that it allows for the variety of explanations of implicit bias that have been discussed in the literature, and can also account for an often overlooked type of implicit bias: cases in which the person is ignorant of what discrimination is. In Sect. 3, I compare the account to other proposals that focus on the nature of the discrimination itself, and point out the advantages of my proposal. In Sect. 4, I make clear that the view can account for the central characteristics of implicit bias: (1) that it is, for a variety of reasons, difficult to control, (2) that we are not necessarily unconscious of implicit

bias but not properly conscious either, and (3) that we can implicitly discriminate regardless of whether we claim to care about fairness. Section 5 concludes the paper.

2 Implicit bias as unintentional discrimination

I am not the first to claim that implicit discrimination has to do with an absence of intention (see, e.g., Brownstein et al., 2020; De Houwer, 2019; Dasgupta, 2013; Payne et al., 2017; Toribio, 2021; Welpinghus, 2020). These scholars have not developed this proposal in much detail, however. In this section, I fill this gap.

In order to develop my account, two aspects of the nature of intentional action need to be set out first. One crucial issue is that intentional action, like implicit bias, is often understood in terms of its psychological causes. According to the dominant causal theory of action, actions are those behaviors that are caused, in the right, way, by (combinations of) psychological states like belief-desire pairs or intentions (see, e.g., Davidson, 1963). One of the persistent problems with this view is deviant causal chains: sometimes our behavior is caused by an intention, but it still does not count as an intentional action (see Asma, 2023 for a more detailed discussion). Davidson (1973, p. 79) gives the following example: a climber may want to rid himself of the weight of another climber, and know that by loosening his hold on the rope he could do so. But this belief and want unnerve him so that he loosens the hold accidentally. So far, no solution to this problem has received general support (e.g., Mayr, 2011, p. 128; Schlosser, 2007, p. 191; Steward, 2012, pp. 57–58). In fact, it could be argued that the causal history is taken to be right or deviant, depending on whether the result is taken to be an intentional action or an accident (see, Asma, 2023). To use the example of the climber: letting go of the rope is an accident or an action *in itself*, and on the basis of that the causal chain, for example the nervousness, is taken to be deviant or not. And if that is the case, we should pay attention to the intentional action itself.

One valuable source for understanding the nature of intentional action is G. E. M. Anscombe's (1963) work on intentional action. She argues that intentional actions have an inherent means-end structure: we do one thing, for example cutting at onion, to do something else, for example cooking spaghetti, at the same time. This view emphasizes that in order to reach our end successfully, spaghetti for dinner for example, we need to identify the relevant facts in the environment and respond to these facts correctly (see Ford, 2016). If we fail to recognize and respond to the facts correctly, for example mistake an apple for an onion, we will not reach our end successfully. Importantly, a defender of the causal theory of action does not need to deny this. They can, and probably do, also maintain that actions are directed at ends, and that agents have to respond to the facts in the environment in order to reach these ends. A difference with Anscombean action theory is that the means-end structure plays a less prominent role in their account of intentional action. For the purposes of this paper, however, how intentional actions can best be characterized and whether they are produced by practical knowledge, as Anscombe maintains, is not of crucial importance. What is important, is that intentional actions have this means-end structure.[1]

[1] Thanks to an anonymous reviewer for pushing me on this point.

A second preliminary point to make is that in the philosophy of action, it is widely accepted that we act intentionally under certain descriptions, but not under others (e.g., Anscombe, 1963; Davidson, 1963, see also Gawronski et al., 2022b, p. 224). To give an example: I could be sitting in the park explaining to a friend what the paper I am writing is about, while at the same time scaring off birds by speaking loudly, alerting a burglar that I am not home, flattening the grass, burning five calories, and speaking for four minutes and 25 seconds. Only under this first description am I acting intentionally, but at the same time I do the other things unintentionally as well. I do not intend to do them, but they happen in virtue of and while I am performing a different intentional action. Typically, these descriptions depend on the larger context in which the person acts—scaring off birds or alerting the burglar –, or specific bodily processes—burning five calories or speaking for four minutes and 25 seconds.

Once we recognize that implicit discrimination is the problem we need to make sense of and explain, it is clear that the nature of discrimination should be the starting point for our analysis. The question is, first and foremost, whether the description 'discrimination' applies to what the agent is doing, saying, feeling, or thinking. And just like whether or not I intended to or recognized that I was scaring off the birds, whether an agent intends to or recognizes their behavior as discriminatory is not relevant for whether the description applies. Scaring off birds involves the birds flying or running away in response to something I did. Similarly, discrimination, the harmful kind that scholars working on implicit bias worry about, involves the unfair or unfavorable treatment of a person in virtue of their membership of a certain social group. Discrimination has a certain character, just like scaring off birds has a certain character. Accordingly, we can behave, decide, think, or feel under the description 'discrimination', independently of whether we acted intentionally under that description. It is not up to me whether what I do or say counts as unfair or unfavorable treatment; our shared understanding of the nature of discrimination forms the starting point.

This brings me back to the first point, the means-end structure of intentional action. Often, discrimination has a structure that is similar to intentional action: it involves a relationship between a fact about a person and the end of the action, or, at least, a fact about a person and a certain response. The end or response are crucial for whether it counts as discrimination. After all, not discriminating—treating a person fairly—is not the same as being blind to facts that imply a person's membership of a certain social group. Sometimes responding to facts that imply membership of a certain social group does not amount to discrimination, but is the fair and right thing to do. Whether a person coming from the other side on the sidewalk is in a wheelchair, for example, is relevant for whether I step off the sidewalk. In order to treat the person fairly, this fact about the person should be taken into account. Similarly, sometimes taking into account pregnancy, nationality, or religious beliefs does not amount to discrimination, but rather to fair treatment.

In cases of discrimination—again, the unfair kind—something is going wrong: agents use facts that they should not be using in light of their end. For example, when selecting a candidate for a job, in most cases, using gender amounts to unfair treatment and, therefore, discrimination. Similarly, deciding whether to shoot a person on the basis of their skin color is unfair and racist. This does not depend on the perspective of the person making the decision or, in an experiment, on the researcher's perspective,

i.e., whether they find it desirable to use these social facts. The point is that from the perspective of unfair and unfavorable treatment, using these facts for this end is discrimination. In order to not discriminate, candidates should be selected because they have the right skills and knowledge, and shooting a person is justified when they pose immanent threat, for example when they threaten to use a gun.[2] That means that often, discrimination takes place within an intentional action: the agent has an end, but responds to a fact that they should not be responding to. In relation to the characteristics of intentional action I put forth, this implies that discrimination is typically wrong in two ways: (1) it involves the unfair or unfavorable treatment of a person in virtue of their membership of a certain social group, and (2) it hampers you from reaching your end, e.g., selecting the best candidate for the job or shooting truly dangerous people, because you are responding to facts that are irrelevant given your end.

Unintentional discrimination, then, involves unintentionally responding to facts that, in light of your end, amounts to treating a person unfairly or unfavorably in virtue of their membership of a certain social group. In the shooter task for example (see, e.g., Payne & Correll, 2020), the task or end of the participants was to shoot people that pose immanent threat, i.e., the ones holding a weapon. The findings show that people responded to a fact that amounted to unfair treatment: the skin color of the persons. Since the findings suggest that they did not intentionally respond to the skin color of the black men, the conclusion is that they discriminated unintentionally.

A similar interpretation applies to Uhlmann and Cohen's (2005) experiment. In this study, participants had to choose between two candidates for a job as police chief. Either they had to choose between Michelle who was more streetwise and Michael who was formally educated, or they had to choose between a formally educated Michelle and a streetwise Michael. That is, their end was to select the best police chief, and the facts they could use were the credential and gender of the candidates. Deciding on the basis of gender would amount to discrimination, while choosing on the basis of the credential would amount to a fair decision—this information is relevant in light of the end, and deciding as such would not, at least not obviously, imply unfair treatment in virtue of membership of a certain social group.

The findings show that a substantial amount of the participants did take into account gender in their decision. The study also suggests that they did so unintentionally. Of course, they must have been conscious of selecting Michael and, with that, selecting the male candidate. But selecting the male candidate does not necessarily amount to unfair treatment; sometimes the male candidate simply is the best choice, given his credentials. The reason why it is discrimination is not that they selected Michael, but that they selected Michael *because* he is male. The problem is that they, given their end, responded to this fact. Since the participants did maintain that they made an objective decision, it is likely that mistakenly took themselves to be responding to the

[2] This can be connected to Payne and Correll's (2020, p. 4) definition of bias, which they take to be a shift in the decision rule that guides an individual's behavior. They use the example of a traffic officer who is monitoring a stretch of highway where the speed limit is 65mph, but who stops black drivers that exceed 70mph, and white drivers when they exceed 80mph. Whether the facts are relevant is implicit in their definition: skin color *should not* change the decision, while speed *should*, and that is why the police officer is biased if skin color does change the decision. I do think their definition is limited because they do not state this explicitly; sometimes decision rules *should* change, for example when someone driving 70mph has no lights on.

credential, not to gender.[3] I say more about how this is possible later in this section. Interestingly, these reflections suggest that unintentional action not only depends on the larger context in which the person acts or on specific bodily processes, as we saw before. It can also depend on mistakenly responding to certain facts while performing an intentional action. Similarly, if my end is to cycle to friends but I take the route to work instead, I am unintentionally cycling towards the university.[4]

Even though we can be conscious of our unintentional actions, e.g., I could be messing up the kitchen unintentionally while making soup but still be conscious of doing so, the participants in the aforementioned studies do not seem to be conscious of making a sexist or racist decision either. This is in line with Gawronski et al.'s (2006) proposal that even though findings show that we are conscious of our implicit attitudes (or can become conscious of them), we may not be conscious of how these attitudes *impact* our judgments, decisions, or (un)intentional actions.

An important question is how this is possible. How can we unintentionally, and unconsciously, be responding to facts that would make what we do an example of discrimination? One possible approach would be, in line with the picture of implicit bias that I set out in the introduction, to ascribe a certain implicit psychological state to these agents. From that perspective, this proposal makes room for the possibility that several psychological states and processes can contribute to unintentional and unconscious discrimination in a variety of ways. Crucially, however, not only psychological processes or mechanisms may contribute to unintentional and unconscious discrimination. My proposal opens up the possibility that other factors may play a role as well, for example how relevant and irrelevant facts are presented, the extent to which they are related, or our failure to understand the nature of discrimination. After all, the proposal entails that the starting point is not our psychology, but unfair treatment. And since whether something counts as unfair treatment, i.e., discrimination, is independent of a certain psychological cause, a variety of explanations are possible.[5]

Firstly, it could be the case that we respond habitually or automatically to social facts (see, e.g., De Houwer, 2019; Toribio, 2021). We encounter persons of a certain ethnicity, gender, nationality, or sexual orientation for example, and we simply find ourselves experiencing fear, disappointment, or even disgust, or make assumptions about their personality, the kind of life they lead, the job they have, etcetera. This explanation probably applies to the shooter task and the IAT, but also to microaggressions. We do not typically choose to smile less or hesitate more when we interact with people from a certain social group, or tense up when members of a certain social group enter the elevator. Even though these automatic discriminatory responses do

[3] For the purposes of this paper, I assume that they are not lying. And even if it would turn out they are, or sometimes agents in such situations are, this is still a useful account of how discrimination can be unintentional.

[4] Thanks to Anna Moltchanova and an anonymous reviewer for pushing me on these points.

[5] A reviewer for this journal pointed out that this proposal may not help to advance the science or offer pathways forward for dealing with the complexities of discriminatory behavior. Even though I agree that the account suggests that there is no easy fix or one unifying explanation, I think the proposal, by explicitly distinguishing between the nature and explanation of unintentional discrimination, directs our attention to alternative explanations and novel directions of research. For example, as I point out later, we could investigate the role of misattribution or examine how people from different backgrounds conceptualize discrimination or unfair treatment.

not take place within an intentional action, they are also examples of unintentional discrimination and can therefore be integrated in this account. What is more, even here the relationship between facts and response is crucial. Some facts justify smiling less or tensing up; skin color, however, is not one of them. What is more, the focus on discrimination as a description of behaviors, broadly understood, makes clear that a person could be understanding their response under a different description. A woman experiencing fear when she sees a black man entering the elevator may think she is responding to the fact that he is a man, even though she is responding to his skin color, i.e., she would have responded differently if he were Asian or white. She does not interpret her response as discrimination, even though it is. Importantly, even though we could, in line with the widely accepted picture of implicit bias that I discussed in the introduction, ascribe an implicit psychological state to explain these automatic responses, that does not have to be the only or full story. Our automatic or habitual discriminatory response may also be the result of embodied habits, for example (see, e.g., Leboeuf, 2020; Ngo, 2016).

Secondly, the problem could be that we are not fully conscious of the facts we are responding to, because of which we run a higher risk of using the wrong facts (or the facts wrongly) in light of our end. Not being conscious of the facts we are responding to and not realizing what effect this information has on how we respond is quite common, however. Arruda and Povinelli (2018) explain this nicely when they describe a berry picker who "is sufficiently adjusted to the delicate touch that one must take with each berry and the spines on the plant" so that she does not need to consciously recognize and respond to the facts about each berry (p. 13). The crucial difference with unintentional discrimination is that the berry picker is an expert; she is accurately responding to the relevant facts in the service of her larger intentional action, and, apparently, she can do so even if she does not consciously register all the facts. In cases of unintentional discrimination, the findings suggest that many of us do not have this kind of unconscious sensitivity. We often respond to irrelevant facts, facts that even we think we should not be using given our end.

Thirdly, the problem could be that perception of the facts is discriminatory. For example, participants in the shooter task could actually be perceiving a weapon instead of an innocuous object when the object is held by a black man (see, e.g., Saul, 2013; Siegel, 2020 for discussion). That is, we, or some of us, may perceive the world through a discriminatory lens, because of which we respond to the facts in such a way that it amounts to unfair treatment.

A fourth possible explanation could be misattribution, which involves mistaking an effect of one source for the effect of another (see, e.g., March, 2018; Payne et al., 2005; Schachter & Singer, 1962). A classic study in this field is one by Dutton and Aron (1974), which shows that men misattribute their arousal, that was in fact caused by a precarious bridge, as a sign of being attracted to the woman directly across it. Similarly, it could be the case that the participants in Uhlmann and Cohen's (2005) study have a positive feeling about Michael because of his gender, attribute this feeling to his credential, and conclude that they chose Michael not because of his gender but because of his credential. What is interesting about this explanation, is that the decision-making situation has to be such that it allows for misattribution. We do not misattribute our evaluation to a fact that is not presented at the same time and

location or is substantially different from the source (see, e.g., March et al., 2018). This explanation thus suggests that in order to fully understand unintentional and unconscious discrimination, we should look beyond psychological states and processes (see, Asma, unpublished manuscript for further discussion).[6]

Finally, unintentional discrimination may also be the result of not understanding what your decisions and actions mean at a higher or different level of description. A person may think that it is not discriminatory to select the male candidate as police chief regardless of the credentials he and the women candidate possess, simply because most police chiefs are men. Or they may think it is okay to ask a women colleague who is more quiet than usual whether she has PMS, or to persistently ask the person with an accent where they are *really* from. This kind of unintentional discrimination involves responding to social facts intentionally and (probably) consciously, but implies not recognizing your action or decision under the description of discrimination, i.e., the unfair treatment of a person in virtue of her membership of a certain social group. In such a case, an agent can still claim that they are not discriminating, but that is because they do not understand what it means to treat people fairly (see, e.g., Kalis & Ghijsen, 2022; Machery, 2016). As I pointed out before, an agent cannot simply decide whether their behavior reflects fairness or discrimination; their behavior can count as such even if the agent does not think of them under this description. This depends on our shared meaning of what discrimination and fairness amount to.[7]

This kind of unintentional discrimination may be the result of blindly follow existing social norms and structures without reflecting on their meaning (e.g., Davidson & Kelly, 2020). We may have grown up in an environment in which it is normal to make jokes about certain social groups or asking women and not men how they plan to combine work and family. Only later we may realize that the norms, practices, and structures in our society are discriminatory, and that we have been acting in line with them without realizing that they are not as innocent as we thought. Understanding implicit bias as unintentional discrimination, then, has the additional advantage of

[6] Even though a full comparison to Gaertner and Dovidio's (e.g., 2005) account of aversive racism goes beyond the scope of this paper, it should be noted that they also strongly emphasize role of contextual factors: "Discrimination will tend to occur in situations in which normative structure is weak, when the guidelines for appropriate behavior are vague, or when the basis for social judgment is ambiguous. In addition, discrimination will occur when an aversive racist can justify or rationalize a negative response on the basis of some factor other than race" (p. 620). It will be interesting to compare this interpretation to a proposal in which misattribution plays a central role. Thanks to an anonymous reviewer for attending me to this model.

[7] I should point out that I do not think that fairness and discrimination have undisputed descriptions that are set in stone. Even though there are clear core cases, some cases are a topic of discussion. This (again) illuminates that ascribing implicit attitudes to agents is not as straightforward as sometimes seems to be suggested in the literature on implicit bias. Thanks to Annemarie Kalis for addressing this. Payne et al. (2017) make an interesting related point about ascribing implicit attitudes: "Suppose, for example, that a research participant is aware that she has stereotypical thoughts passing through her mind but does not think that means that she dislikes the group in question. The researcher, meanwhile, thinks that the presence of stereotypical thoughts does indicate prejudice. If that participant displays bias on an implicit test but reports low levels of prejudice on an explicit questionnaire, the stage is set for 'introspectively unidentified (or inaccurately identified) traces of past experience' that could constitute unconsciousness (Greenwald & Banaji, 1995). However, we are in no position to know whether the inaccurate identification is on the part of the participant or the researcher. Distilling the 'real meaning' of concept accessibility requires an act of interpretation—by both the participant and the researcher—and sometimes they will disagree." (p. 243).

bridging the gap between implicit bias as it is studied by psychologists, and structural explanations of discrimination (see, e.g., Haslanger, 2015; Lauer, 2019).

In fact, my proposal can be seen as an addition to recent views that aim to move beyond the dichotomy between individual and structural explanations of injustice, or at least soften the distinction between the two (e.g., Ayala-López & Beeghly, 2020; Davidson & Kelly, 2020; Madva, 2016; Soon, 2019; Zheng, 2018). These scholars try to find a middle ground between the view that we should mainly focus on individual minds in combatting injustice, and the view that the causes of injustice are unjust social structures, and "biased minds are merely symptomatic of a deeper problem" (Soon, 2019, p. 1858). Roughly, these scholars maintain "that there are continuous feedback loops of mutual influence between individuals and structures, minds and social worlds" (Davidson & Kelly, 2020, p. 193). My account of unintentional discrimination shows that social structures and individual psychologies not only interact, but that the world in which we act can determine the unintentional meaning certain actions have. Selecting the only male candidate as CEO has a different meaning in a world where eighty percent of CEOs are men compared to a world in which fifty percent is, and cancelling a bus line that is mostly used by minorities that go to work has a different meaning than cancelling one that is used by students going to the beach. I would argue, then, that social structures and individual minds are not only deeply intertwined and both causally contribute to injustice, but that whether we even have a case of unintentional discrimination can be, at least partly, constituted by the context in which the agent acts.

An important implication is that unintentional discrimination does not have to be caused by an implicit psychological state with the same content; it may as well be the result of a lack of understanding of the nature of discrimination. In order to be egalitarian, especially in a society in which sexist and racist practices and structures exist and persist, we have to work to overcome unintentional discrimination (see, e.g., Huebner, 2016). It also implies that I cannot simply study my conscious and unconscious mind, and conclude that I treat people fairly if I do not encounter biased attitudes. In order to avoid discrimination, we have to realize that the meaning of our actions goes beyond our intentions and unconscious influences. Treating people fairly is not a trait we just have and we are done with at some point; we have to reflect on how we act, behave, feel, and think, what the wider meaning of our behavior is, and which institutionalized patterns and norms need to be intervened on (see Lauer, 2019).

My proposal to understand implicit bias as unintentional discrimination, then, implies casting a wider net. We should not merely focus on the unintentional responses that are measured by means of the IAT or the shooter task for example, responses of which participants even turn out to be conscious or are able to become conscious of (see, e.g., Gawronski et al., 2006; Payne & Correll, 2020; see also Reis-Dennis & Yao, 2021), but also pay attention to unintentional and unconscious discrimination that takes place within or that are implied in actions that are conscious and intentional under a different description.

Critics may think this is a step too far, for two reasons. First of all, they may think that we end up ascribing unintentional discrimination to actions and decisions of agents that has nothing to do with their individual psychologies and that they may not, or hardly, have any control over. I think, however, that it would not be bad to move

away from identifying an individual as having a certain characteristic, i.e., as *having* an implicit attitude or *being* implicitly biased. My suggestion is that implicit bias is first and foremost a characteristic of behavior, decisions, thoughts, feelings, and also of norms and practices, and that agents that are actually committed to fairness can have blind spots.

Secondly, critics may think that taking the aforementioned cases to be examples of unintentional discrimination is unacceptable. That would mean that people can simply fail to recognize their actions as discrimination, and claim that they are responding to the relevant facts, because in their view gender and skin color are relevant. I do not see this as a problem for the view, however. The main reason is that my aim, in line with much of the literature on implicit bias, is to explain how discrimination can persist even if people take themselves to be objective. From that perspective, this type of ignorance is of crucial importance. What is more, I do not mean to imply that people are off the hook. They are still discriminating, and the extent to which they are excused would depend on the specific circumstances of the case.

An interesting question is whether any conscious and intentional discrimination remains. After all, given what I said before it is likely that someone who treats a person differently on the basis of their skin color, gender, body weight, or age for example may think, for whatever reason, that this amounts to fair treatment. My account does not exclude the possibility that people can discriminate explicitly, however. A person may know, for example, that rejecting someone on the basis of skin color or gender is discrimination, but decide to choose the white male anyway because they do not want their work environment to change. They know that they should not be using this information, but yet they do, because for them another end prevails over the end of preventing discrimination or even choosing the best candidate for the job. It definitely is an upshot of the account, however, that many more cases of discrimination may in some sense be unintentional and unconscious. People may think that their action or decision is justified, while in fact it amounts to unfair treatment. In my view, this is a valuable insight, and may help us understand how conflicts in this area occur, i.e., how people may have completely opposed perspectives on the nature of certain actions and decisions. It also emphasizes that preventing discrimination and the harms associated with it is a matter of learning what discrimination and fairness amount to. We often do not simply know what the right thing to do is. In that sense, in line with what I argued before, unconscious and unintentional discrimination comes in many forms. There is an important difference between not realizing that you are taking into account gender or skin color in your decision, and not realizing that taking into account gender or skin color, in light of your specific end, is an act of discrimination. Also, in certain cases, we may have to reject a person's claim to ignorance. Just like claiming that poisoning someone is not (attempted) murder, claiming that considering skin color while hiring a candidate is not discrimination simply doesn't make sense.[8]

[8] Thanks to an anonymous reviewer for pushing me on this point.

3 Comparison to other proposals

Other scholars have recently also argued that in order to understand implicit bias, we should focus on what agents actually do (broadly understood) instead of merely analyzing a hidden psychological state that may be causally responsible for it (De Houwer, 2019; Gawronski et al., 2022a; Toribio, 2021, see also Baston & Vosgerau, 2016; Machery, 2016; Payne et al., 2017). In this section, I compare my account to these other proposals, and argue that my account has advantages over the others.

3.1 Implicit discrimination as habitual action

The first proposal I will discuss is Toribio's (2021). Toribio argues that "the discriminatory behavior triggered by implicit biases is best understood as a type of habitual action—as a harmful, yet deeply entrenched, passively acquired, and socially relevant type of habit" (p. 2).[9] Implicit discrimination or, in her words, habitual behavior that results from implicit attitudes, is "unintentional relative to at least one of its features, and it is, to some degree and relative to some standard, goal-independent, uncontrolled, autonomous, purely stimulus driven, unconscious, efficient, and fast" (p. 6).

Even though habitual behavior may capture an important sense in which implicit discrimination manifests itself, as a characterization of implicit bias it is incomplete, since it does not account for implicit discrimination in considered decisions, like in Uhlmann and Cohen's (2005) study. Selecting a candidate as police chief is not habitual behavior we just happen to perform; it is not an uncontrolled, autonomous, purely stimulus driven, unconscious, efficient, or fast response. These participants made a conscious choice, and could take their time to do so. Research shows that these types of considered, and yet implicitly discriminating, decisions are common (see, e.g., Antony, 2016; Gawronski et al., 2022b, p. 227; Welpinghus, 2020), and therefore a full account of implicit bias should account for those as well. An understanding of implicit discrimination as unintentional includes these cases.

Secondly, in certain situations, implicit discrimination seems to be the result of a *disruption* of habit. Our social interactions are to a substantial extent driven by habitual behavior. We do not reflect on how often we look into the other person's eyes, the distance between ourselves and the other, or whether we smile for example. However, if we encounter a person with a different skin color, for some people, a person who does not regularly interact with people of color for example, the context of the habitual action may have substantially changed, because of which her habitual way of interacting may be suspended (see, e.g., Wood et al., 2005). Research indeed shows that larger IAT scores predicted "greater speaking time, more smiling, more extemporaneous social comments, fewer speech errors, and fewer speech hesitations in interactions with the White (vs Black) experimenter" (McConnell & Leibold, 2001, p. 439). Especially the speech errors and hesitations suggest that the interaction with the white experimenter is habitual, while the interaction with the black experimenter is not. Seemingly, the fact that the person the participants were interacting with was

[9] Leboeuf (2020) and Ngo (2016) also emphasize the role of racist habits in (implicitly) biased behavior. Since their goal is not to give an (exhaustive) account of implicit bias, I will not discuss their accounts here.

of color disrupted rather than activated their habitual way of interacting (see Leboeuf, 2020, pp. 46–47). Even though habits may be one explanation of why unintentional discrimination occurs and persists, then, it does not capture all kinds of implicit bias.

3.2 Implicit discrimination as automatic discrimination

Another proposal on how to make sense of implicit discrimination has recently been developed by De Houwer (2019; De Houwer & Boddez, 2022). De Houwer (2019) argues that implicit bias "is seen as behavior that is automatically influenced by cues indicative of the social group to which others belong" (p. 1), and that "[t]he influence of these social cues can be labeled as implicit when it occurs quickly, effortlessly, unintentionally, unconsciously, or in a way that is difficult to control" (p. 2).

This proposal is in line with the view of automaticity De Houwer has developed with Moors (2006). In this influential paper, they argue that automaticity should be seen as an umbrella term: neither of the aforementioned features is necessary for a process to count as automatic. Discriminatory behavior, then, would be implicit if it is automatic in (at least) one of these ways. Relatedly, in more recent work, De Houwer and Boddez (2022) argue that implicit discrimination can involve several automaticity features, that all refer to different ways in which conditions for cognitive processing are suboptimal, for example a lack of awareness or motivation, or deciding under time pressure.

One advantage of this proposal is that it does not explain implicit discrimination in terms of one underlying psychological state, and therefore allows for different (psychological) states, processes, or circumstances to do the explanatory work. In line with that, it emphasizes that the problem of implicit discrimination is, in first instance, about how people respond to the facts in the environment. I do not think, however, that the proposal paints a full picture, because the central focus is on the causal history of the behavior. Automatic influences, however, could be and often are part of a well-executed and conscious intentional action. Think of cycling to work or making breakfast; these are full-blown intentional actions, even though we respond to the facts largely automatically (e.g., Arruda & Povinelli, 2018). Similarly, I could quickly and effortlessly, using minimal attentional capacity (Moors & De Houwer, 2006, p. 298), step off the sidewalk to make room for a person in a wheelchair, but still do so intentionally and consciously. Or, the participants in Uhlmann and Cohen's (2005) study could, as a result of the name Michelle, have a bad feeling about this candidate, consciously decide that men simply are better suited to be police chiefs than women, and intentionally and consciously choose the male candidate because of his gender. Automatic processes played a role, but the action itself is intentional and conscious. An advantage of understanding implicit bias as unintentional discrimination is that it specifically captures the peculiar nature of the behavior itself. From that perspective, De Houwer's (2019) proposal does contribute to developing important insights on how we can respond to irrelevant facts unintentionally and unconsciously, but as a characterization of implicit discrimination it would be incomplete: not only is

our response to the facts automatic, unintentional discrimination is unintentional and often unconscious in itself as well.[10]

Secondly, De Houwer and Boddez (2022) make a connection between implicitness and automaticity on the hand, and suboptimal conditions on the other. Implicit discrimination is biased behavior under suboptimal conditions. On first sight, it makes perfect sense to think of lack of motivation, lack of awareness, or time pressure as suboptimal conditions, but why exactly are they suboptimal? They are not suboptimal in and of themselves. After all, we can think of examples where lack of awareness of certain aspects of the action is optimal, think of running down the stairs or driving a car. That suggests that the conditions are suboptimal *because* they lead to discriminatory, i.e., suboptimal, behavior. That invites a further question: why is discriminatory behavior taken to be suboptimal? An obvious answer is that something has gone wrong: the agent behaved in a way that is taken to be problematic, either by themselves, the researcher, or in virtue of our shared meaning of discrimination as unfair treatment. This suggests that implicit discrimination *and* the conditions, like lack of awareness or time, are suboptimal because they are or lead to behaviors that are out of line with what (we think) is the right thing to do, i.e., they are suboptimal in virtue of their contribution to unintentional discrimination.

3.3 Implicit discrimination as unconscious discrimination

Finally, I want to compare my account to a proposal that has recently been defended by Gawronski et al., (2022a, 2022b). They argue that implicit discrimination should be understood as unconscious: implicit bias involves making a distinction on the basis of skin color for example, and being unconscious of the fact that your decision or behavior was influenced as such. Accordingly, they define implicit bias as "unconscious effects of social category cues […] on behavioral responses" (p. 140). The unconscious *effects* part is crucial. It is in line with earlier work, in which Gawronski et al. (2006) have argued that the characteristic feature of implicit attitudes is not how they are formed or whether we are conscious of the content, but the impact they have on what we do. We are unconscious of the fact that a person's gender, skin color, or sexual orientation plays a role in our judgments, decisions, or behaviors.

As I have pointed out, responding to facts unconsciously may indeed be an important explanation of how it is possible that we unintentionally and unconsciously discriminate. One problem I see with the proposal, however, is that it implies that the absence of discrimination is the same as being color blind (see Dovidio & Kunst, 2022; Norman & Chen, 2022; Schmader et al., 2022). It does not consider that sometimes we should respond to social facts; we should step off the sidewalk when someone in a wheelchair comes from the other side, for example. A different way to put this would be that we need to distinguish between equality and equity: sometimes we should make distinctions to not treat people unfairly; think of supporting young mothers or people with

[10] This line of criticism also applies to Payne et al.'s (2017) account of implicit bias, according to which implicit attitudes are gut reactions and thoughts that merely momentarily pass through the person's mind, which we do not experience as arising from stable attitudes and traits (p. 241), and do not depend upon intent (p. 240). Again, even though these biased thoughts and feelings just 'pop up', that does not mean that the resulting decision is implicitly biased as well.

certain disabilities in the workplace. Indeed, the examples Gawronski et al. (2022a) use in the beginning of their paper are all about unfair treatment, why not explicitly make that part of a definition of implicit discrimination?

In response to this line of criticism, Gawronski et al. (2022b) argue that the behavioral response and the undesirability of the effects should be kept apart, because whether the behavioral response has undesirable consequences depends on values and goals. They use the example of a woman calling the police because families are barbecuing in the park, but only when they have dark brown skin (p. 221). As Gawronski et al. state, whether this is desirable or not depends whether you want to maintain or reduce existing social hierarchies.

But whether someone has these goals or values does not matter when it comes to whether we have a case of discrimination. Calling the police only if the people have dark brown skin *is discrimination*, it is treating a person unfairly in virtue of their membership of a certain social group. Given the end, making sure that people do not barbecue in the park, skin color is irrelevant and taking it into account amounts to unfair treatment. This does not depend on whether the person cares about fairness or not. Of course, an agent could think that unfair treatment of members of certain social groups is a good thing or not something to explicitly take into account, but that is irrelevant for whether it counts as discrimination or not. What is more, I do not think that discriminatory behavior receives its nature from its effects (p. 221): even if micro-aggressions would have no substantial impact, it would still be discrimination. It is inherently unfair.

I agree with Gawronski et al. (2022b) that whether something counts as discrimination does depend on context and history to a certain extent, but that is exactly why the field is, and should be, interested in gender and skin color playing a role in how we treat people. Gender and skin color are in most cases and for most decisions irrelevant, and taking such a factor into account often implies unfair treatment. At the same time, these facts often have played a role in how a person was treated, and these decisions and actions in the past still have a substantial influence today. But that does not at all imply that discrimination depends on an agent's goals and values; it depends on our shared understanding of fairness and discrimination.

What is more, if we, as they and others suggest (see De Houwer, 2019; Payne & Correll, 2020), let go of the distinction between discrimination, understood as unfair and unfavorable treatment, and discrimination in a neutral sense, shooting a person who is threatening with a weapon is discrimination as well, and so is choosing a candidate who has the right credentials. Every decision in which we use facts about a person to distinguish or make a choice would count as discrimination. As a result, we run the risk of losing track of what we were worried about in the first place. The problem of discrimination is, essentially, that we use social facts that we, given our ends and in relation to what it means to treat someone (un)fairly, should not be using. Discrimination only depends on our ends in the sense that some ends imply that responding to skin color or gender is unfair and discriminatory, while for other goals it is justified to use these social facts—in case of gender and skin color, most of the time it isn't. The nature of discrimination, the unfair or unfavorable treatment of a person in virtue of their membership of a certain social group, forms the starting

point, and that is how it should be. Whether I care about discrimination or whether an equal society is my goal, is irrelevant.

In relation to this, I take it to be an advantage of my proposal that it captures that sometimes it is difficult to identify unintentional discrimination, and that we can disagree about whether certain behavior or decisions count as such. Some cases are very clear; we should not reject a woman for a job as police chief simply because she is a woman or shoot a person because they are black. But sometimes things are less straightforward. Was a person treated differently because of their gender, skin color, or sexual orientation, or was the rejection based on relevant information? And in which cases, given which ends, does membership of a certain social group count as information that should be used? In real life, these are common discussions, and we should take them seriously and try to understand why they take place. An account that can capture this aspect of implicit discrimination seems to me to be the stronger contender.

What is more, Gawronski et al., (2022a, p. 145 & p. 146) want to exclude cases where people are conscious but are not able to do something about their discriminatory behavior. They argue for this mainly to support their view that implicit bias and bias on implicit measures are distinct, and want to avoid that the strong focus on the IAT leaves out cases they are particularly worried about. I completely agree that the IAT should play a less central role in our understanding of implicit bias. But why not try to come up with an account that captures all these cases? If we understand implicit bias as unintentional discrimination, we can account for automatic discriminatory responses and discrimination in considered decisions.

4 Unintentional discrimination and the characteristics of implicit bias

In the previous sections, I argued that we should understand implicit bias as unintentional discrimination. I made clear how this notion makes room for a variety of explanations, and still provides a unitary account of implicit bias under its different guises: it includes automatic and habitual responses, but also considered decisions. What is more, it even includes a kind of implicit bias that has often been overlooked: unintentional discrimination that is the result of ignorance of what discrimination and fairness amount to. In this final section, I show how my proposal accounts for some central characteristics of implicit bias.

Firstly, my proposal shows why implicit bias is difficult to control. In many cases, agents have a hard time preventing their behavior to be discriminatory, because they have to respond to the facts quickly, for example in the IAT or shooter tasks, or because they happen to experience fear and have a hard time controlling their emotional and physical response. However, my proposal shows that there is a different sense in which agents in situations such as these, but also when they have to make a considered decision, can have a hard time controlling their unintentional discrimination: they are not conscious of acting under a, for example, sexist or racist description, either because they do not fully understand what discrimination is, or because they are not conscious of responding to irrelevant social facts that implies unjust treatment. The latter may be the result of unconsciously responding to the facts, discriminatory perception (e.g.,

Siegel, 2020 or misattribution (e.g., March et al., 2018). What is more, as Uhlmann and Cohen's (2005) experiment makes clear for example, unintentional discrimination is often not reflected in one thought, feeling, or decision, but in a *pattern* of thoughts, feelings, behaviors, and decisions. The problem is not that participants take a certain amount of time to press a certain key, sit five feet away from a black person, or do not choose the women candidate. The problem is that their response *differs* from what it would be if the face or person would have been white or male. This is difficult to recognize in individual expressions. In order to control unintentional discrimination, then, it would not simply be a matter of becoming conscious of one act or decision, but of interpreting patterns in your responses. Even if there is room for reflection and conscious decision making, then, the individual participant may not realize something is off, that her choice is out of line with her convictions.

Secondly, the proposal accounts for the insight that we are not necessarily unconscious of implicit bias but not properly conscious either. First of all, it provides an interesting perspective on Gawronski et al.'s (2006) claim that we are conscious of our implicit attitudes (see also Goedderz et al., 2023; Hahn et al., 2014; Hall & Payne, 2010; Nier, 2005; Olson & Fazio, 2003), but not of the impact they have on what we do; i.e., we are not conscious of discriminating. We should recognize, however, that sometimes we are able to become conscious of discriminating, while in other cases we seem to be unable to recognize our action as such. In the shooter task for example, participants recognize responding to the wrong facts and are able to correct themselves (see Payne & Correll, 2020), while in Uhlmann and Cohen's (2005) study participants seem to lack this kind of consciousness. We should reflect on how this difference can be accounted for. But more important in relation to my proposal: even if agents become conscious of their unintentional discrimination, this would not amount to explicit discrimination (see, e.g., Levy, 2014). Many philosophers of action maintain that, at least most of the time, we have direct, non-observational knowledge of our intentional actions (e.g., Anscombe, 1963; Davidson, 1963; Marcus, 2012; Setiya, 2017). When I intentionally walk to the supermarket or buy ice cream for example, I know this is what I am doing without observing my bodily movements, the context, or reflecting on my psychological states. We do not have this unique kind of knowledge of our unintentional actions. As a result, we could become conscious of our discriminatory unintentional actions, but this is a different kind of consciousness than we have of our intentional actions. The latter type of consciousness is observational and third-personal; it is something we merely become *conscious of* (see Finkelstein, 2003, see also Berger, 2020; Levy, 2014; Rosenthal, 2005). In order to find out that you are unintentional discriminating, you would have to pay attention to which facts you may be responding to, reflect on other meanings your decisions and behavior, or recognize patterns in your behavior.

Thirdly and finally, understanding implicit bias as unintentional discrimination captures why unintentional discrimination is often not in line with our explicit beliefs, but is not necessarily out of line with them either (see, e.g., Holroyd, 2016; Nier, 2005; Zheng, 2016). For an explicit and convinced sexist, it would be in line with their explicit beliefs to select Michael as the new police chief, because according to them the male candidate is the best candidate. If the discrimination is unintentional, we have two possibilities. First of all, a person could be convinced that fairness is important.

They think they should select a candidate because this person has the right skills and knowledge. If it turns out that they do not choose this way, their unintentional discrimination would be out of line with their explicit beliefs. However, a person could also be explicitly sexist, but this time decide to select on the basis of the candidates' credentials. Unbeknownst to them, however, they may display unintentional discrimination by being influenced by the name Michael. In this case, unintentional discrimination and explicit beliefs do line up (e.g., Holroyd, 2016; Nier, 2005; Zheng, 2016).

5 Conclusion

In this paper, I proposed to understand implicit bias as unintentional discrimination, and argued that this approach has several advantages. It paints a clear picture of what goes wrong when we display implicit bias: that we, given our end, respond to facts that are irrelevant and imply unfair treatment unintentionally and often unconsciously. In doing so it offers a broad perspective on implicit bias, and accounts for the different ways in which it can display itself and can be explained. Most notably, it makes room for explanations beyond implicit psychological states, and considers a kind of unconscious discrimination that is often overlooked in the field: discrimination that is the result of ignorance. Furthermore, the view can account for the central characteristics of implicit bias: (1) that it is, for a variety of reasons, difficult to control, (2) that we are not necessarily unconscious of implicit bias but not properly conscious either, and (3) that we can unintentionally discriminate regardless of whether we claim to care about fairness. A central upshot of the account is that the field of implicit bias should explicitly discuss the phenomenon they are trying to explain: the nature of persisting unfair treatment.

Acknowledgements An earlier version of this paper was given at the Moral Psychology and Rationality Workshop in 2021, the OSZW conference 2021, and at the HHU Düsseldorf. I would like to thank the audiences for their helpful questions and comments. Also, I would like to thank Olivier Burtin, Jan De Houwer, Anna Moltchanova, Annemarie Kalis, Josefa Toribio and two anonymous reviewers for their helpful feedback on the paper.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Deutsche Forschungsgemeinschaft—AS 667/1-1.

Declarations

Conflict of interest No financial interest or benefit has arisen from the direct applications of this research.

References

Anscombe, G. E. M. (1963). *Intention* (2nd ed.). Basil Blackwell.

Antony, L. M. (2016). Bias: Friend or foe? Reflections on Saulish skepticism. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology* (pp. 157–190). Oxford University Press.

Arruda, C. T., & Povinelli, D. J. (2018). Two ways of relating to (and acting for) reasons. *Mind & Language, 33*(5), 441–459. https://doi.org/10.1111/mila.12185

Asma, L. J. F. (2023). From causation to conscious control. *Philosophical Explorations, 26*(3), 420–436. https://doi.org/10.1080/13869795.2023.2223200

Asma, L. J. F. (unpublished manuscript). Unintentional discrimination and misattribution.

Ayala-López, S., & Beeghly, E. (2020). Explaining injustice. In E. Beeghly & A. Madva (Eds.), *An introduction to implicit bias: Knowledge, justice, and the social mind* (pp. 211–232). Routledge.

Baston, R., & Vosgerau, G. (2016). Implicit attitudes and implicit prejudices. *Philosophical Psychology, 29*(6), 889–903. https://doi.org/10.1080/09515089.2016.1181260

Beeghly, E., & Madva, A. (2020). *An introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*. Routledge.

Berger, J. (2020). Implicit attitudes and awareness. *Synthese, 197*, 1291–1312. https://doi.org/10.1007/s11229-018-1754-3

Brownstein, M., Mavda, A., & Gawronski, B. (2020). Understanding implicit bias: Putting the criticism into perspective. *Pacific Philosophical Quarterly, 101*(2), 276–307. https://doi.org/10.1111/papq.12302

Brownstein, M., & Saul, J. (2016). *Implicit bias and philosophy, volume1: Metaphysics and epistemology*. Oxford University Press.

Byrd, N. (2019). 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

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology, 47*, 233–279. https://doi.org/10.1016/B978-0-12-407236-7.00005-X

Davidson, D. (1963). Actions, reasons and causes. *The Journal of Philosophy, 60*(23), 685–700. https://doi.org/10.2307/2023177

Davidson, D. (1973). Freedom to act. In T. Hondrich (Ed.), *Essays on freedom of action* (pp. 63–81). Routledge & Kegan Paul Ltd.

Davidson, L. J., & Kelly, D. (2020). Minding the gap: Bias, soft structures, and the double life of social norms. *Journal of Applied Philosophy, 37*(2), 190–210. https://doi.org/10.1111/japp.12351

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science, 14*(5), 835–840. https://doi.org/10.1177/1745691619855638

De Houwer, J., & Boddez, Y. (2022). Bias in implicit measures as instances of biased behavior under suboptimal conditions in the laboratory. *Psychological Inquiry, 33*(3), 173–176. https://doi.org/10.1080/1047840X.2022.2106755

Dovidio, J. F., & Kunst, J. R. (2022). Delight in disorder: Inclusively defining and operationalizing implicit bias. *Psychological Inquiry, 33*(3), 177–180. https://doi.org/10.1080/1047840X.2022.2106756

Dovodio, J. F., Gaertner, S. E., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? Interpersonal biases and interracial distrust. *Cultural Diversity and Ethnic Minority Psychology, 8*(2), 88–102. https://doi.org/10.1037/1099-9809.8.2.88

Dutton, D. G., & Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology, 23*, 510–517. https://doi.org/10.1037/h0037031

Finkelstein, D. H. (2003). *Expression and the Inner*. Harvard University Press.

Ford, A. (2016). On what is in front of one's nose. *Philosophical Topics, 44*(1), 141–161. https://doi.org/10.5840/philtopics20164419

Gaertner, S. L., & Dovidio, J. F. (2005). Understanding and addressing contemporary racism: From aversive racism to the common ingroup identity model. *Journal of Social Issues, 61*(3), 615–639. https://doi.org/10.1111/j.1540-4560.2005.00424.x

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

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 283–310). Cambridge University Press.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition, 15*(3), 485–499. https://doi.org/10.1016/j.concog.2005.11.007

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022a). Implicit bias≠ bias on implicit measures. *Psychological Inquiry, 33*(3), 139–155. https://doi.org/10.1080/1047840X.2022.2106750

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022b). Reflections on the difference between implicit bias and bias on implicit measures. *Psychological Inquiry, 33*(3), 219–231. https://doi.org/10.1080/1047840X.2022.2115729

Goedderz, A., Azad, Z. R., & Hahn, A. (pre-print). Awareness of implicit attitudes revisited: A meta-analysis on replications across samples and settings. https://doi.org/10.31234/osf.io/frwcy

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.74.6.1464

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Hall, D. L., & Payne, B. K. (2010). Unconscious influences of attitudes and challenges to self-control. In R. Hassin, K. Ochsner, & Y. Trope (Eds.), *Self control in society, mind, and brain* (pp. 221–242). Oxford University Press.

Haslanger, S. (2015). Distinguished lecture: Social structure, narrative and explanation. *Canadian Journal of Philosophy, 45*(1), 1–15. https://doi.org/10.1080/00455091.2015.1019176

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy, 43*(3), 274–306. https://doi.org/10.1111/j.1467-9833.2012.01565.x

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., & Sweetman, J. (2006). 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.

Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology* (pp. 47–79). Oxford University Press.

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology, 108*(2), 187–218. https://doi.org/10.1037/a0038557

Kalis, A., & Ghijsen, H. (2022). Understanding implicit bias: A case for regulative dispositionalism. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2022.2046261

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *The Quarterly Journal of Experimental Psychology, 63*(3), 595–619. https://doi.org/10.1080/17470210903076826

Krickel, B. (2018). Are the states underlying implicit biases unconscious?—A Neo-Freudian answer. *Philosophical Psychology, 31*(7), 1007–1026. https://doi.org/10.1080/09515089.2018.1470323

Lauer, H. (2019). Implicit racist epistemology. *Angelaki, 24*(2), 34–47. https://doi.org/10.1080/0969725X.2019.1574076

Leboeuf, C. (2020). The embodied biased mind. In E. Beeghly & A. Madva (Eds.), *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind* (pp. 41–56). Routledge.

Levy, N. (2014). Consciousness, implicit attitudes and moral responsibility. *Noûs, 48*(1), 21–40. https://doi.org/10.1111/j.1468-0068.2011.00853.x

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.

Madva, A. (2016). A plea for anti-anti-individualism: How oversimple psychology misleads social policy. *Ergo, an Open Access Journal of Philosophy, 3*(27), 701–728. https://doi.org/10.3998/ergo.12405314.0003.027

Madva, A. & Brownstein, M. (2018). Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Noûs, 52*(3), 611–644. https://doi.org/10.1111/nous.12182

McFarland, S. G. & Crouch, Z. (2002). A cognitive skill confound on the Implicit Association Test. *Social Cognition, 20*(6). https://doi.org/10.1521/soco.20.6.483.22977
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
March, D. S., Olson, M. A., & Fazio, R. H. (2018). The implicit misattribution model of evaluative conditioning. *Social Psychological Bulletin, 13*(3), 1–25. https://doi.org/10.5964/spb.v13i3.27574
Marcus, E. (2012). *Rational causation*. Harvard University Press.
Mayr, E. (2011). *Understanding human agency*. Oxford University Press.
McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 375*, 435–442. https://doi.org/10.1006/jesp.2000.1470
Meissner, F., & Rothermund, K. (2015). The insect-nonword IAT revisited: Dissociating between evaluative associations and recoding. *Social Psychology*. https://doi.org/10.1027/1864-9335/a000220
Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297
Ngo, H. (2016). Racist habits: A phenomenological analysis of racism and the habitual body. *Philosophy & Social Criticism, 42*(9), 847–872. https://doi.org/10.1177/0191453715623320
Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A Bogus Pipeline approach. *Group Process & Intergroup Relations, 8*(1), 39–52. https://doi.org/10.1177/1368430205048615
Norman, J. B., & Chen, J. M. (2022). Grappling with social complexity when defining and assessing implicit bias. *Psychological Inquiry, 33*(3), 193–198. https://doi.org/10.1080/1047840X.2022.2106760
Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Frontiers of social psychology: Social psychology and the unconscious. The automaticity of higher mental processes* (pp. 265–292). Psychology Press.
Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*(6), 636–639. https://doi.org/10.1046/j.0956-7976.2003.psci_1477.x
Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277
Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. *Advances in Experimental Social Psychology, 62*, 1–50. https://doi.org/10.1016/bs.aesp.2020.04.001
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. https://doi.org/10.1080/1047840X.2017.1335568
Reis-Dennis, S., & Yao, V. (2021). 'I *love* women': An explicit explanation of implicit bias test results. *Synthese, 199*, 13861–13882. https://doi.org/10.1007/s11229-021-03401-3
Rosenthal, D. M. (2005). *Consciousness and mind*. Clarendon Press.
Saul, J. (2013). Scepticism and implicit bias. *Disputatio, 5*(37), 243–263.
Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379–399. https://doi.org/10.1037/h0046234
Schlosser, M. E. (2007). Basic deviance reconsidered. *Analysis, 67*(3), 186–194. https://doi.org/10.1093/analys/67.3.186
Schmader, T., Bareket-Shavit, C., & Baron, A. S. (2022). Beyond awareness: The many forms of implicit bias and its implications. *Psychological Inquiry, 33*(3), 156–161. https://doi.org/10.1080/1047840X.2022.2106752
Setiya, K. (2017). *Practical knowledge: Selected essays*. Oxford University Press.
Siegel,. (2020). Bias and perception. In E. Beeghly & A. Madva (Eds.), *An Introduction to implicit bias: Knowledge, justice, and the social mind* (pp. 99–115). Routledge.
Soon, V. (2020). Implicit bias and social schema: A transactive memory approach. *Philosophical Studies, 177*, 1857–1877. https://doi.org/10.1007/s11098-019-01288-y
Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.
Toribio, J. (2018). Implicit bias: From social structure to representational format. *Theoria, 33*(1), 41–60. https://doi.org/10.1387/theoria.17751
Toribio, J. (2021). Responsibility for implicit discrimination: A habit-based approach. *Journal of Social Philosophy, 53*(2), 239–254. https://doi.org/10.1111/josp.12442

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16*(6), 474–480. https://doi.org/10.1111/j.0956-7976.2005.01559.x

Welpinghus, A. (2020). The imagination model of implicit bias. *Philosophical Studies, 177*, 1611–1633. https://doi.org/10.1007/s11098-019-01277-1

Wood, W., Tam, L., & Witt, M. G. (2005). Changing circumstances, disrupting habits. *Journal of Personality and Social Psychology, 88*(6), 918–933. https://doi.org/10.1037/0022-3514.88.6.918

Zheng, R. (2016). Attributability, accountability and implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, volume 2: Moral Responsibility, Structural Injustice, and Ethics* (pp. 62–89). Oxford University Press.

Zheng, R. (2018). Bias, structure, and injustice: A reply to Haslanger. *Feminist Philosophy Quarterly*. https://doi.org/10.5206/fpq/2018.1.4