Implicit Bias: What Is It and How Does It Matter for Diversity, Equity, and Inclusion?

Bertram Gawronski
University of Texas at Austin

References to implicit bias are abundant in initiatives to increase diversity, equity, and inclusion (DEI). Common claims about implicit bias are that it is widespread (Greenwald et al., 2022) and pervasive (Nosek et al., 2007); that everyone has it (Staats, 2016); and that it is a major obstacle to DEI in virtually all aspects of life, including organizations (Jost et al., 2009), the legal system (Levinson & Smith, 2012), education (Staats, 2016), and medical care (Hall et al., 2015). But what exactly is implicit bias, and how does it matter for DEI? A closer look at the literature reveals that there is no straightforward answer to these questions, because (1) the term *implicit bias* has been used with different meanings and (2) the conclusions suggested by the available evidence differ depending on the meaning of the term.

To provide a basis for informed discussions about implicit bias and its significance for DEI, the current chapter discusses two dominant ideas of what constitutes implicit bias, relevant empirical evidence, and the implications of this evidence for DEI. In the first part, I discuss (1) the idea that people can behave in a biased manner without being aware that their behavior is biased, (2) two potential mechanisms that may lead to biased behavior without awareness, and (3) the significance of these mechanisms for DEI. In the second part, I discuss (1) the idea that implicit bias is what is being measured by indirect measures of bias, (2) why bias on indirect measures is different from unconscious bias, (3) what is currently known about the relation between bias on indirect measures and discriminatory behavior, (4) recent accounts that treat bias on indirect measures as an indicator of systemic (rather than individual) bias, and (5) the implications of the available evidence for DEI. In the final section, I provide an integrative discussion of (1) what we know about implicit bias, (2) important questions that still need to be addressed, and (3) implications of the available evidence for initiatives to increase DEI. I conclude with a list of recommendations for researchers, practitioners, and educators.

Implicit Bias as Unconscious Bias

A common conception of implicit bias involves the idea that people can behave in a biased manner without being aware that their behavior is biased (Gawronski et al., 2022a). Examples illustrating this idea can be found under the hashtag *#LivingWhileBlack*, which includes a long list of mundane, noncriminal activities for which police had been called on Black people (e.g., waiting for a friend at Starbucks, shopping for prom clothes; see Griggs, 2018). The critical assumptions underlying descriptions of these incidents as instances of implicit race bias are that (1) police would not have been called if the same activities had been performed by a White person and (2) people were unaware that their decision to call the police was influenced by race-related characteristics of the target person (e.g., skin color). Similar concerns have been raised about instances of implicit gender bias, in that (1) people often show different responses to a target person depending on the gender of the target and (2) people may not be aware that their responses are influenced by the target's gender. For the sake of conceptual clarity, I will use the term *unconscious bias* to refer to cases where people behave in a biased manner without being aware that their behavior is biased.

Underlying Mechanisms

The available evidence suggests two psychological mechanisms that can lead to unconscious bias: (1) biased interpretation of ambiguous information and (2) biased weighting of mixed information (see Gawronski, Ledgerwood, et al., 2020; Gawronski et al. 2022a). Biased interpretation occurs when people construe the same information about a target differently depending on the social group membership of the target. This idea resonates with the concerns expressed under the hashtag *#LivingWhileBlack*, in that the individuals who called the police construed the mundane activities of Black people as suspicious and threatening, and that they presumably would not have construed these activities in the same way if the targets had been White people. These concerns are supported by evidence of experimental studies showing that the same behavior is often interpreted differently depending on the social group membership of the target (e.g., Darley & Gross, 1983; Duncan, 1976; Gawronski et al., 2003; Kunda & Sherman-Williams, 1993; Sagar & Schofield, 1980). For example, in research on face perception, White participants have been found to perceive the same neutral facial expression as friendly when the target was White and as unfriendly when the target was Black (Bijlstra et al., 2014; Hugenberg & Bodenhausen, 2003; Hutchings & Haddock, 2008; see also Halberstadt et al., 2018). Moreover, consistent with the hypothesis that biased interpretations can occur outside of awareness (Fazio & Olson, 2014; Trope, 1986), effects of social group membership on the interpretation of ambiguous target information have been found even when participants were motivated and able to respond in an unbiased manner (Gawronski et al., 2003).

To appear in: V. M. Esses, J. F. Dovidio, J. Jetten, D. Sekaquaptewa, & K. West (Eds.). *The Sage Handbook of Psychological Perspectives on Diversity, Equity, and Inclusion.* Thousand Oaks, CA: Sage.

Biased weighting occurs when people weigh the same information about a target differently depending on the social group membership of the target. An illustrative example is the biased weighting of credentials in hiring decisions. For example, in a hiring scenario involving a highly qualified man with superior credentials in terms of a Criterion A and highly qualified woman with superior credentials in terms of another Criterion B, decision-makers may give more weight to Criterion A than Criterion B, leading them to hire the man and not the woman. Yet, in a scenario where the credentials of the two candidates are reversed, the decision-makers may give more weight to Criterion B than Criterion A, thus leading them to hire the man regardless of who is superior in terms of the two criteria (e.g., Hodson et al., 2002; Norton et al., 2004; Uhlmann & Cohen, 2005; see also Régner et al., 2019). In both cases, the decision-makers may justify their preference with whatever qualification makes the man superior to the woman, suggesting that they weighed the candidates' credentials in a manner that merely served to rationalize a pre-existing preference instead of generating a preference based on the candidates' credentials. Some studies suggest that self-perceptions of objectivity in such cases are associated with greater (rather than smaller) bias (Uhlmann & Cohen, 2005). While this finding could be due to self-presentational concerns, it is consistent with the idea that differential weighting may bias decisions outside of awareness.

Significance for DEI

The significance of unconscious bias arising from biased interpretation and biased weighting is clear and straightforward. Potential examples of their impact are abundant. In policing, ambiguous actions may be more likely to be interpreted as threatening when the target is Black rather than White; in hiring and promotion, mixed credentials may be weighed in a manner that favors members of dominant over members of non-dominant groups; in medical decision-making, diagnoses based on ambiguous symptoms may contribute to health disparities via different treatment recommendations for members of different groups; and in legal decision-making, the same illegal activity may be perceived differently for members of different social groups, leading to different sentencing decisions by judges and juries. The notion of unconscious bias suggests that decisions in these cases may be biased without the decision-makers being aware that their decisions are influenced by the social group membership of the target. While decision-makers may be convinced that their decisions are based on objective facts, they may not realize that their subjective perception of these facts is biased by the social group membership of the target. Needless to say, DEI will be difficult to achieve as long as members of historically disadvantaged groups remain the target of biased decisions, and discrimination is arguably more difficult to combat when people are not aware of their biased decisions. If ambiguous actions are more likely to be interpreted negatively when they are performed by members of historically disadvantaged groups and if mixed information about members of historically disadvantaged groups is more likely to be weighed in an unfavorable manner, diversity will remain low, inequities will remain common, and members of historically disadvantaged groups will continue to feel excluded.

Implicit Bias as Bias on Indirect Measures

Another common conception equates implicit bias with what is being measured by a particular type of indirect measures (Greenwald & Banaji, 2017), the most prominent examples being the Implicit Association Test (IAT; Greenwald et al., 2021), the Evaluative Priming Task (EPT; Fazio et al., 1995), and the Affect Misattribution Procedure (AMP; Payne et al., 2005). For the sake of conceptual clarity, I will use the term *bias on indirect measures* to refer to bias captured by indirect measures such as the IAT (for reviews of indirect measures, see Gawronski & De Houwer, 2014; Greenwald & Lai, 2020).

Unconscious vs. Unintentional Bias

There is considerable confusion about whether bias on indirect measures is unconscious. This confusion is at least partly due to seemingly contradictory statements by the inventors of the IAT. While some of their publications explicitly state that indirectly measured bias is not the same as unconscious bias (Greenwald & Banaji, 2017), other publications include claims that bias on the IAT operates outside conscious awareness (Greenwald et al., 2022; Morehouse & Banaji, 2024) and that the IAT uncovers hidden biases that people do not know they have (Banaji & Greenwald, 2016). The latter claims echo the authors' original conceptualization of implicit constructs as "introspectively unidentified (or inaccurately identified) trace of past experience that mediates [responses]" (Greenwald & Banaji, 1995, p. 5). However, equating *bias on indirect measures* with *unconscious bias* is problematic on conceptual and empirical grounds (see Gawronski et al., 2022a).

First, people are typically aware that their responses on indirect measures are influenced by the social group membership of the target individuals used as stimuli in these tasks. For example, when participants complete a race IAT, they are typically aware that their responses are slower and that they make more errors in the bias-incongruent block compared to the bias-congruent block (Monteith et al., 2001). Similar findings have been obtained with other indirect measures (Hughes et al., 2023; Kurdi et al., in press). These findings conflict with the notion that people behave in a biased manner without being aware that their behavior is biased, as discussed in the first part of this chapter.

Second, counter to the idea that indirect measures capture biases that people do not know they have (Banaji & Greenwald, 2016), people are highly accurate in predicting their biases on indirect measures (e.g., Hahn et al., 2014; Hahn & Gawronski, 2019; Morris & Kurdi, 2023; Rahmani Azad et al., 2023). For example, when participants were asked to predict their scores on multiple IATs involving different social groups before completing these IATs, participants showed high accuracy in predicting their IAT scores regardless of their prior experience with the IAT, regardless of how much information they received about the IAT in the instructions for the prediction task, and regardless of whether the IAT was introduced as a measure of true beliefs or cultural associations (Hahn et al., 2014). These findings conflict with the idea that indirect measures capture biases that people do not know they have.

Third, counter to the idea that surprise reactions in response to feedback about one's biases on indirect measures indicate unawareness (Banaji, 2011; Krickel, 2018; Ratliff & Smith, 2022), such surprise reactions can be explained as the product of statistical distortions in the calculation of numeric measurement scores (Wolsiefer et al., 2017) and arbitrary conventions in the verbal description of these scores (Gawronski et al., 2022b). These issues undermine interpretations of surprise reactions to bias feedback as evidence for unawareness, reconciling the apparent conflict with findings that people are highly accurate in predicting their biases on indirect measures (e.g., Hahn et al., 2014; Hahn & Gawronski, 2019; Morris & Kurdi, 2023; Rahmani Azad et al., 2023).

Fourth, meta-analytic evidence suggests that bias on indirect measures is not uniquely related to biased behavior that occurs outside of awareness, in that associations between bias on indirect measures and discriminatory behavior do not differ depending on whether the focal behaviors do or do not involve awareness (Kurdi et al., 2019). Hence, there is nothing a priori about a person's bias on an indirect measure that would justify claims that this person engages in biased behavior without being aware that their behavior is biased (because the person could be perfectly aware of their biased behavior). This conclusion conflicts with the idea that bias on indirect measures could be treated as an indicator for the kind of unconscious biases discussed in the first part of this chapter.

If indirect measures do not capture unconscious bias, what do they measure? Despite disagreements on specific details, there is growing consensus that indirect measures capture biased behavior that is expressed without intention (De Houwer & Boddez, 2022; Gawronski et al., 2022b; Melnikoff & Kurdi, 2022; Ratliff & Smith, 2022; see also De Houwer, 2019). Yet, unintentional bias is not the same as unconscious bias, because people may be aware that their behavior toward a target is biased by the target's social group membership even when they do not intend to behave in a biased manner (Gawronski et al., 2022b). These considerations raise important questions about the significance of unintentional bias on indirect measures for DEI. Different from the reviewed instances of unconscious bias arising from biased interpretation and biased weighting, there is no real-world counterpart to unintentional biases in the categorization of stimuli on indirect measures. Hence, the relevance of unintentional biases on indirect measures for DEI has to be evaluated based on their functional properties, the most significant being their predictive relation to discriminatory behavior. The basic idea is that bias on indirect measures and discriminatory behavior are rooted in the same underlying mental representations (e.g., automatically activated associations), and that indirect measures provide a tool to capture these representations in a manner that reduces intentional influences (Fazio et al., in press).

Prediction of Behavior

Several independent meta-analyses suggest that predictive relations between bias on indirect measures and discriminatory behavior are modest at best, with meta-analytic correlations ranging between .14 and .28 (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013). Extant dual-process theories (e.g., Fazio & Olson, 2014; Strack & Deutsch, 2004) suggest that these average correlations may conceal more complex patterns, in that bias on indirect measures should be predictive of spontaneous (but not deliberate) behavior, behavior under suboptimal (but not optimal) processing conditions, and behavior of individuals with a disposition to engage in superficial (but not elaborate) processing (for a review, see Friese et al., 2008). While the available evidence for these predictions is mixed (Greenwald et al., 2022; but see Gawronski, 2019), two properties of bias on indirect measures raise questions about the extent to which bias on indirect measures may show meaningful predictive relations with discriminatory behavior. First, different from the high temporal stability of bias on traditional self-report measures, bias on indirect measures has been found to be highly unstable over time (see Gawronski et al., 2017). Second, bias on indirect measures has been found to be highly context-sensitive, in that even minor changes in a person's social context can influence that person's level of bias on an indirect measure (see Gawronski & Sritharan, 2010). Together, the two aspects suggest that a person's level of bias measured with an indirect measure at one time point provides little information about that person's level of bias at a different time point, especially when the social contexts at the two time points are different (see Gschwendner et al., 2008).

To appear in: V. M. Esses, J. F. Dovidio, J. Jetten, D. Sekaquaptewa, & K. West (Eds.). *The Sage Handbook of Psychological Perspectives on Diversity, Equity, and Inclusion.* Thousand Oaks, CA: Sage.

While these issues undermine predictive relations between bias on indirect measures and discriminatory behavior over time and across contexts (Gawronski, 2019), they do not necessarily question the presumed behavioral impact of the mental representations underlying bias on indirect measures. After all, it seems possible that the mental representations underlying bias on implicit measures influence discriminatory behavior in the moment even when these representations fluctuate from one moment to the other as a result of changes in the context. If that were the case, contextually induced changes in bias on indirect measures should be associated with concurrent changes in discriminatory behavior. However, a meta-analysis on this question found no evidence for this assumption: there is no evidence for the idea that changing bias on indirect measures would lead to concurrent changes in discriminator behavior (Forscher et al., 2019). Together, these findings pose a challenge to the idea that bias on indirect measures provides insights into the underpinnings of discriminatory behavior.

Some have argued that the low temporal stability of bias on indirect measures is the product of measurement error, and that measurement error could potentially be reduced by aggregating data from multiple measurements (e.g., by asking participants to complete the same IAT multiple times and averaging the scores from all measurements; see Greenwald et al., 2021). While the available evidence on the effectiveness of this approach is mixed (Carpenter et al., 2023; Hannay & Payne, 2022), some studies found that aggregating multiple IAT scores from the same person can indeed improve the identification of temporally stable biases at the trait level (Carpenter et al., 2023). However, another notable finding of these studies is that aggregating multiple IAT scores from the same person substantially increased the overlap between bias on the IAT and bias on traditional self-report measures. Because large overlap between the outcomes of direct and indirect measures undermines the basis for the distinction between explicit and implicit bias, these findings pose a challenge to the idea that bias on indirect measures represents a unique obstacle to DEI that is distinct from bias on direct measures.

Individual vs. Systemic Bias

The weak associations between bias on indirect measures and discriminatory behavior at the individual level may seem to conflict with evidence for rather strong associations between bias on indirect measures and social disparities at the regional level (for a review, see Calanchini et al., 2020). Examples of the latter include associations between aggregate scores of racial bias on indirect measures at the regional level with use of lethal force by police officers against African Americans (Hehman et al., 2018), racial disparities in traffic stops by police (Ekstrom et al., 2022), and mortality rates among African Americans (Leitner et al., 2016) at the same regional level. Together with the available evidence for low temporal stability and high context sensitivity of bias on indirect measures, these findings have led some researchers to suggest that bias on indirect measures reflects bias at the systemic level rather than bias at the individual level (Payne & Hannay, 2021). The basic idea underlying this account is that bias on indirect measures may not be a causal force that leads to discriminatory behavior, but instead reflects momentary thoughts elicited by a person's environment (Payne et al., 2017). Such environmental influences may involve proximal factors such as family members (e.g., Castelli et al., 2009), one's community (Vuletich & Payne, 2019), and media portrayals (Weisbuch et al., 2009), but also distal factors such as historical inequalities (Payne et al., 2019).

The idea that bias on indirect measures reflects systemic biases reconciles three sets of paradoxical findings in the literature on indirect measures. First, it explains how bias on indirect measures can be widespread and robust on average (Nosek et al., 2007), yet highly unstable over time at the individual level (Gawronski et al., 2017). Second, it explains how bias on indirect measures can be highly stable across age starting from early childhood (Dunham et al., 2008) despite being highly unstable over just a few weeks (Gawronski et al., 2017). Third, it explains why regional differences in bias on indirect measures show strong associations with societal disparities (Calanchini et al., 2022), although bias on indirect measures shows rather weak associations with discriminatory behavior at the individual level (Cameron et al., 2012; Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013). Robust average levels of bias over time and across age groups are assumed to reflect the persistence of societal disparities, while short-term fluctuations at the individual level reflect incidental aspects of a person's momentary context. Moreover, strong associations between societal disparities and regional levels of bias on indirect measures are assumed to reflect the elicitation of bias-related thoughts by biased environments, while associations between bias on indirect measures and discriminatory behavior at the individual level are regarded as spurious.

Accounts that treat bias on indirect measures as indicators of systemic bias are important for DEI initiatives, because they turn the dominant narrative about indirect measures on its head. A common assumption in the literature on indirect measures is that, to increase DEI, researchers have to develop interventions that effectively reduce bias on indirect measures at the individual level (e.g., Lai et al., 2014, 2016). The idea underlying this assumption is that such interventions will promote DEI by reducing discriminatory behavior. Systemic accounts suggest that

such interventions are destined to fail, because bias on indirect measures is merely a reflection of bias at the systemic level, with bias on indirect measures not being causally involved in producing discriminatory behavior (Payne et al., 2018). According to this view, DEI initiatives require interventions that effectively reduce bias at the systemic level. To the extent that these interventions are effective, bias on indirect measures will show parallel effects (Sawyer & Gampa, 2023). However, these effects are mere reflections of the reduced levels of bias at the systemic level; they are not causally involved in bringing about the observed changes at the societal level. These assumptions have important implications for the presumed significance of bias on indirect measures for initiatives to promote DEI.

Significance for DEI

Different from the rather obvious significance of unconscious bias, the available evidence poses a challenge to the dominant narrative about the significance of unintentional bias on indirect measures. After more than a quarter century of research using indirect measures (Gawronski, De Houwer, et al., 2020), there is no solid evidence for the idea that bias on indirect measures poses a meaningful obstacle to DEI. Associations with discriminatory behavior are modest at best; a person's bias on an indirect measure at one time point provides little information about that person's bias at later time points and in other contexts; and there is no evidence that changes in bias on indirect measures lead to corresponding changes in discriminatory behavior. Systemic accounts explain these findings by assuming that bias on indirect measures reflects momentary thoughts elicited by a person's environment. However, according to these accounts, bias on indirect measures is a mere reflection of bias at the systemic level that does not itself contribute to societal disparities via discriminatory behavior.

While systemic accounts suggest that indirect measures could still be valuable as indicators of biases at the regional level, an important caveat is that virtual all findings involving bias at regional levels replicate on both indirect and direct measures of bias (Calanchini et al., 2022), with correlations between the two measures reaching levels as high as $r = .85$ (Hehman et al., 2019). There is virtually no evidence for dissociations between indirect and direct measures at the regional level, rendering the significance of the distinction obsolete. These findings cast further doubts about the unique significance of research with the IAT and other indirect measures for DEI initiatives.

Open Questions

The available evidence suggests that, while unconscious bias likely has important implications for DEI, the widely proclaimed significance of bias on indirect measures seems questionable. Although the discussion in the first part of this chapter included various examples of how people may behave in a biased manner without being aware that their behavior is biased, I deliberately used the qualifier "likely" with reference to the presumed significance of unconscious bias because we still know very little about this important phenomenon. Somewhat ironically, this lack of knowledge is primarily due to the rise of indirect measures such as the IAT and the mistaken assumption that these measures capture unconscious bias (Gawronski et al., 2022a). Although this assumption has been disputed since the early days of indirect measures (Fazio & Olson, 2003; Gawronski et al., 2007), it gave rise to the mistaken idea that unconscious bias could be studied by having participants complete an IAT or other indirect measures of bias. Because administering an indirect measure is much easier than studying unconscious effects of a target's social group membership, researchers interested in unconscious bias have devoted most of their resources to studies with indirect measures, thereby ignoring the actual phenomenon of unconscious bias. Thus, as a first step to gaining a better understanding of unconscious bias, it seems prudent to reallocate resources from studies on bias on indirect measures to studies that investigate actual instances of unconscious bias, which can be formally defined as unconscious effects of social category cues on behavioral responses (Gawronski et al., 2022a, 2022b).

An important aspect in this research will arguably be the mechanisms underlying unconscious bias. While there is considerable evidence for biased interpretation (e.g., Darley & Gross, 1983; Duncan, 1976; Gawronski et al., 2003; Hugenberg & Bodenhausen, 2003; Kunda & Sherman-Williams, 1993; Sagar & Schofield, 1980) and biased weighting (e.g., Hodson et al., 2002; Norton et al., 2004; Uhlmann & Cohen, 2005), the majority of these studies have been conducted decades ago and there is barely any research on the question of whether the two mechanisms influence behavior outside of awareness (for a review, see Gawronski & Corneille, in press). Although establishing unawareness can be a methodologically difficult endeavor, evidence for unawareness seems critical if one wants to corroborate the presumed significance of unconscious (as opposed to conscious) bias for DEI. Thus, an important task for future research is to gain a better understanding of unconscious bias by investigating the presumed unawareness of the effects of biased interpretation and biased weighting.

Expanding on this work, three important questions are: (1) How prevalent is unconscious bias? (2) What are the boundary conditions of unconscious bias? (3) What can be done to reduce unconscious bias? Although implicit bias has been claimed to be widespread (e.g., Greenwald et al., 2022) and pervasive (e.g., Nosek et al.,

2007), such claims are based on research with indirect measures, which provides no information regarding the prevalence of unconscious bias. Future research on this question would also benefit from distinguishing between dispersed and concentrated discrimination (Campbell & Brauer, 2021). According to the notion of dispersed discrimination, experiences of social discrimination come from interactions with a large number of individuals who behave in slightly biased ways. In contrast, according to the notion of concentrated discrimination, experiences of social discrimination come from interactions with a small number of individuals who behave in strongly biased ways. While claims such as "everyone has implicit biases" are common among practitioners (e.g., Staats, 2016, p. 30), there is no evidence to date on the presumed pervasiveness of unconscious bias, and whether unconscious bias involves dispersed or concentrated patterns of discriminations.

Regarding the boundary conditions of unconscious bias, there is evidence that biased interpretation is more common for ambiguous information (e.g., ambiguous facial expressions shown by a Black vs. White person), and that biased weighting is more common for mixed information (e.g., mixed credentials of male vs. female job applicants). Otherwise, extant knowledge about their boundary conditions is very limited. At this point, there is also very little evidence on what could be done to reduce unconscious bias arising from biased interpretation and biased weighting. Evidence on the effectiveness of interventions in reducing bias on indirect measures (e.g., Lai et al., 2014, 2016) provides no information on this question, because unconscious bias is different from unintentional bias on indirect measures (Gawronski et al., 2022b). This conclusion echoes broader concerns about the ineffectiveness of implicit-bias trainings to increase DEI (e.g., Carter et al., 2020; Greenwald et al., 2022; Kim & Roberson, 2022; Onyeador et al., 2021). Although the demand for diversity trainings in public and private organizations gave rise to a multibillion-dollar industry, the available evidence suggests that these investments had little to no impact (Dobbin & Kalev, 2016). Applied to current question, these findings raise the question of what could be done combat effects of biased interpretation and biased weighting.

One potential strategy to eliminate effects of biased interpretation and biased weighting is to make decision-makers "blind" about the social group membership of the target(s) of their decisions (Gawronski, Ledgerwood, et al., 2020). While this approach may work well for some decision contexts (e.g., blind peer review of scientific manuscripts), it is not applicable to the majority of contexts where the two mechanisms may bias decisions. Another obvious strategy would be to change the mental representations that give rise to biased interpretations and biased weighting. However, the literature on attitude change suggests that this may be easier said than done, in that the effectiveness of interventions to change attitudes depends on numerous contextual factors (Albarracín & Shavit, 2018). Based on these considerations, some scholars in this area suggested that it may be more effective to focus on interventions that aim to change discriminatory behavior via structural aspects of decision contexts rather than underlying mental representations (Brauer, 2024; Onyeador et al., 2021).

Regarding effects of biased weighting in hiring decisions, one example is to identify unambiguous selection criteria prior to the review of application materials. Although there is evidence supporting the effectiveness of this approach (Uhlmann & Cohen, 2005), it might be less effective when the criteria-relevant credentials are interpreted in a biased manner. Another limitation is that the applicability of this approach seems limited to contexts that involve evaluations of credentials. An alternative approach with broader applicability would be to educate people about the effects of biased interpretation and biased weighting, and about how the two mechanisms can lead to biased decisions (Gawronski, Ledgerwood, et al., 2020). While such educational interventions would only convey knowledge about the two mechanisms without granting conscious access to their operation in the moment (Wilson & Brekke, 1994), the relevant knowledge would provide a basis for counterfactual reasoning about whether the same information might be interpreted or weighted differently if the target(s) belonged to a different social group (see Hirt & Markman, 1995; Lord et al., 1984). Although the broader literature on bias correction suggests that such educational interventions could be quite effective (Strack & Hannover, 1996; Wegener & Petty, 1997), there is no research to date that has tested their effectiveness in reducing effects of biased interpretation and biased weighting in DEI-related contexts. Past research on bias correction also suggests that knowledge of the two mechanisms may be insufficient without a corresponding motivation to make unbiased decisions (Wegener & Petty, 1997). Yet, relevant evidence for these ideas is still lacking. Thus, future research tackling these important questions will be critical for the development of effective interventions to reduce unconscious bias and, by extension, increase DEI.

Conclusions

The current chapter started with the question of how implicit bias matters for DEI. The answer to this question is: it depends on what is meant with *implicit bias*. On the one hand, it is conceivable that people can behave in a biased manner without being aware that their behavior biased. Such instances of unconscious bias may arise from biased interpretations of ambiguous information

and biased weighting of mixed information, which can contribute to discrimination in policing, hiring and promotion, medical decision-making, and legal sentencing. However, evidence for unawareness in the relevant studies is scarce and the boundary conditions and properties of unconscious bias are largely unknown—primarily due to a lack of research on these questions. On the other hand, the enormous body of research with indirect measures raises doubts about whether unintentional bias on indirect measures has any unique significance for understanding discriminatory behavior. Thus, while unconscious bias is an understudied but potentially significant obstacle to DEI, the widely presumed relevance of bias on indirect measures seems questionable, if there is any at all.

In moving forward, researchers, practitioners, and educators might consider the following list of five recommendations based on the current analysis:

1) Unconscious bias should not be equated with bias on indirect measures, and vice versa.
2) Unconscious bias should be explained with the known mechanisms of biased interpretation and bias weighting, and the significance of the two mechanisms for discrimination in real-world contexts.
3) To avoid confusion about the difference between unconscious bias and bias on indirect measures, references to the IAT and other indirect measures should be avoided when explaining unconscious bias.
4) Interventions that aim to increase DEI by tackling unconscious bias should be based on scientific evidence that directly speaks to unconscious bias.
5) To obtain a solid empirical basis for the development of such interventions, researchers should reallocate resources from studying bias on indirect measures to studying actual instances of unconscious bias.

I hope that this list of recommendations (and the analysis it is based on) provides a basis to effectively tackle implicit bias as an obstacle to DEI.

References

Albarracín, D., & Shavitt, S. (2018). Attitudes and attitude change. *Annual Review of Psychology*, *69*, 299-327.

Banaji, M. R. (2011). A vehicle for large-scale education about the human mind. In J. Brockman (Ed.), *How is the internet changing the way you think?* (pp. 392-395). New York: Harper Collins.

Banaji, M. R., & Greenwald, A. G. (2016). *Blindspot: Hidden biases of good people*. New York: Bantam.

Bijlstra, G., Holland, R. W., Dotsch, R., Hugenberg, K., & Wigboldus, D. H. J. (2014). Stereotype associations and emotion recognition. *Personality and Social Psychology Bulletin, 40,* 567-577.

Brauer, M. (2024). Stuck on intergroup attitudes: The need to shift gears to change intergroup behaviors. *Perspectives on Psychological Science*, *19*, 280-294.

Calanchini, J., Hehman, E., Ebert, T., Esposito, E., Simon, D., & Wilson, L. (2022). Regional intergroup bias. *Advances in Experimental Social Psychology, 66*, 281-337.

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

Campbell, M. R., & Brauer, M. (2021). Is discrimination widespread? Testing assumptions about bias on a university campus. *Journal of Experimental Psychology: General, 150*, 756-777.

Carpenter, T. P., Goedderz, A., & Lai, C. K. (2023). Individual differences in implicit bias can be measured reliably by administering the same Implicit Association Test multiple times. *Personality and Social Psychology Bulletin*, *49*, 1363-1378.

Carter, E. R., Onyeador, I. N., & Lewis Jr, N. A. (2020). Developing and delivering effective anti-bias training: Challenges and recommendations. *Behavioral Science & Policy, 6*, 57-70.

Castelli, L., Zogmaister, C., & Tomelleri, S. (2009). The transmission of racial attitudes within the family. *Developmental Psychology, 45*, 586–591.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20-33.

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science, 14*, 835-840.

De Houwer, J., & Boddez, Y. (2022). Bias in implicit measures as instances of biased behavior under suboptimal conditions in the laboratory. *Psychological Inquiry*, *33*, 173-176.

Dobbin, F., & Kalev, A. (2016). Why diversity programs fail. *Harvard Business Review*, *94*, 52-59.

Duncan, B. L. (1976). Differential perception and attribution of intergroup violence: Testing the lower limits of stereotyping of Blacks. *Journal of Personality and Social Psychology, 34,* 590-598.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences, 12,* 248-253.

Ekstrom, P. D., Le Forestier, J. M., & Lai, C. K. (2022). Racial demographics explain the link between racial disparities in traffic stops and county-level racial attitudes. *Psychological Science, 33,* 497–509.

Fazio, R. H., Granados Samayoa, J. A., Boggs, S. T., & Ladanyi, J. (in press). Implicit bias: What is it? In

J.A. Krosnick, T. H. Stark, & A.L. Scott (Eds.), *The Cambridge Handbook of Implicit Bias and Racism*. Cambridge, England: Cambridge University Press.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013-1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297-327.

Fazio, R. H., & Olson, M. A. (2014). The MODE model: Attitude-behavior processes as a function of motivation and opportunity. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), Dual-process theories of the social mind (pp. 155-171). New York: Guilford.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*, 522-559.

Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behavior? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology, 19*, 285-338.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14,* 574-595.

Gawronski, B., & Corneille, O. (in press). Unawareness of attitudes, their environmental causes, and their behavioral effects. *Annual Review of Psychology*.

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* (2nd edition, pp. 283-310). New York: Cambridge University Press.

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition, 38,* s1-s25.

Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology, 33,* 573-589.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2*, 181-193.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2020). Implicit bias and anti-discrimination policy. *Policy Insights from the Behavioral and Brain Sciences, 7,* 99-106.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022a). Implicit bias ≠ bias on implicit measures. *Psychological Inquiry*, *33*, 139-155.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022b). Reflections on the difference between implicit bias and bias on implicit measures. *Psychological Inquiry*, *33*, 219-231.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43,* 300-312.

Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216-240). New York: Guilford Press.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4-27.

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, *72*, 861–871.

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., ... & Wiers, R. W. (2021). Best research practices for using the Implicit Association Test. *Behavior Research Methods, 54*, 1161-1180.

Greenwald, A. G., Dasgupta, N., Dovidio, J. F., Kang, J., Moss-Racusin, C. A., & Teachman, B. A. (2022). Implicit-bias remedies: Treating discriminatory bias as a public-health problem. *Psychological Science in the Public Interest*, *23*, 7-40.

Greenwald, A G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology, 71,* 419-445.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17-41.

Griggs, B. (2018, December 28). Living while black: Here are all the routine activities for which police were called on African-Americans this year. *CNN*. Retrieved from https://www.cnn.com/2018/12/20/us/living-while-black-police-calls-trnd/index.html (January 16, 2024).

Gschwendner, T., Hofmann, W., & Schmitt, M. (2008). Differential stability: The effects of acute and chronic construct accessibility on the temporal stability of the Implicit Association Test. *Journal of Individual Differences, 29,* 70-79.

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116,* 769-794.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*, 1369.

Halberstadt, A. G., Castro, V. L., Chu, Q., Lozada, F. T., & Sims, C. M. (2018). Preservice teachers' racialized emotion recognition, anger bias, and hostility attributions. *Contemporary Educational Psychology*, *54*, 125-138.

Hall, W. J., Chapman, M. V., Lee, K. M., Merino, Y. M., Thomas, T. W., Payne, B. K., ... & Coyne-Beasley, T. (2015). Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *American Journal of Public Health*, *105*, e60-e76.

Hannay, J. W., & Payne, B. K. (2022). Effects of aggregation on implicit bias measurement. *Journal of Experimental Social Psychology*, *101*, Article 104331.

Hehman, E., Calanchini, J., Flake, J. K., & Leitner, J. B. (2019). Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of Experimental Psychology: General, 148*, 1022-1040.

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, *9*, 393-401

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69,* 1069-1086.

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin, 28,* 460-471.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14,* 640-643.

Hughes, S., Cummins, J., & Hussey, I. (2023). Effects on the Affect Misattribution Procedure are strongly moderated by influence awareness. *Behavior Research Methods*, *55*, 1558-1586.

Hutchings, P. B., & Haddock, G. (2008). Looking black in anger: The role of implicit prejudice in the categorization and perceived emotional intensity of racially ambiguous faces. *Journal of Experimental Social Psychology, 44,* 1418-1420.

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior*, *29*, 39-69.

Kim, J. Y., & Roberson, L. (2022). I'm biased and so are you. What should organizations do? A review of organizational implicit-bias training programs. *Consulting Psychology Journal, 74*, 19–39.

Krickel, B. (2018). Are the states underlying implicit biases unconscious? A neo-Freudian answer. *Philosophical Psychology, 31,* 1007-1026.

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin, 19,* 90-99.

Kurdi, B., Melnikoff, D. E., Hannay, J. W., Korkmaz, A., Lee, K. M., Ritchie, E., ... & Ferguson, M. J. (in press). Testing the automaticity features of the Affect Misattribution Procedure: The roles of awareness and intentionality. *Behavior Research Methods*.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 74*, 569-586.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., … & Nosek, B. A. (2014). A comparative investigation of 18 interventions to reduce implicit racial preferences. *Journal of Experimental Psychology: General, 143,* 1765-1785.

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

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Racial bias is associated with ingroup death rate for Blacks and Whites: Insights from Project Implicit. *Social Science & Medicine, 170,* 220-227.

Levinson, J. D., & Smith, R. J. (Eds.). (2012). *Implicit racial bias across the law*. Cambridge, MA: Cambridge University Press.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Consider the opposite: A corrective strategy for social judgments. *Journal of Personality and Social Psychology, 47,* 1231-1243.

Melnikoff, D. E., & Kurdi, B. (2022). What implicit measures of bias can do. *Psychological Inquiry*, *33*, 185-192.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*, 395-417.

Morehouse, K. N., & Banaji, M. R. (2024). The science of implicit race bias: Evidence from the Implicit Association Test. *Daedalus, 153,* 21-50.

Morris, A., & Kurdi, B. (2023). Awareness of implicit attitudes: Large-scale investigations of mechanism and scope. *Journal of Experimental Psychology: General, 152*, 3311–3343.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology, 87,* 817-831.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36-88.

Onyeador, I. N., Hudson, S. K. T., & Lewis Jr, N. A. (2021). Moving beyond implicit bias training: Policy insights for increasing organizational diversity. *Policy Insights from the Behavioral and Brain Sciences, 8,* 19-26.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*, 171-192.

Payne, B. K., Cheng, S. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277-293.

Payne, B. K., & Hannay, J. W. (2021). Implicit bias reflects systemic racism. *Trends in Cognitive Sciences, 25,* 927-936.

Payne, B. K., Niemi, L., & Doris, J. M. (2018). How to think about "implicit bias". *Scientific American*. Retrieved from https://www.scientificamerican.com/article/how-to-think-about-implicit-bias/ (January 26, 2021).

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences*, *116*, 11693-11698.

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,* 233-248.

Rahmani Azad, Z., Goedderz, A., & Hahn, A. (2023). Self-awareness and stereotypes: Accurate prediction of implicit gender stereotyping. *Personality and Social Psychology Bulletin*, *49*, 1695-1708.

Ratliff, K. A., & Smith, C. T. (2022). Implicit bias as automatic behavior. *Psychological Inquiry*, *33*, 213-218.

Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., & Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour, 3,* 1171-1179.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39,* 590-598.

Sawyer, J. E., & Gampa, A. (2023). Social movements as parsimonious explanations for implicit and explicit attitude change. *Personality and Social Psychology Review*, *27*, 28-51.

Staats, C. (2016). Understanding implicit bias: What educators should know. *American Educator*, *39*, 29-33.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8,* 220-247.

Strack, F., & Hannover, B. (1996). Awareness of the influence as a precondition for implementing correctional goals. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 579-596). New York: Guilford Press.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93,* 239-257.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16,* 474-480.

Vuletich, H. A., & Payne, B. K. (2019). Stability and change in implicit bias. *Psychological Science*, *30*, 854-862.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. *Advances in Experimental Social Psychology, 29,* 141-208.

Weisbuch, M., Pauker, K., & Ambady, N. (2009). The subtle transmission of race bias via televised nonverbal behavior. *Science*, *326*, 1711-1714.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116,* 117-142.

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*, 1193-1209.