Article

Implicit Bias and Antidiscrimination Policy

Policy Insights from the Behavioral and Brain Sciences
2020, Vol. 7(2) 99–106

DOI: 10.1177/2372732220939128
journals.sagepub.com/home/bbs
SAGE

Bertram Gawronski[1], Alison Ledgerwood[2], and Paul W. Eastwick[2]

Abstract

The science behind implicit bias tests (e.g., Implicit Association Test) has become the target of increased criticism. However, policymakers seeking to combat discrimination care about reducing bias in people's actual behaviors, not about changing a person's score on an implicit bias test. In line with this argument, we postulate that scientific controversies about implicit bias tests are irrelevant for antidiscrimination policy, which should instead focus on implicit bias in actual discriminatory behavior that occurs outside of awareness (in addition to instances of explicit bias). Two well-documented mechanisms can lead to implicit bias in actual discriminatory behavior: biased weighting and biased interpretation of information about members of particular social groups. The policy relevance of the two mechanisms is illustrated with their impact on hiring and promotion decisions, jury selection, and policing. Implications for education and bias intervention are discussed.

Keywords

bias intervention, biased information processing, discrimination, implicit bias, policy

Tweet

The significance of #implicitbias for #policy is rooted in #unconscious mechanisms leading to actual #discrimination; this significance remains unqualified by scientific controversies about #implicitbiastests; #IAT, #policy, #stereotype, #racism, #gender

Key Points

- The science behind implicit bias tests has become the target of increased criticism.
- Policymakers seeking to combat discrimination care about reducing bias in people's actual behaviors, not about changing a person's score on an implicit bias test.
- Instead of equating implicit bias with responses on implicit bias tests, implicit bias should be conceptualized in terms of actual discriminatory behavior.
- A behavioral conceptualization reveals two mechanisms that can lead to discriminatory behavior outside of awareness: biased weighting and biased interpretation of information.
- A psychological analysis of the two mechanisms leads to evidence-based recommendations for policy, education, and bias intervention.
- The recommendations can be included in extant diversity trainings and implemented into organizational decision-making procedures with little or no extra costs.

Introduction

On April 12, 2018, two African American men asked to use the restroom at a Starbucks in Philadelphia. A barista told them that the bathrooms were for customers only (Park, 2018). When the two men were asked to leave the premises after they occupied a table without making a purchase, they declined to leave, saying they were waiting for an acquaintance. In response, the store manager called the police, who escorted the two men out of the coffee shop. When a video of the incident taken by a customer went viral on social media, Starbucks apologized and closed all of its brand-operated stores for half a day to provide mandatory implicit bias training for its 175,000 employees (Chapell, 2018).

In line with Starbucks's response to the described incident, an increasing number of public institutions and private corporations offer or require implicit bias training for their employees. Yet, at the same time, the science behind implicit bias tests has become the target of increased criticism. This criticism is based on research suggesting that (a) relations between people's responses on implicit bias tests (e.g., Implicit Association Test; Greenwald et al., 1998) and actual discriminatory behavior (e.g., biased hiring decisions) are

[1]The University of Texas at Austin, USA
[2]University of California, Davis, USA

Corresponding Author:
Bertram Gawronski, The University of Texas at Austin, 108 E Dean Keeton A8000, Austin, TX 78712, USA.
Email: gawronski@utexas.edu

rather weak (Oswald et al., 2013), (b) many lab-based interventions influence responses on implicit bias tests without affecting discriminatory behavior (Forscher et al., 2019), and (c) responses on implicit bias tests may reflect the level of bias in a person's social environment rather than personal animosities (Payne et al., 2017). Over the past few years, these concerns have also received increased attention in the popular media, which is reflected in critical headlines such as *Can We Really Measure Implicit Bias? Maybe Not* (Bartlett, 2017) or *The False "Science" of Implicit Bias* (MacDonald, 2017).

Although some of the arguments against implicit bias tests can be criticized for ignoring important theoretical, empirical, and methodological issues (see Brownstein et al., 2020; Gawronski, 2019; Kurdi et al., 2019), the ongoing controversies surrounding these tests raise the question of whether it is wise to base antidiscrimination policies on the notion of implicit bias. In this article, we argue that criticism of implicit bias tests have implications for antidiscrimination policy only if implicit bias is equated with responses on these tests (e.g., when implicit bias is equated with people's responses on the Implicit Association Test). Although this conceptualization is widespread in the scientific literature, it is problematic for various reasons (see Calanchini & Sherman, 2013; Corneille & Hütter, 2020; De Houwer, 2019; Gawronski, 2019; Payne & Correll, in press).

An alternative conceptualization that seems superior for applied questions (i.e., policy) defines implicit bias in terms of actual discriminatory behavior. According to this conceptualization, discriminatory behavior represents an instance of implicit bias to the extent that the person showing the behavior is unaware that their behavior is biased. The central argument is that antidiscrimination policy should consider evidence for implicit bias in terms of this behavioral conceptualization instead of dismissing the notion of implicit bias because of extant controversies about implicit bias tests.

Sources of Implicit Bias in Behavior

From a psychological perspective, discrimination can be said to occur when a person's behavior toward a target individual is influenced by the target's group membership, including (but not limited to) the target's race, gender, or sexual orientation. Of particular concern for policy are instances of discrimination that involve negative outcomes for the target individual.[1] Examples include discrimination based on race or gender in hiring, recruitment, compensation, promotion, and termination; other examples include discrimination in housing and police support (Greenwald & Pettigrew, 2014). In terms of the above conceptualization, discriminatory behaviors in these cases are instances of implicit bias to the extent that a person is unaware that their behavior is influenced by the category membership of the target (e.g., the target's race or gender; see Gawronski & Bodenhausen, 2012). Research in social psychology has documented two mechanisms that can lead to implicit bias in terms of the proposed conceptualization: (a) biased weighting of mixed information and (b) biased interpretation of ambiguous information. The general pattern underlying both instances is that people show an initial response to the target that is influenced by the target's category membership, and this initial response influences the subsequent processing of information about the target.

Biased Weighting

One mechanism that can lead to implicit bias in terms of the proposed behavioral conceptualization is biased weighting of information (e.g., Hodson et al., 2002; Norton et al., 2004; Uhlmann & Cohen, 2005). Such biases tend to be particularly pronounced in cases involving judgments and decisions about multiple targets when the available information about these targets is mixed. For example, in hiring decisions involving a male and a female candidate with distinct job-relevant qualifications, an interviewer may attribute greater weight to the unique strengths of the male candidate (e.g., better grades) compared to the unique strengths of the female candidate (e.g., more experience). However, the differential weighting of strengths might be biased in the sense that it merely serves as a post hoc justification for hiring the male candidate rather than as an a priori criterion. For example, an interviewer might have an "intuitive" preference for a male over a female candidate, because there is a greater fit between social stereotypes about men and the qualities believed to be necessary for successful performance (Heilman, 2012). In such cases, the interviewer might rationalize their "intuitive" preference for the male candidate by focusing on unique strengths of the male candidate and/or unique weaknesses of the female candidate. To the extent that people are unaware of their bias in weighting mixed information in a manner that merely justifies a pre-existing preference, it can lead to discriminatory behavior in terms of the proposed conceptualization of implicit bias.

Empirical evidence for biased weighting of mixed information comes from a number of decision-making studies in which (a) participants were presented with sets of distinct information about two (or more) target individuals who differ in terms of their category membership (e.g., race and gender), and (b) the assignment of the information sets to the two targets was experimentally manipulated, such that participants in one condition saw Information X about Target A and Information Y about Target B, while participants in the other condition saw Information Y about Target A and Information X about Target B. A key aspect of these studies is that the two sets of information suggest distinct qualities in the sense that one set suggests a unique strength in one domain, whereas the other set suggests a unique strength in a different domain. A biasing effect of the target's category membership on participants' relative weighting of these strengths can be inferred when participants (a) show a

preference for the same target regardless of the information paired with the target (e.g., a preference for a male over a female candidate regardless of the information about the two candidates) and (b) justify their preference with the unique strength that happens to characterize the preferred candidate in the experimental condition randomly assigned.

For example, in a study by Norton et al. (2004), participants viewed application materials of a male and a female job candidate and indicated which of the two candidates they would prefer for particular job. In one condition, the male candidate had less work experience but more education than the female candidate did. In another condition, the male candidate had less education but more work experience than the female candidate did. Consistent with the idea of biased weighting, participants showed a preference for the male candidate in both experimental conditions and justified their responses with whatever qualification made him superior to the female candidate. That is, when the male candidate excelled in terms of education, participants listed education as the most significant criterion. Yet, when the male candidate excelled in terms experience, participants listed experience as the most significant criterion (for similar findings, see Hodson et al., 2002; Uhlmann & Cohen, 2005). Further research suggests that biasing effects of differential weighting occur outside of awareness, in that participants' self-perceptions of objectivity in their decision were associated with greater (rather than smaller) bias (Uhlmann & Cohen, 2005).

Although biased weighting can lead to discrimination in a wide range of real-world contexts, its effects are most prominently reflected in selective choice decisions, such as admission, hiring, and promotion decisions. In such cases, decision-makers often have to identify a small number of candidates (or only one) among a large number of highly qualified candidates. What makes these decisions particularly difficult is that the relevant evaluation criteria are often multidimensional rather unidimensional, forcing decision-makers to compare "apples and oranges" when candidates differ in term of their relative strengths. Thus, to the extent that the relative importance of evaluation criteria remains unspecified, decision-makers have to come up with their own weighting schema, leaving considerable room for arbitrary weightings that merely justify a decision-maker's biased preference (Bragger et al., 2002; Uhlmann & Cohen, 2005). Such biases are difficult to address, because decision-makers tend to think of their decisions as being based on their impressions of specific individuals rather than beliefs about the social groups to which these individuals belong (see Ledgerwood et al., 2020). For example, people may deny that gender had any influence on their preference for a male over a female candidate and refer primarily to unique strengths of the male target without realizing that they would justify their preference with whatever criterion makes the male candidate seem superior.

Another example of biased weighting in real-world contexts is bias in jury selection. In 1986, the U.S. Supreme Court ruled that prospective jurors could not be challenged on the basis of being a member of a cognizable racial group (*Batson v. Kentucky*, 1986). Subsequent rulings have extended this rule to preemptory challenges based on gender (*J.E.B. v. Alabama*, 1994). However, questions have been raised about whether requiring attorneys to justify suspicious challenges—which has become common practice since *Batson v. Kentucky*—is effective in preventing bias in jury selection (Sommers & Norton, 2008). Similar to the concern about biased weighting in the justification of hiring decisions, attorneys may justify their preemptory challenges by referring to race- and gender-neutral characteristics, but this does not mean that their challenges are unaffected by a juror's race and gender. In line with this concern, experimental studies found that race influenced preemptory challenges by advanced law students and practicing attorneys, but their justifications were entirely race-neutral (Sommers & Norton, 2007). Although participants might have been aware of their biased reasoning, biased weighting of information to justify a particular decision would qualify as an instance of implicit bias, to the extent that attorneys are unaware of the influence of race or gender on their preemptory challenges.

Biased Interpretation

Even when two individuals do the same thing, people often perceive the behavior differently depending on the category membership of the behaving person (e.g., Darley & Gross, 1983; Duncan, 1976; Gawronski et al., 2003; Hugenberg & Bodenhausen, 2003; Kunda & Sherman-Williams, 1993; Sagar & Schofield, 1980; Trope, 1986). Such biased perceptions are particularly pronounced when the observed behavior is ambiguous. For example, a teacher may perceive a student's essay for an English class as stronger when the student is white than when the student is black, but the student's race may have little impact on the teacher's perceptions of objectively correct or incorrect responses on a math exam (Darley & Gross, 1983). Because people tend to treat their subjective perceptions as direct reflections of objective reality rather than the product of active construal processes that are prone to perceptual biases (Trope & Gaunt, 1999), attempts to correct one's biased perceptions are relatively rare, leading to discriminatory behavior without people being aware of their biases (see Strack & Hannover, 1996; Wegener & Petty, 1997; Wilson & Brekke, 1994).

Empirical evidence for biased interpretations comes from a number of studies in which (a) participants were presented with ambiguous information about a target person and (b) the target person's category membership was experimentally manipulated, such that the target belonged to one social category (e.g., white) in one condition and a different social category (e.g., black) in another condition. A key aspect of these studies is that the ambiguous information is exactly the same in the two experimental conditions, the only difference being the category membership of the target. A biasing effect of the

target's category membership on participants' interpretations of the ambiguous behavior can be inferred when participants judge the behavior differently in the two experimental conditions.

For example, in a study by Hugenberg and Bodenhausen (2003), participants watched short video clips of either black or white targets whose facial expressions changed either from smiling to frowning or from frowning to smiling. The experimenters created the target faces with a three-dimensional (3D) computer program, such that the facial structure was identical for matched black and white targets, the only difference being their skin color and hairstyle. Participants' task was to press a key (a) as soon as they saw hostility in the target's face, when the facial expression was changing from smiling to frowning, and (b) as soon as they do not see any hostility in the target's face, when the facial expression was changing from frowning to smiling. Consistent with the hypothesis that even perceptions of basic emotional expressions can be biased by category membership, participants perceived hostility earlier and for longer durations when the target faces were black than when they were white (see also Bijlstra et al., 2014; Hutchings & Haddock, 2008). Further research suggests that such biasing effects occur outside of awareness, in that even people who are highly motivated to respond in a nonprejudicial manner show the same bias in their perceptions of ambiguous information (Gawronski et al., 2003).

The real-world relevance of biased interpretations can be illustrated with the cases listed under hashtag *#LivingWhileBlack*, which describe ordinary activities for which police have been called on African Americans (Griggs, 2018). In addition to the aforementioned case of waiting for an acquaintance at Starbucks, the list includes mundane activities such as moving into an apartment, making a phone call in a hotel lobby, shopping for prom clothes, not waving while leaving an Airbnb, eating lunch on a college campus, working as a home inspector, and delivering newspapers. The general theme underlying these cases is that, while the described behaviors tend to be perceived as ordinary when a white person does them, they are perceived as suspicious (and potentially threatening) when a black person does them.

A lethal variant of such biased perceptions is the tendency to more frequently misidentify harmless objects as weapons when they are held by a black person than when they are held by a white person (for a review, see Payne & Correll, in press). Although early research suggested that this tendency is rooted in impulsive response tendencies that can be intentionally controlled, given sufficient time and mental resources (Payne et al., 2005), more recent evidence supports the idea that the greater tendency to shoot unarmed black (vs. white) men is at least partly driven by unconscious visual processes leading to biased perceptions of ambiguous objects (Correll et al., 2015). Beyond racially biased identifications of harmless objects as weapons, unconscious perceptual biases have also been implicated in divergent perceptions of video evidence (Granot et al., 2018).

Another illustrative example is the concern that the same agentic behavior is often perceived less favorably when a woman does it than when a man does (Rudman et al., 2012). For example, while self-promoting, assertive, and dominant behavior is often interpreted positively in a man (e.g., reflecting confidence and leadership), the same behavior is more likely to be interpreted negatively in a woman (e.g., reflecting neuroticism and disagreeableness). In work contexts, such biased perceptions can lead to gender discrimination in promotions for leadership roles, given that promotion decisions depend on perceptions of leadership-relevant traits. Yet, unlike the idea that gender influences such decisions in a direct manner, the notion of biased interpretation suggests a more subtle, indirect effect. That is, a person's gender influences people's perceptions of the person's behavior, which in turn influences overall impressions of that person's suitability for a leadership role (Trope, 1986). As with the effects of biased weighting, such biases are difficult to address, because decision-makers tend to think of their decisions as being based on their impressions of a specific person rather than their beliefs about men and women in general (see Ledgerwood et al., 2020). Thus, people may deny that a target's category membership had any influence on their decision and refer primarily to their perceptions of the specific target person, without realizing that their perception of the target's behavior is influenced by the target's category membership (see Dovidio & Gaertner, 2004).

For example, a manager might carefully select a set of qualities that an employee should display to get a promotion (e.g., assertiveness and strong leadership potential) and then evaluate each employee with respect to those traits. Yet, implicit bias could creep into this decision if the manager perceives the same behavior differently depending on the group membership of the employees (e.g., Mark and Maria both express anger toward someone who missed a deadline, but Mark's behavior is interpreted as assertive, whereas Maria's behavior is interpreted as volatile; Mark is then evaluated as more assertive and thus more deserving of a promotion). Thus, even when people are careful to be evenhanded in their decision-making process, biased interpretations of ambiguous behavior may have already shaped upstream impressions of the individuals being evaluated.

Implications for Education and Intervention

Organizational efforts to combat bias have created a multibillion-dollar industry (Lipman, 2018). Yet, empirical assessments of their effectiveness in increasing diversity suggest a bleak conclusion (Kalev et al., 2006). Although the identified reasons for this outcome are complex and beyond the scope

of this article (for a discussion, see Carter et al., in press), the reviewed effects of biased weighting and biased interpretation suggest that extant interventions would benefit from considering their contributions to discrimination in the workplace and various other contexts.

Raising Awareness

A first step in this regard is to increase public awareness of the two sources of bias by educating people how biased weighting and biased interpretation can lead to discriminatory behavior. Examples of suitable contexts for this endeavor are organizational trainings and dedicated lectures in high-school classes, which may include presentations on the evidence reviewed above. Hands-on exercises that replicate experimental demonstrations of the two mechanisms could be particularly helpful to illustrate their impact. Popular media may also contribute to increasing public awareness by communicating the scientific evidence for biased weighting and biased interpretation to nonacademic audiences. Because describing bias as unconscious can lead people to feel less accountable for biased actions (Daumeyer et al., 2019; Payne et al., 2010), discussions of implicit bias should emphasize the responsibility of individuals and organizations to create policies and procedures to prevent expressions of implicit bias in individual behavior. To avoid implying that bias only exists at the level of individuals, these discussions should also contextualize the issue of implicit bias at the individual level in a broader understanding of systemic and historical bias (see Bonam, et al., 2019; Salter et al., 2018).

Strategies for Individuals

Although knowledge of the two mechanisms that we have described is an important first step in combatting their effects, such knowledge alone seems unlikely to eliminate their impact without additional hands-on strategies (Carter et al., in press). For example, a person may be aware that biased weighting can lead to discrimination in hiring decisions, but the person may not be aware that biased weighting influences their own hiring decision in a particular case. Regarding bias correction at the individual level, some research suggests that a strategy termed *consider-the-opposite* (Hirt & Markman, 1995; Lord et al., 1984) can be helpful to combat effects of biased weighting. The strategy involves a reconsideration of the same information assuming that the target differed on a potentially biasing characteristic. For example, in cases involving a choice between a male and a female job candidate, people may mentally simulate whether they would make a different choice if the qualifications of the two candidates were swapped. If people realize that their preference for the male candidate would be unaffected by a swap of qualifications, their formerly "implicit" bias would become "explicit" in the sense that they are now aware of the biasing effect of gender on their hiring preference. This insight allows decision-makers to "re-compute" their judgments taking the identified source of bias into account (Strack, 1992).[2]

Although mental simulations considering the opposite can be helpful in identifying effects of biased weighting, identifying effects of biased interpretation is more difficult. For example, in cases involving interpretations of ambiguous behavior shown by an African American person, people may mentally simulate how they would perceive the behavior if the target was white. To the extent that the behavior would be perceived differently for a white target, people would become aware of the biasing effect of race on their perception of the target's behavior, providing a basis to "re-compute" their judgments taking the identified source of bias into account (Strack, 1992). However, the likelihood of such awareness-raising effects is relatively low because such mental simulations are based, not on "objective" features of the observed behavior, but subjective interpretations of the behavior, which are prone to the bias described above. For example, a person may conclude that they should call the police on anyone who is trying to break into a house regardless of whether person is white or black. However, they may not realize that they are interpreting the target's ambiguous behavior as "trying to break into a house" only because the target is black and that they would not have interpreted the same behavior in this way if the target had been white. This intricate link makes it difficult to determine if one's perception of a person's behavior is biased by the person's category membership.

Strategies for Organizations

As the discussion above makes clear, identifying and correcting for implicit bias at the individual level can be challenging. Indeed, a more effective way to combat implicit bias is to change structures and procedures to create contexts in which discrimination is less likely to occur (Carter et al., in press; Salter et al., 2018). One of the most effective strategies in decision-making contexts is to "remove" potentially biasing category information, as is the case in the practice of blinded evaluation. Such a policy can effectively prevent effects of both biased weighting and biased interpretation (Goldin & Rouse, 2000). If there is no category information to begin with, it cannot bias the weighting of mixed information or the interpretation of ambiguous information.

To the extent that blinding is not feasible, an alternative strategy to prevent effects of biased weighting is to specify unambiguous decision criteria before decision-makers review any information about the relevant target individuals (Bragger et al., 2002; Uhlmann & Cohen, 2005). In hiring contexts, clear specifications and prior commitment to specific criteria can reduce arbitrary weightings that serve to merely justify a pre-existing preference independent of the actual information about the candidates (Uhlmann & Cohen, 2005). Similar effects occur for highly structured (compared to informal) interviews, which have proven their

effectiveness in reducing biases against pregnant job applicants (Bragger et al., 2002).

However, prior specification of evaluation criteria will increase diversity only if the identified criteria are unbiased in the sense that they do not favor members of certain groups. For example, a manager might select a set of qualities for evaluating employees that includes the traits *assertive*, *confident*, and *leadership potential*. Such a list can lead to biased outcomes if the identified qualities are more readily inferred from behaviors when the person performing the behavior is a man rather than a woman (e.g., via biased interpretations of ambiguous behavior). To combat this source of bias, decision-makers would need to be accountable for adding equally desirable qualities that fit better with stereotypes of women than men (e.g., *excellent communicator* and *inspires effective teamwork*), so that the resulting list of desired criteria became more balanced. It may also help to create procedures that increase the amount of time that evaluation committees spend discussing attributes that favor systematically disadvantaged candidates (e.g., asking committees to spend as much time discussing candidate warmth as they spend discussing candidate competence), although additional research is needed to test this intervention idea in real-world hiring contexts (Chang & Cikara, 2018).

Of course, the psychological processes underlying discrimination do not take place in a vacuum. Individual decisions and behaviors are always situated in a broader historical and societal context. Strategies designed to combat implicit bias at the individual level can only go so far (Bonam et al., 2019; Payne & Vuletich, 2018). It will be important for organizations to invest in long-term training (rather than expecting a single training to have long-term behavioral consequences), monitor training effectiveness in particular contexts, and develop organizational structures that increase accountability for diversity (e.g., diversity committees and staff positions; Carter et al., press; Kalev et al., 2006). Even perfectly evenhanded behavior at the individual level can perpetuate inequalities produced by long periods of systemic discrimination (see Kendi, 2017; Rothstein, 2017). Because such processes involve societal factors that go beyond the psychological mechanisms discussed in this article, they require additional strategies to combat bias at the systemic level (e.g., affirmative action policies).

Conclusion

Research on implicit bias has become the target of increased criticism, raising questions about whether antidiscrimination policy should be based on a controversial construct. In response to this concern, we argued that extant criticism of implicit bias tests (e.g., Implicit Association Test) affects antidiscrimination policy only if implicit bias is equated with responses on these tests; it remains unaffected if implicit bias is defined behaviorally in terms of actual discriminatory behavior. This alternative conceptualization highlighted the role of two well-understood mechanisms that can lead to discriminatory behavior outside of awareness: biased weighting of mixed information and biased interpretation of ambiguous information. Of course, either type of bias may be systematically related to responses on implicit bias tests, which is a question for basic scientific research (for a review, see Gawronski et al., 2006). However, this question is entirely irrelevant for antidiscrimination policy on implicit bias. What matters for such policy is implicit bias in actual discriminatory behavior.

The social psychological literature offers valuable insights into the mechanisms underlying implicit bias in actual discriminatory behavior and potential strategies to combat their effects. Some of these strategies have already proven their effectiveness in reducing bias (e.g., blinded evaluations, prior specification of evaluation criteria, and structured as opposed to informal interviews); others were derived from laboratory-based findings that await further testing in real-world contexts (e.g., public knowledge of the two mechanisms, consider-the-opposite, and stereotypically balanced evaluation criteria). Yet, all of them can be easily included in extant diversity trainings, and organizational executives can implement them into their decision-making procedures with little or no extra costs (e.g., blinded evaluations, prior specification of evaluation criteria, structured as opposed to informal interviews, stereotypically balanced evaluation criteria). Although effective interventions will require approaches that target individual, organizational, and systematic aspects of discrimination, neither approach will succeed without considering implicit bias in discriminatory behavior that occurs outside of awareness.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this article was supported by National Science Foundation Grant (grant no. BCS-1941440). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID iDs

Bertram Gawronski https://orcid.org/0000-0001-7938-3339
Alison Ledgerwood https://orcid.org/0000-0002-4535-6276

Notes

1. Note that a purely psychological definition of discrimination does not cover systemic aspects (e.g., the lingering consequences of slavery, redlining, and the denial of civil rights), which we deem equally important for policy, yet are beyond

the scope of this article. Although this article focuses mainly on the psychological level, we deem policies that treat everyone equal regardless of group membership as insufficient, because such policies tend to perpetuate existing inequalities rooted in systemic discrimination (see Rothstein, 2017).

2. Although it is possible that some people respond defensively to the outcomes of their mental simulations and try to justify their initial preference, any such justifications will differ from the initial ones because people would have to justify a bias that is now explicit (e.g., they would have to justify why they would hire a male over a female candidate regardless of their qualifications).

References

Bartlett, T. (2017, January 5). Can we really measure implicit bias? Maybe not. *The Chronicle of Higher Education*. https://www.chronicle.com/article/Can-We-Really-Measure-Implicit/238807

Batson v. Kentucky. (1986) 476 U.S. 79.

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.

Bonam, C. M., Nair Das, V., Coleman, B. R., & Salter, P. (2019). Ignoring history, denying racism: Mounting evidence for the Marley hypothesis and epistemologies of ignorance. *Social Psychological and Personality Science*, *10*, 257–265.

Bragger, J. D., Kutcher, E., Morgan, J., & Firth, P. (2002). The effects of the structured interview on reducing biases against pregnant job applicants. *Sex Roles*, *46*, 215–226.

Brownstein, M., Madva, A., & Gawronski, B. (2020). Understanding implicit bias: Putting the criticism into perspective. *Pacific Philosophical Quarterly*, *101*, 276–307

Calanchini, J., & Sherman, J. W. (2013). Implicit attitudes reflect associative, non-associative, and non-attitudinal processes. *Social and Personality Psychology Compass*, *7*, 654–667.

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

Chang, L. W., & Cikara, M. (2018). Social decoys: Leveraging choice architecture to alter social preferences. *Journal of Personality and Social Psychology*, *115*, 206–223.

Chapell, B. (2018, May 29). Starbucks closes more than 8,000 stores today for racial bias training. *National Public Radio*. https://www.npr.org/sections/thetwo-way/2018/05/29/615119351/starbucks-closes-more-than-8-000-stores-today-for-racial-bias-training

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*. Advance online publication. https://doi.org/10.1177/1088868320911325

Correll, J., Wittenbrink, B., Crawford, M., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate complex visual stimuli. *Journal of Personality and Social Psychology*, *108*, 219–233.

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

Daumeyer, N. M., Onyeador, I., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, *84*, 103812.

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

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. *Advances in Experimental Social Psychology*, *36*, 1–52.

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.

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.

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

Gawronski, B., & Bodenhausen, G. V. (2012). Self-insight from a dual-process perspective. In S. Vazire & T. D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 22–38). Guilford Press.

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., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition*, *15*, 485–499.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review*, *90*, 715–741.

Granot, Y., Balcetis, E., Feigenson, N., & Tyler, T. (2018). In the eyes of the law: Perception versus reality in appraisals of video evidence. *Psychology, Public Policy, and Law*, *24*, 93–104.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, *69*, 669–684.

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*. https://www.cnn.com/2018/12/20/us/living-while-black-police-calls-trnd/index.html

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, *32*, 113–135.

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.

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.

J.E.B. v. Alabama. (1994). 511 U.S. 127.

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review*, *71*, 589–617.

Kendi, I. X. (2017). *Stamped from the beginning: The definitive history of racist ideas in America*. Penguin Random House.

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

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.

Ledgerwood, A., Eastwick, P. W., & Gawronski, B. (2020). Experiences of liking versus ideas about liking. *Behavioral and Brain Sciences*, *43*, e136.

Lipman, J. (2018, January 25). How diversity training infuriates men and fails women. *Time*. http://time.com/5118035/diversity-training-infuriates-men-fails-women/

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.

MacDonald, H. (2017, October 9). The false "science" of implicit bias. *Wall Street Journal*. https://www.wsj.com/articles/the-false-science-of-implicit-bias-1507590908

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

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.

Park, M. (2018, April 18). What the Starbucks incident tells us about implicit bias. *CNN*. https://www.cnn.com/2018/04/17/health/implicit-bias-philadelphia-starbucks/index.html

Payne, B. K., Cameron, C. D., & Knobe, J. (2010). Do theories of implicit race bias change moral judgments? *Social Justice Research*, *23*, 272–289.

Payne, B. K., & Correll, J. (in press). Race, weapons, and the perception of threat. *Advances in Experimental Social Psychology*.

Payne, B. K., Shimizu, Y., & Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon identifications. *Journal of Experimental Social Psychology*, *41*, 36–47.

Payne, B. K., & Vuletich, H. A. (2018). Policy insights from advances in implicit bias research. *Policy Insights From the Behavioral and Brain Sciences*, *5*, 49–56.

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.

Rothstein, R. (2017). *The color of law: A forgotten history of how our government segregated America*. Liveright.

Rudman, L. A., Moss-Racusin, C. A., Glick, P., & Phelan, J. E. (2012). Reactions to vanguards: Advances in backlash theory. *Advances in Experimental Social Psychology*, *45*, 167–227.

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.

Salter, P. S., Adams, G., & Perez, M. J. (2018). Racism in the structure of everyday worlds: A cultural-psychological perspective. *Current Directions in Psychological Science*, *27*, 150–155.

Sommers, S. R., & Norton, M. I. (2007). Race-based judgments, race-neutral justifications: Experimental examination of peremptory use and the Batson challenge procedure. *Law and Human Behavior*, *31*, 261–273.

Sommers, S. R., & Norton, M. I. (2008). Race and jury selection: Psychological perspectives on the peremptory challenge debate. *American Psychologist*, *63*, 527–539.

Strack, F. (1992). The different routes to social judgments: Experiential versus informational strategies. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 249–275). Lawrence Erlbaum Associates.

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). Guilford Press.

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

Trope, Y., & Gaunt, R. (1999). A dual-process model of overconfident attributional inferences. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 161–178). Guilford Press.

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

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.

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