**Debiasing**

*The holy grail of implicit race bias research is to change the underlying associations that form the basis of implicit bias.*


While implicit biases are deeply entrenched in the subconscious, researchers generally agree that biases are malleable and that implicit associations may be unlearned (see, e.g., Blair, 2002; Blair, Ma, & Lenton, 2001; Dasgupta & Greenwald, 2001; Devine, 1989; Kang, 2009; Kang & Lane, 2010). As discussed in this chapter, the debiasing process can take many different forms and yield varying results depending on factors such as individual motivation and context, as these influence what associations are brought to the foreground of one’s mind (Foroni & Mayr, 2005).

Debiasing is far from a simple task, as it involves the construction of new mental associations. Devine writes, “Inhibiting stereotype-congruent or prejudice-like responses and intentionally replacing them with non-prejudiced response can be likened to the breaking of a bad habit” (Devine, 1989, p. 15). She adds how “intention, attention, and time” are needed so that new responses are learned well enough to compete with the formerly automatically activated responses (Devine, 1989, p. 16). Given how strongly rooted implicit biases tend to be, debiasing efforts have to compete against stimuli that can, in effect, “re-bias” (Kang, et al., 2012, p. 1170).

The first inclination for many people who realize they hold implicit racial/ethnic biases may be to attempt to debias by repressing these biased thoughts; however, this notion generally has not been supported by the literature due to “rebound effects.” Suppressing automatic stereotypes does not reduce them and may even amplify them by making them hyper-accessible (Galinsky & Moskowitz, 2000, 2007; Macrae, Bodenhausen, Milne, & Jetten, 1994). Studies have shown that “instead of repressing one’s prejudices, if one openly acknowledges one’s biases, and directly challenges or refutes them, one can overcome them” (bstan-’dzin-rgya-mtsho & Cuttler, 2009, p. 70). Similarly, Blair and Banaji found that conscious efforts to counter stereotypes can inhibit the activation of automatic associations (Blair & Banaji, 1996).
Numerous researchers have used the IAT to demonstrate the malleability of implicit attitudes. For example, Kawakami and colleagues found that proximity, as examined through approach-avoidance orientations, affected implicit racial bias scores across a series of four studies (Kawakami, Phillips, Steele, & Dovidio, 2007). Ito and colleagues published another example of the IAT documenting the malleability of implicit racial attitudes that they achieved by manipulating participants’ emotional cues. Some participants were surreptitiously induced to smile by holding a pencil in their mouths while viewing photos of unfamiliar Black or White men; others were not instructed to perform any such somatic manipulation. IAT results showed that this manipulation influenced IAT outcomes, as individuals who surreptitiously smiled while viewing Black faces earlier displayed less racial bias against Blacks (Ito, Chiao, Devine, Lorig, & Cacioppo, 2006). Richeson and Ambady considered the role of situational power (i.e., whether one is regarded as superior or subordinate in an interracial dyad) was reflected by changes in implicit racial attitudes in interracial interactions but not in same-race interactions (Richeson & Ambady, 2003). These studies, among others, declare implicit biases to be malleable.

With this in mind, it is logical that subsequent attention has been devoted to using this malleability property to counter existing biases. The following sections examine various debiasing techniques.

**Interventions that may debias successfully**

*Counter-stereotypic training*

A significant portion of the debiasing research centers on interventions that counter stereotypes and train individuals to develop new associations. One such example, advanced by Wittenbrink, Judd, and Park (2001) focused on how modifying the situational context may influence racial attitudes. By juxtaposing ordinary people in counter-stereotypic situations, such as depicting young White and Black males in scenes that included a church and a graffiti-strewn street corner, researchers found that the context condition affected participants’ racial attitudes on a subsequent sequential priming task (Wittenbrink, Judd, & Park, 2001). The data from this counter-stereotypic training indicated that social category clues may affect individuals’
automatic responses and racial attitudes.

Taking the notion of countering stereotypes rather literally, Kawakami et al., 2000 studied the effects of training people to negate stereotypic associations, including racial associations. By instructing participants to verbally respond “no” when presented with a stereotypic trait that matched a category representation and “yes” when viewing non-stereotypic associations, they found that participants who received this stereotype negation training displayed diminished stereotype activation (Kawakami, et al., 2000).

Notably, this effect remained salient 24 hours after the training ended (Kawakami, et al., 2000). These findings emphasize the importance of not just counter-stereotypic instruction, but also the need for consistent repetition of this instruction. Kawakami and colleagues later extended this work to examine the effect of training on non-stereotypic traits of men and women in the context of hiring decisions and found similar results supporting the effectiveness of counter-stereotypic training (see Kawakami, Dovidio, & Kamp, 2005).

Finally, Blair and colleagues researched the strategy of using mental imagery as a way to moderate implicit stereotypes. Over the course of five experiments that used mental imagery to target gender stereotypes, they found compelling evidence that counter-stereotypical mental imagery yielded notably weaker implicit stereotypes as compared to the implicit stereotypes assessed in individuals who either engaged in other forms of mental imagining or no mental imagery whatsoever (Blair, et al., 2001).

Rather than only mental, imagery in other forms may be used to debias. For courtroom settings, Kang and colleagues suggest the use of posters, pamphlets, photographs, and similar materials that would provoke counter-typical associations in the minds of jurors and judges (Kang, et al., 2012). The effects of this would likely vary based on the amount of exposure, with the underlying intention that even momentarily activating a different association may help decrease the presence of implicit bias during legal processes.
Exposure to counter-stereotypic individuals

Another type of intervention focuses on exposing people to individuals who contradict widely-held stereotypes. One fascinating study by Dasgupta and Greenwald (2001) investigated whether exposure to counter-stereotypic exemplars could decrease automatic preferences, such as that for White over Black Americans. They found that exposure to pro-Black exemplars (e.g., Michael Jordan, Colin Powell, Martin Luther King, Jr.) as opposed to nonracial or pro-White exemplars (e.g., Tom Hanks, Jay Leno, John F. Kennedy) significantly decreased the automatic White preference effect, as measured by the IAT (Dasgupta & Greenwald, 2001). Similar to the findings of Kawakami et al., 2000, this effect had staying power, as the IAT effect of this pro-Black condition remained 24 hours after the exposure to images of admired Blacks and disliked Whites (e.g., Jeffrey Dahmer, Timothy McVeigh, Al Capone). Emphasizing the malleability of implicit biases, the authors suggest that “creating environments that highlight admired and disliked members of various groups … may, over time, render these exemplars chronically accessible so that they can consistently and automatically override preexisting biases” (Dasgupta & Greenwald, 2001, p. 807).

This scholarship aligns well with the concept of debiasing agents, which refers to individuals whose traits contrast with the stereotypes typically associated with that particular category (Kang & Banaji, 2006). The presence of debiasing agents decreases the implicit biases of those they encounter due to their unique positioning. Example debiasing agents would include male nurses, elderly athletes, and female scientists. In many cases, debiasing agents change individuals’ implicit stereotypes, not just their implicit attitudes (Kang & Banaji, 2006). However, to be effective, debiasing agents must be viewed as not merely an exception but rather connect that individual to relevant categories, regardless of any counter-stereotypical traits they may also possess (Kang & Banaji, 2006).

The success of the exposure to counter-stereotypic individuals intervention has been echoed by other studies that do not focus explicitly on race. For example, one study found that exposure to women in leadership positions at a women’s college led to students being less likely to express automatic gender stereotypes about women,
compared to students from a coeducational college (Dasgupta & Asgari, 2004).

Moreover, exposure to counter-stereotypic exemplars does not even need to occur through in-person interactions. Renowned implicit bias scholar Mahzarin Banaji works to offset her own implicit biases through viewing images of counter-stereotypical individuals on her computer screensaver (Lehrman, 2006). Photos and other wall décor can serve a similar purpose (Kang, et al., 2012; National Center for State Courts).

However, some researchers question this counter-stereotypic exemplar debiasing method. First, an extensive 2010 study by Schmidt and Nosek examined whether Barack Obama, as a high-status famous Black exemplar shifted implicit or explicit racial attitudes during his candidacy and early presidency. Results from a heterogeneous sample of nearly 480,000 individuals led to the conclusion that there was minimal evidence that implicit racial attitudes changed systematically due to Obama’s presence as a counter-stereotypic exemplar (Schmidt & Nosek, 2010).

The authors suggest that the mere presence of a high-status counter-stereotypic exemplar may be inadequate to shift implicit or explicit racial attitudes (Schmidt & Nosek, 2010). Building on this work, Lybarger and Monteith’s research concluded that President Obama’s saliency alone did not have a debiasing effect, as one individual may be inadequate to shift long-standing implicit racial associations (Lybarger & Monteith, 2011).

Second, Joy-Gaba and Nosek provide a word of caution regarding the perceived malleability of implicit biases through exposure to counter-stereotypic exemplars. Their efforts to replicate and expand upon the work of Dasgupta and Greenwald (2001) noted earlier in this subsection yielded conspicuously less convincing results. Notably, while the degree of malleability found by Dasgupta and Greenwald was quite high on both the initial measure (d = 0.82) and follow-up 24 hours later (d = 0.71), Joy-Gaba and Nosek found significantly weaker effect magnitudes (d = 0.17 and 0.14, respectively) (Joy-Gaba & Nosek, 2010). This discrepancy leads the authors to broadly conclude that the extent to which the literature declares implicit biases to be malleable may have been overstated (Joy-Gaba & Nosek, 2010).
Intergroup Contact

Championed by American psychologist Gordon W. Allport in 1954, intergroup contact theory asserts that four key conditions are the necessary for positive effects to emerge from intergroup contact (Allport, 1954). Allport stipulated that optimal intergroup contact involves individuals of equal status, which explains why some relationships, such as that of student and teacher, do not necessarily lead to reductions in bias. Other conditions that yield positive intergroup contact effects include sharing common goals, interacting in a cooperative rather than competitive setting, and being supported by authority figures, laws, or customs. Allport’s theory has been supported consistently in the literature, including through large-scale meta-analyses (see, e.g., Pettigrew & Tropp, 2006, 2011).

Beyond simply gaining familiarity with outgroups through intergroup contact, these interactions have been shown to reduce implicit bias. For example, Thomas Pettigrew’s multi-national study found that “the reduction in prejudice among those with diverse friends generalizes to more positive feelings about a wide variety of outgroups” (Pettigrew, 1997, p. 180-181). Moreover, ten years later in a meta-analytic test of intergroup contact theory, Pettigrew and Tropp examined 713 samples and concluded that ingroup contact generally reduces intergroup prejudice (Pettigrew & Tropp, 2006).

With respect to the realms of criminal justice and health care previously discussed, intergroup contact can play a debiasing role in specific contexts. Peruche and Plant studied police officers and noted that “high levels of negative contact with Black people at work were related to negative expectations regarding Black suspects and marginally more negative attitudes toward Black people generally;” however, intergroup contact with Blacks outside of the workplace countered these effects (Peruche & Plant, 2006, p. 197). Similarly, diverse clinical care teams are vital to health care, because in a diverse team where members are granted equal power, “a sense of camaraderie develops that prevents the further development of stereotypes based on race/ethnicity, gender, culture or class” (Betancourt, 2004, p. 108).
Education about Implicit Bias

Efforts aimed at raising awareness of the phenomenon of implicit bias can also debias. This education can take several forms. For example, U.S. district judge Mark W. Bennett educates potential jurors about implicit bias during his time with them during the juror selection process. Judge Bennett aims to explain implicit bias and make jurors skeptical of their own objectivity through a 25-minute lesson that concludes by asking each juror to sign a pledge against bias. The text of this pledge is prominently displayed in the jury room. Then, at the beginning of the trial, Judge Bennett reiterates how implicit bias can taint jurors’ judgment by giving a short speech before the lawyers’ opening statements. He believes in the positive outcomes studies have documented regarding individuals’ responses to awareness of their own implicit biases (Bennett, 2010).

A recent article by Anna Roberts strongly supports the idea of using the IAT to educate jurors about implicit bias while dismissing the notion that the IAT should be used to “screen” prospective jurors (Roberts, 2012). Much like Judge Mark Bennett, Roberts recommends that this educational component be integrated into juror orientation, preferably with jurors receiving hands-on experiential learning that includes taking the IAT, as this is more impactful than passively learning about the IAT and its findings (Roberts, 2012).

Judges can also benefit from implicit bias education, which in large part involves persuading them of the presence of this problem (Kang, et al., 2012; Saujani, 2003). Organizations such as the National Center for State Courts (NCSC) have begun creating resources, such as films and assessments, designed to raise judicial awareness of implicit biases and their implications in a courtroom setting. Results from pilot sites showed promising preliminary results (Kang, et al., 2012). Research suggests that educating judges about implicit bias is most effective under three circumstances: 1) training should start early, such as during new judge orientation when

2. See Appendix A for the text of this brief speech.

3. The National Center for State Courts have posted some of their educational films and materials here: http://www.ncsconline.org/D_Research/ref/implicit.html
people are most open to new ideas; 2) the training should be presented in such a way that judges do not feel defensive; it should not be accusatory in nature; and 3) judges should be encouraged to take the IAT, as the results often prompt action (Kang, et al., 2012).

An entire industry around diversity education and trainings has proliferated in recent years, offering participants promises of reduced prejudice and greater appreciation of various cultures. Studies have examined whether diversity education can counter implicit biases, though the results are mixed. One study found that Whites who volunteered for diversity education forums showed lower levels of implicit and explicit anti-Black prejudice, with the change in implicit orientations “predicted by emotion-based factors, including reduced fear of Blacks, and liking for the Black professor who taught the course” (Rudman, 2004, p. 136; Rudman, Ashmore, & Gary, 2001). Conversely, other research finds diversity training yield minimal effects, particularly from a long-term perspective (Rynes & Rosen, 1995).

Accountability

Having a sense of accountability, meaning “the implicit or explicit expectation that one may be called on to justify one’s beliefs, feelings, and actions to others,” can be another powerful measure to combat bias (Lerner & Tetlock, 1999, p. 255). Research finds that having a sense of accountability can decrease the influence of bias (Kang, et al., 2012; Reskin, 2005). When decision makers are not held accountable for their actions, they are less likely to self-check for how bias may affect their decision-making (National Center for State Courts). Jurors’ feelings of being held accountable by the judge to produce unbiased decisions can help jurors keep their implicit biases in check (Kang, et al., 2012).

Fostering Egalitarian Motivations

Considerable research has shown that fostering egalitarian motivations can counter the activation of automatic stereotypes (Dasgupta & Rivera, 2006; Moskowitz, Gollwitzer, Wasel, & Schaal, 1999). Stone and Moskowitz write, “When activated, egalitarian goals inhibit stereotypes by undermining and counteracting the implicit
nature of stereotype activation, thereby cutting stereotypes off before they are brought to mind” (Stone & Moskowitz, 2011, p. 773). For example, work by Dasgupta and Rivera found that automatic biases are not necessarily inevitable, as the relationship between automatic antigay prejudice and discrimination was moderated by individuals’ conscious holding of egalitarian beliefs (Dasgupta & Rivera, 2006).

Taking the Perspective of Others

Another debiasing strategy that has gained some traction is when individuals take the perspective of someone who is different from them. Across three experiments, Galinsky and Moskowitz found that perspective-taking was effective at debiasing, as it “tended to increase the expression of positive evaluations of the target, reduced the expression of stereotypic content, and prevented the hyperaccessibility of the stereotype construct” (Galinsky & Moskowitz, 2000, p. 720).

Benforado and Hanson support perspective-taking as a debiasing tool, noting that considering opposing perspectives and fostering recognition of multiple perspectives are good techniques for reducing automatic biases (Benforado & Hanson, 2008). They caution, however, that this approach may have limited effects if individuals believe they have taken the perspective of others when in fact they have not been as successful at this venture as they judge themselves to be.

Later empirical work by Todd et al. shed light on effects of perspective taking. The researchers employed five experiments designed to assess whether taking the perspective of others could counter automatic expressions of racial bias. Their findings found that this debiasing technique yielded “more favorable automatic interracial evaluations” (Todd, Bodenhausen, Richeson, & Galinsky, 2011, p. 1038).

Taking the perspective of others can also be used in debiasing exercises. When Stone and Moskowitz outlined the components of a cultural competency workshop for medical professionals that sought to educate them about implicit bias, the authors suggested that the medical professionals imagine themselves as a minority group patient and write a story about that person’s life (Stone & Moskowitz, 2011). Finally, it is worth noting that perspective taking has benefits that extend beyond
debiasing. For example, in the realm of healthcare, studies have shown that encouraging practitioners to take the perspective of others cultivates empathy, which leads to positive outcomes for patient satisfaction and treatment (see, e.g., Blatt, LaLacheur, Galinsky, Simmens, & Greenberg, 2010; Drwecki, Moore, Ward, & Prkachin, 2011).

Deliberative Processing

Another technique that can counter implicit biases is to “engage in effortful, deliberative processing” (Kang, et al., 2012, p. 1177). This is particularly important for individuals who may be operating under time constraints or a weighty cognitive load, such as doctors and judges, because spontaneous judgments can provoke reliance on stereotypes (Burgess, 2010; Kang, et al., 2012). To that end, Betancourt suggests that medical professionals constantly self-monitor their behaviors in an effort to offset implicit stereotyping (Betancourt, 2004).

In another manner of deliberative processing, Stone and Moskowitz encourage medical professionals to rethink the standard ways that patients are classified (e.g., race/ethnicity, gender, etc.) and instead focus on a common identity that they share with each patient (Stone & Moskowitz, 2011). By activating this shared identity, the patient’s other identities (e.g., race/ethnicity) are not as prevalent in the medical professional’s mind, thus helping to counter the enactment of the implicit biases and stereotypes associated with those identities (Stone & Moskowitz, 2011).

The significance of deliberative processing is reinforced by research that finds that even one’s emotional state can influence the activation and nature of implicit biases (Dasgupta, DeSteno, Williams, & Hunsinger, 2009). For example, DeSteno and colleagues examined how the creation of automatic outgroup prejudice can be affected by emotional states, such as anger or sadness. Using both an evaluative priming measure and the IAT, they found that an angry emotional state led to automatic prejudice against outgroups, which the researchers attributed to anger’s association with intergroup competition and conflict (DeSteno, Dasgupta, Bartlett, & Cajdric, 2004). Thus, deliberate processing, including self-awareness of one’s own emotional state, plays a role in individuals’ ability to counter implicit biases.
Other Interventions

As discussed in this final subsection, some researchers have developed specific interventions as a means of debiasing.

As an extension of work that relied on counter-stereotypic exemplars, Foroni & Mayr showed how short fictional scenarios designed to present a counter-stereotypic example (in this case, flowers were regarded noxious while insects were positively regarded) had an immediate and automatic modulation of the IAT effect (Foroni & Mayr, 2005). This same effect was not observed when subjects were simply asked to think of flowers as negative and insects as positive. “These results suggest that a newly acquired knowledge structure targeting the abstract, category level can produce behavioral effects typically associated with automatic categorization” (Foroni & Mayr, 2005, p. 139).

A recent publication by Devine et al. highlights an intervention that is founded on the premise that “implicit bias is like a habit that can be broken through a combination of awareness of implicit bias, concern about the effects of that bias, and the application of strategies to reduce bias” (Devine, Forscher, Austin, & Cox, 2012, p. 1267). The logic here is that “breaking the habit” of implicit bias requires awareness of the contexts that can activate bias and knowledge of how to replace biased reactions with ones that reflect a non-prejudiced mindset. Devine and colleagues sought to assess whether interventions could yield long-term reductions in implicit racial bias. They used a randomly controlled experimental design in which participants assigned to the intervention group engaged in a bias education and training program that taught participants five strategies they could apply to different situations in their lives as appropriate (stereotype replacement, counter-stereotypic imaging, individuation, perspective taking, and increasing opportunities for contact). Results showed that the participants who had received this training had lower IAT scores than the control group participants, and unprecedentedly, this reduction in implicit race bias endured for at least eight weeks following the intervention (Devine, et al., 2012). Devine et al. attribute this decline in implicit bias to the multifaceted nature of the intervention rather than any specific aspect of the intervention.