The Psychology of Implicit Intergroup Bias and the Prospect of Change

Calvin K. Lai[1] & Mahzarin R. Banaji[2]

[1] Washington University in St. Louis

[2] Harvard University

January 10, 2019

Correspondence should be addressed to: Calvin Lai at calvinlai@wustl.edu.

Paper should be cited as:

Lai, C. K., & Banaji, M. R. (2019). The psychology of implicit intergroup bias and the prospect of change. In D. Allen & R. Somanathan (Eds.), *Difference without Domination: Pursuing Justice in Diverse Democracies*. Chicago, IL: University of Chicago Press.

1. Introduction

Over the course of evolution, human minds acquired the breathtaking quality of consciousness which gave our species the capacity to regulate behavior. Among the consequences of this capacity was the possibility of internal dialogue with oneself about the consistency between one's intentions and actions. This facility to engage in the daily rituals of deliberative thought and action is so natural to our species that we hardly reflect on it or take stock of how effectively we are achieving the goal of intention-action consistency. We do not routinely ask at the end of each day how many of our actions were consistent with the values so many individuals hold: a belief in freedom and equality for all, in opportunity and access for all, in fairness in treatment and justice for all.

Even if we wished to compute the extent to which we succeed at this task, how would we go about doing it? As William James (1904) pointed over a century ago, the difficulty of studying the human mind is that the knower is also the known, and this poses difficulties in accessing, in modestly objective fashion, the data from our own moral ledger. Ask and you will learn that people believe themselves to be good moral actors who may not be perfect but are largely behaving as they wish they would (Aquino & Reed, 2002). This result is especially likely when we probe people's attitudes and beliefs about social groups (Banaji & Greenwald, 2013). And yet, since the implicit revolution (Greenwald & Banaji, 2017), the data from the social and behavioral sciences have repeatedly shown significant discrepancies between the lack of expression of conscious prejudice and stereotypes and data on group disparities in hiring and promotion, medical treatment, access to financial resources, education, and basic living conditions (e.g., by gender, age, race, ethnicity, sexuality). In this chapter, we report on what is known from psychological science about the mental limits to fair treatment with a special focus on what we know about the possibilities for change in mind and behavior.

A humble three-pound organ, the brain, gives rise to the grandest thinking apparatus called the mind. *Beliefs* and *attitudes* originate here, and these thoughts and feelings shape societies and the course of civilization itself. Among the most significant roles that beliefs and attitudes play is to help humans formulate concepts of good and bad, right and wrong. Modern humans in democratic societies set a premium on particular values: on freedom, on equal opportunity, and on fairness. These values are broadly deemed to be the foundations of a just society. And yet as scientists like ourselves have studied the extent to which *implicit* or unconscious attitudes and beliefs subvert these ideals, we have discovered just how early in life, how quickly, and how subtly they do so. For this reason, any discussion about social difference without domination must begin with understanding the act of an individual mind making ordinary and everyday decisions about others – decisions about core features such as a person's worth, competence, and goodness.

A quick glance at the past century of race relations in the United States reveals dramatic changes in how Americans think about race. Racially egalitarian principles infuse American laws and institutions. The workplace, transportation, and public spaces in many parts of the country are visibly more diverse in race and ethnicity. Quite the reverse of the America of only a few decades ago, major corporations, governmental agencies, and NGOs publicly promote their commitment to egalitarian merit-based treatment. They showcase their commitment through the affinity groups they sponsor, the awards they receive for being leaders on diversity, and the representations of groups among their ranks who were previously absent in their industry.

Healthcare systems devote time and attention to reducing disparities in health outcomes. Educational institutions set aside sizable funding to ensure that the most meritorious can attend. Government agencies, whether they involve law and law enforcement, labor, health, education, housing, or the military speak enthusiastically about their efforts to increase access and opportunities. Indeed the military, in its amicus brief for the Supreme Court cases on affirmative action, Gratz vs. Bollinger and Grutter v. Bollinger (Becton et al., 2003), made the case that diversity in the military is necessary for the military to do its job – to keep the country safe. These principles are also present in public discourse with demands for civil speech that does not rely on group-based prejudices and stereotypes and chastising on social media of expressions of views that curtail freedom, opportunity, and fairness in treatment of all.

These public shifts in opinions and actions that characterize race relations today are simultaneously reflected in private shifts in attitudes and beliefs. A series of studies focused on racial/ethnic stereotypes known as the Princeton Trilogy shows this change most dramatically. At various moments since the 1930s, social psychologists asked White American undergraduates to report privately on what they believed to be true of various racial/ethnic groups. In 1933, 75% of students endorsed the stereotype that Black Americans were lazy (Katz & Braly, 1933). In 1951, that percentage was reduced by more than half to 31% (Gilbert, 1951). By 2001, that percentage was cut by two-thirds more. Only 12% of White students reported that African Americans were lazy (Madon et al., 2001).

Similar research on White Americans' opinions about various policies have been conducted over the decades, and those opinions have changed dramatically as well. Figure 1 shows four decades of results for White Americans' answers to two questions, one about school segregation and another about residential segregation. In the early 1960s, only about 60% of White Americans supported racial integration in schools. By 1995, support for school integration had grown to almost 100 percent. Shortly after, the question was removed from surveys because the question ceased to be informative. Similar increases in support were found for racial integration of neighborhoods. Whereas only 40% of White Americans supported housing integration in the early 1960s, more than 80% did so by the 1990s.

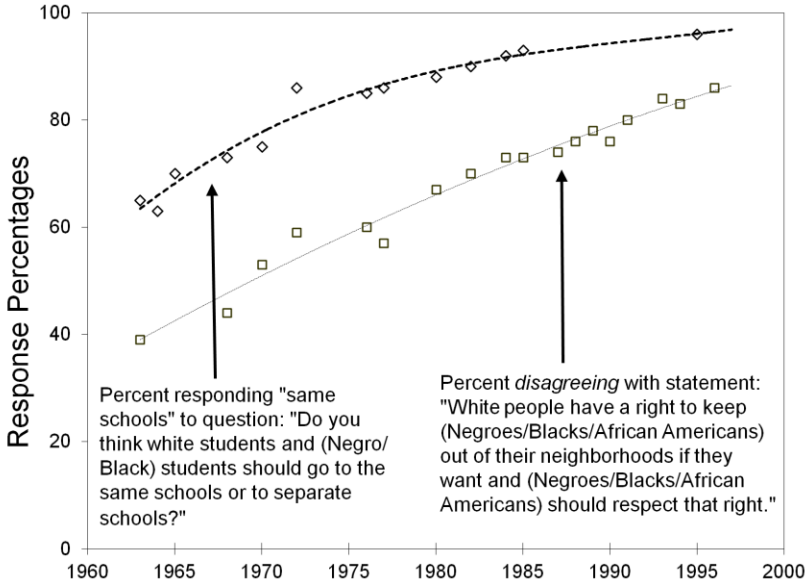

Figure 1. White Americans' increasing support for racial integration (1963-1996). Data source: Schuman, H., Steeh, C., Bobo, L., & Krysan, M. (1997). Racial attitudes in America. Cambridge, Mass.: Harvard University Press. Table 3.5A.

A conclusion from studies like these is that racial animus has largely disappeared. There is good reason to believe that such responses are not "just for show" but that Americans have changed their privately held conscious attitudes and beliefs as well. But this account belies the continued presence of discrimination in American life. There is a mismatch between the attitudes and beliefs that are expressed and the facts of Black lives. Focusing on race alone, we have many examples of racial segregation (Iceland, Weinberg, & Steinmetz, 2002), job discrimination (Bertrand & Mullainathan, 2004), and inequalities in access to basic constitutional rights and their entailments – from education (U.S. Department of Education, 2016), to housing (U. S. Census, 2015), to employment (U. S. Bureau of Labor Statistics, 2016), and to health (LaVeist, 2003).

A large tradition of research addresses this disconnect between what people say and what they do. Our particular strand of this research has looked at the earliest possible feelings and thoughts that can be measured behaviorally[1] with the intent of measuring less conscious forms of racial attitudes and beliefs. This research tells a different story about how present day Americans think and feel about social groups (Banaji & Greenwald, 2013; Devine, 1989; Macrae & Bodenhausen, 2000; Gawronski & Payne, 2010). Much of the research on implicit cognition taps into introspectively unidentified thoughts using measures that examine the strength of association between a *concept* (e.g., Black or White) and an *attribute* (e.g., good-bad, smart-dumb, rich-poor) by relying on the relative speed to make pairings between the concept and attribute. Using a variety of methods to get at these associations has led to a striking set of discoveries. Among the most central of these discoveries is that within the same individual mind there exist multiple actors: a deliberative decision-maker who aspires to egalitarian ideals and a less conscious partisan who is attentive to the similarity, familiarity, and social standing of those who are judged. Such lack of consistency within the same mind is psychologically interesting and morally consequential. It is easier to understand that a society may consist of people who vote for equal opportunity and fairness in treatment and those who do not. But to discover that one and the same mind can hold these differing values is perplexing. It raises questions such as why such mental states exist, where they come from, how pervasive they are, and their range of influence.

To set a manageable scope for our discussion, and because research on implicit social cognition has been covered elsewhere (Gawronski & Payne, 2010), we focus on a particular thread of this research: implicit associations and their openness to change. To understand the possibilities and limits of change is profoundly important for understanding human minds in social context and for serving as a foundation for the discussion in this volume about how difference is to be attended to within democratic societies. Social policies may seek to address the many divisions and segments of any society, but the work is incomplete unless we consider the mental barriers to equal treatment. Our intent is to lay bare what we know today in order to

[1] By behaviorally, we mean any action outside of neural and physiological responses. By that definition, survey responses and responses on reaction time tasks are behaviors.

provoke a discussion about the standards we set for ourselves when we say that we value providing equal access and fair treatment.

2. The Implicit Association Test

The most commonly used method to measure implicit group-based cognition is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1988). In the Race IAT, participants begin by seeing images appear on the screen. Some images represent the faces of people of African origin, whereas other faces represent the faces of people of European origin. The participant's task is to sort faces into one of two categories using a key on a standard computer keyboard: "Black people" or "White people." Then, the participant learns to classify words into two categories, 'good' or 'bad' words also using the same two keys on the keyboard. The two tasks after that require combining race categories and valence attributes (See Figure 2). Participants complete combined blocks of trials where they are presented with all four types of images and words (Black faces, White faces, bad words, and good words). In one set of blocks, participants must mentally associate White + Good and Black + Bad by using a common key to respond to the pair. In a different set of blocks, participants must make the opposite association of White + Bad and Black + Good. A measure of ease or effort is obtained by measuring the relative speed with which the two types of pairings are made. In mathematical terms, a score for the Race IAT is computed with the following equation (RT = Reaction Time):

$$\frac{Mean(\text{RT pairing White \& Bad/Black \& Good }) - Mean(\text{RT pairing White \& Good/Black \& Bad})}{Standard\ Deviation(\text{Overall RT})}$$

The idea that the speed of a response tells us about the strength of mental association is non-controversial, beginning with the insight that reaction time could be used as a measure of mental processes from the Dutch scientist Franciscus Donders (1868). Two concepts that are closely associated together in the mind are responded to more quickly than two concepts that are less associated together. To the extent that there is a stronger association between White + Good / Black + Bad than White + Bad / Black + Good, the test reveals faster response to the former pairings than the latter. A majority of White Americans are faster in pairing White + Good / Black + Bad, suggesting that White Americans tend to hold relatively more pro-White/anti-Black implicit preferences (Nosek et al., 2002, 2007).

The IAT's methodological limitations have been well-documented. A commonly observed challenge is the possibility that the order in which blocks are administered may matter. That is, people who are first assigned to associate White + Good and Black + Bad may perform differently on the IAT than people are first assigned to associate White + Bad and Black + Good. To address this possibility, IAT standard practice is to randomly assign people to receive one order or the other so that it can be statistically controlled for (Nosek, Greenwald, & Banaji, 2005). Another common concern is whether the images and words used to represent the categories may bias the results. To address this, why researchers often take care to use images and words that are clear and representative of the categories that seek to represent (Nosek et al., 2005). Over ten potential constraints to the IAT's validity have been identified, and many of these have been addressed through methodological refinements or are routinely taken into account in the interpretation of results (see Nosek, Greenwald, & Banaji, 2007 for a full review).

The IAT has been used in hundreds of psychological laboratories as well as having been available online since 1998 (see Nosek et al., 2002; implicit.harvard.edu). From these data, we know that implicit biases are pervasive and often larger in magnitude than those observed on self-report measures (Nosek et al., 2007). In the aggregate, visitors to Project Implicit show implicit preferences in favor of culturally dominant groups. In addition to showing a pro-White/anti-Black implicit preference, participants show implicit preferences for the concept *straight* rather than *gay*, *thin* rather than *overweight*, and *young* over *elderly*. Data also show implicit stereotypes associating *career* with *men* and *home* with *women* and *European American* with the concept *American* more so than *Native American* with *American*. Data also show associations of *Black American* with *danger* and *White American* with *safety* (Correll, Park, Judd, & Wittenbrink, 2007; Donders, Correll, & Wittenbrink, 2008).

We know that although there is a dissociation between consciously expressed attitudes and the data on the IAT, the two measures are nevertheless correlated. Implicit and explicit racial attitudes are moderately correlated at $r = .31$ (Cohen, 1992; Nosek et al., 2007). At the same time, they are psychologically distinct from each other with each type of attitude accounting for unique statistical variance (Cunningham, Nezlek, & Banaji, 2004).

Sethi (this volume) suggests that such implicit stereotypes might play a role in how criminal offenders select victims and in issues of how the law is enforced. He extends the argument about stereotypes to explain how they create incentive structures and how these incentive structures create larger social patterns. Sethi's argument, in other words, and Loury's argument in *The Anatomy of Racial Inequality* (2002), shows how mental barriers to equal treatment can come to be linked to the divisions in society that social policy often seeks to address.

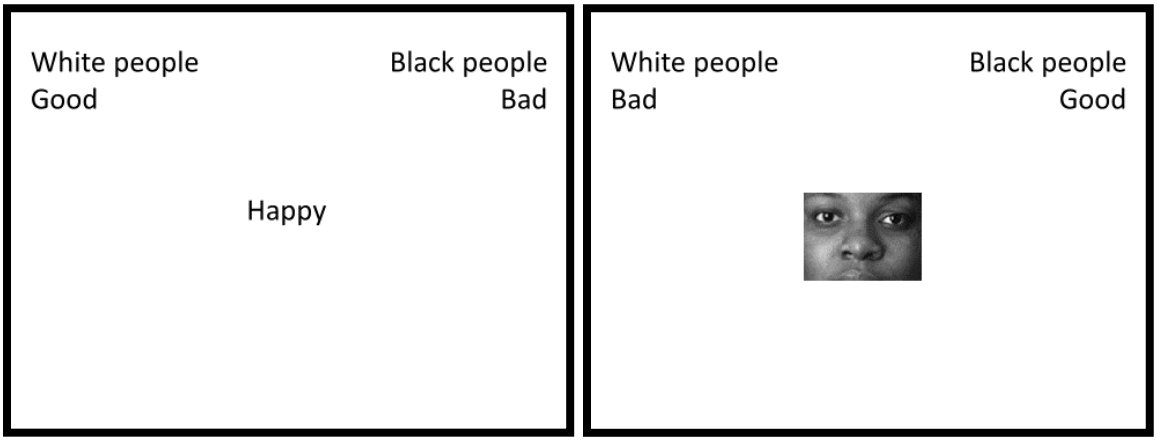

Figure 2. The Race Implicit Association Test.

What are the factors that seem to predict preference for one group or another? For many decades, psychological research had shown the power of group membership in group attitudes and beliefs. We like our own kind, we believe our own group to be superior to others, and we are more willing to justify advantages to our own group over other groups. It is received wisdom

that ingroup favoritism is ubiquitous and accounts for a large share of one's group preferences. But recognition that not all groups show equally robust ingroup preference is not a new idea (see Jost & Banaji, 1994; Jost, Banaji, & Nosek, 2004). Groups that sit higher in the social hierarchy tend to show greater ingroup-favoring attitudes and beliefs than those who sit lower in the social hierarchy.

Research on implicit social cognition has tested the contributions of group membership and recognition of social hierarchy. From a simple *ingroup favoritism* account, members of two different groups, say Black and White people, should show equal and opposite preferences for their own group over the other. For instance, White participants are expected to show pro-White preference and Black participants are expected to show equal and opposite pro-Black preference because such preferences reflect in-group favoring attitudes. In contrast, third-party participants such as Asians should show neutrality on a Black-White Race IAT because they do not belong to either group. From a *social standing* account, the test should reflect the degree to which a group has greater standing in society. This account would suggest that all groups would show a stronger pro-White implicit bias, reflecting the existence of higher status of White Americans in American society.

Examining the data can unpack the relative weight of both accounts. From looking at White participants at Project Implicit (total N = 1,716,521), 73% show a pro-White implicit preference, 16% show no overall preference, and 11% show a pro-Black preference, which is consistent with both *ingroup preference* and *social standing* accounts (Nosek et al., 2007; Xu, Nosek, & Greenwald, 2014).

To get better traction on the relative contribution of these accounts, we can look at the data of Black participants (total N = 279,612). If implicit preference reflects ingroup membership, then the data from Black participants should be the mirror image of that of White participants with 73% of Black participants showing pro-Black implicit preference and 11% showing pro-White implicit preference. In contrast, if implicit preference reflects social standing, then the data should be near-identical to White participants with 75% showing pro-White implicit preference and 10% showing pro-Black preference. The actual results show the presence of both factors on implicit attitudes. 40% of Black participants show pro-Black implicit preference, 35% of Black participants show pro-White implicit preference, and 25% show no overall preference. Although Black participants show slightly more preference for their group than do White participants, the striking result here is the lack of the sort of ingroup preference observed in the data from White participants. Knowledge of Blacks' lower standing in society is visible in the implicit preferences of Black participants themselves (despite Black participants self-reporting robust preferences for their own group).

Asian participants (total N = 135,338) also show pro-White preference almost at the level of White participants. 69% show a pro-White implicit preference, 13% show a pro-Black implicit preference, and 18% show no overall preference. These results again show the importance of social standing in implicit attitudes. Theoretically, Asians could have shown neutrality on the White-Black test of implicit attitude, but they did not. Together, results from these three groups suggest that two factors account for implicit preferences: *ingroup favoritism* is visible in the data of White and Black participants; *social standing* is visible in the far lower ingroup preference of Black Americans and the lack of neutrality among Asians. This simultaneous contribution of ingroup favoritism and social standing is evident on other tests of

implicit bias as well. For example, Christian Americans show stronger implicit preference for their own religion compared with Jewish Americans and Muslim Americans, and straight individuals show stronger preferences for straight people than gay individuals do for gay people (Axt, Ebersole, & Nosek, 2014; Nosek et al., 2007; Westgate, Riskind, & Nosek, 2015).

The signature result from this research is the gap between what we explicitly report when asked about our feelings toward various social groups and what is observed on such tests of implicit attitudes. White participants report less explicit ingroup preference than they show on measures on implicit preference, whereas Black participants report more ingroup preference on survey measures than they reveal on measures of implicit preference. From such data, investigators have concluded that self-reported attitudes are products of reflection to a greater extent whereas implicit attitudes are products of impulse to a greater extent. When the two are not in sync, implicit and explicit attitudes reveal a psychological dissociation – a schism between two parts of the same mind.

3. Implicit Bias and Behavior

Measures of implicit cognition have been used to predict behavior in many domains, from racial discrimination to mental health to consumer preferences (Cameron, Brown-Iannuzzi, & Payne, 2012; Carlsson & Agerstrom, 2016; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald et al., 2013). Although most studies measuring the link between implicit cognition and behavior are laboratory based, interest has also focused on whether measures of implicit cognition can predict behavior in naturalistic settings. As an example of such a study, researchers evaluated how cashiers interacted with managers in a chain of French grocery stores (Glover, Pallais, & Pariente, 2016). Managers in this study took an IAT assessing the degree to which they associated Europeans and North Africans with being good or poor employees. The researchers also measured the work performance of majority White and minority North African employees. On average, minority and majority cashiers performed about equally well. However, this differed based on who the presiding manager was during a particular work shift. When the presiding manager's anti-minority bias was low, the African cashiers performed better than White cashiers. When the presiding manager's anti-minority bias was high, African cashiers performed worse than White cashiers. African cashiers (but not White cashiers) who were scheduled to work when the manager on duty had high anti-minority bias were more likely to be absent from work, left work earlier, scanned items at the counter more slowly, and took more time between customers.

These racial biases in employee management extend to hiring. In a field experiment, researchers sent over 1,500 resumes to positions listed in job ads in Sweden (Rooth, 2007). These resumes were matched to be almost identical except for two words: the name of the applicant, which signaled whether the applicant was Swedish or Arab. The study found that simply having an Arabic-sounding name reduced the probability of an interview callback by 9%. Moreover, implicit racial/ethnic stereotypes were predictive of biases in the callback rate even though explicit racial/ethnic attitudes and stereotypes were not.

Implicit biases are also implicated in critical life-or-death decisions to shoot (or not shoot) criminal suspects. In the one series of studies, participants played a simulated game by observing images of White and Black men situated in everyday places like parks or city sidewalks. Some of the men are armed with guns, while others are unarmed and carrying

mundane objects like wallets or cell phones. Participants are instructed to press a button to "shoot" (if the man is holding a gun) or press another button to "don't shoot" (if the man is not holding a gun). Research on civilian participants finds that they tend to mistakenly shoot unarmed Black targets more than unarmed White targets (Correll, Park, Judd, & Wittenbrink, 2002). Similar research on police officers finds that the racial shooting bias depends on the type of experiences that police officers have, such as whether they primarily interact with members of the community or gang members (Correll et al., 2007; Sim, Correll, & Sadler, 2013), fatigue (Ma et al., 2013), and aspects of the interaction (James, Vila, & Daratha, 2013).

Implicit cognition also predicts how we make sense of political information. One study examined political partisans and independents within the United States (Hawkins & Nosek, 2012). Participants read about two competing special education policies, one proposed by the Democrats, the other by Republicans. Political independents were found to be mixed in their support of these policies, with some preferring the Democrat version and others the Republican one. However, this mixed support did not indicate an objective appraisal of policy facts. Independents who implicitly identified with Democrats (Republicans) were more likely to support policies proposed by the Democratic (Republican) party, regardless of policy details. The significance of this result of course rests on the fact that the participants believe themselves to be independent. Implicit cognition may provide a window into patterns of behavior that are missed by explicit endorsements and can reflect hidden partisan affiliations (Dennis, 1988; Keith et al., 1992).

Implicit biases predict how voters seek and consume political media. Undecided voters in Northern Italy took an IAT assessing implicit attitudes toward integrating Turkey into the European Union (EU; Galdi, Gawronski, Arcuri, & Friese, 2012). These implicit attitudes predicted how voters selectively exposed themselves to biased information. Undecided voters who had more pro-integrationist implicit attitudes were more likely to choose to read pro-integrationist news articles, leading to more pro-integrationist conscious attitudes. Meanwhile, undecided voters who had more anti-integrationist implicit attitudes were more likely to choose to read anti-integrationist news articles, leading to more anti-integrationist conscious attitudes.

The impact of implicit biases on discrimination has generated considerable interest in understanding how to combat them (e.g., Sethi, this volume). But how does one do so? In this chapter, we analyze the process of change by focusing on aspects of the situation that can elicit or retard the expression of implicit bias, on the actual implicit associations themselves, and on conscious strategies for self-regulation. These three "locations" – situations, mental associations, and regulation – may serve as a useful way to organize what is currently known about the process of changing biases as they affect the allocation of opportunities and interpersonal dealings (See Figure 3).

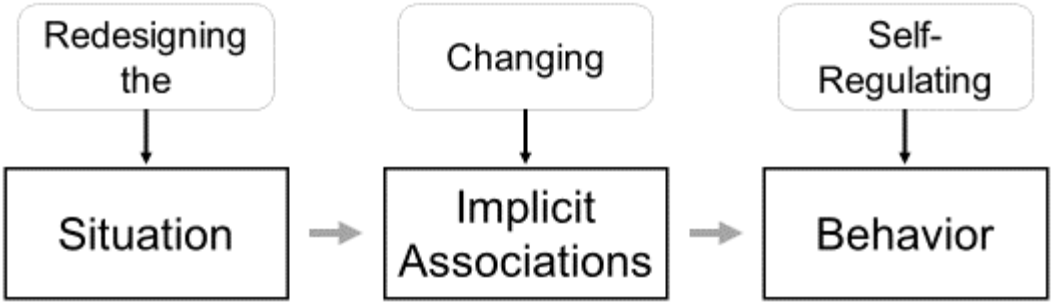

Figure 3. Three "locations" for addressing the impact of implicit bias.

4. Three Approaches to Addressing Implicit Bias

4.1 Redesigning Situations

The concept of an *affordance* is an old chestnut in psychological theorizing, originally proposed by J. J. Gibson (1979) who promoted the idea that specific physical objects, situations, and the environment create specific possibilities for action. Open spaces "afford" walking, a thread "affords" sewing, handles "afford" turning. In other words, particular objects directly drive particular actions. In a loose way, we suggest that social situations "afford" or prompt particular behaviors. Situations can create or demand behavior that relies on use of group-based information to a greater or lesser extent. Here we give two examples of features of situations that can be designed to promote or hinder use of social group information.

Blinding. Many group memberships are signaled to us by visual and auditory cues. In most cases, a person's gender, ethnicity, and age are easily available from visual and auditory perception. Even a person's name can cue us about one's group memberships. As dozens of audit studies in the field show, merely knowing a person's name can lead to group-based discrimination. Studies exist in the domains of hiring (Bertrand & Mullainathan, 2004), financial access and remuneration (Ayres, Vars, & Zakariya, 2005; Ladd, 1998), and healthcare (Kugelmass, 2016) that reveal the role of an individual's group membership on decisions made about them.

To the extent that attributes of a person such as their name, physical body, dress, and speech produce discrimination, it is worth considering when such cues are undesirable and should be removed from perception. As one example, a highly cited case study of symphony orchestras suggests that blind auditions (i.e., of using a curtain to visually block musicians from view of selecting judges) led to the increased hiring of female musicians (Goldin & Rouse, 2000). There is no question that blinding decision-makers to individual demographic characteristics can remove the pernicious and privileging effects of irrelevant information. In modern democracies, many will agree that, if a prospective employee's group membership enters into decision-making about the individual's competence and thwarts two ideals -- the ideal of selecting the best possible employee and the ideal of fair treatment –it is in a society's interest to consider corrective measures.

If we wish to evaluate competently, removing irrelevant information is clearly a good practice. But current practices seem largely unaffected by the evidence. A recent study found that law firms were much more likely to invite male applicants from high-income backgrounds for an interview than equally qualified female applicants or male applicants from low-income backgrounds (Rivera & Tilcsik, 2016). Another recent study shows that African Americans are 16% less likely to be accepted as a renter than White Americans on Airbnb, a short-term housing rental service (Edelman, Luca, & Svirsky, 2017). Such studies suggest that if access to cues for inferring characteristics like a person's gender, race, or social class were hidden from view, evaluators will home in on the criteria that matter most for job performance. Consider the field experiment described in the beginning of the chapter, where CVs with Arabic-sounding names were 9% less likely to get an interview callback than CVs with Swedish-sounding names (Rooth, 2007). To combat subtle biases like these, institutional policies may mandate the removal of names in job applications in some rounds of hiring. In 2010, Germany's Federal Anti-Discrimination Agency conducted such a field experiment testing the effects of procedural blinding (Krause, Rinne, & Zimmermann, 2012). Some companies used anonymized job applications and other companies did not. They found that procedural blinding worked: companies that used anonymized job applications reduced bias against women and migrant applicants in many cases.

As the reader may have already guessed, blinding is possible only in limited circumstances such as grading in the classroom and reading applications for jobs. And even in these circumstances, it is not airtight. Information about one's group membership can indirectly leak through. We take this up shortly in describing the many interactions in which human decision makers cannot avoid a person's group information. Moreover, if one wishes to look at differences in training (two candidates are only slightly apart in skill but the one with slightly less skill has had far less schooling; in such a case it may be smarter to make an argument to select the slightly less good candidate based on potential to soar with training or experience. Such decisions cannot be implemented in blind reviews where differences of all kinds are explicitly obfuscated in order to zoom in on specific attributes.

Reducing subjectivity. When the criteria for making a decision are ambiguous or uncertain, and when the decision requires reliance on one's judgment (e.g., a gut feeling, or a sense of what is expected to work best), biased judgments are to be expected (Darley & Gross, 1983).To understand the extent to which minds will bend to create the outcome that is wished for rather than the outcome that rewards merit, consider a series of studies by Uhlmann and Cohen (2005). Participants were tasked with selecting one of two candidates, Michael and Michelle, for promotion to the position of police chief. All the materials for Michael and Michelle were identical with one difference. One set of participants learned that Michael is known for his practical knowledge (street smart), while Michelle is known for being formally educated (book smart). After reviewing the candidates, participants were more likely to select Michael for promotion, citing his street smarts as essential to success as a police chief.

Another set of participants in the same experiment also evaluated Michael and Michelle, but were assigned the opposite attributes: Michael was formally educated (book smart) and Michelle was the practical one (street smart). If participants were evaluating solely on credentials, then one would expect that these participants would now pick Michelle given that the quality of being street smart was prized. However, that was not the outcome in the

experiment. Participants in this condition were also more likely to pick Michael, citing formal education as essential to success as a police chief. These results demonstrate the slipperiness of decision-making criteria that allow us to hire the candidate we want while believing that we are selecting based on objective criteria for judgment. Participants' notions of what qualities mattered for the position were confabulated in order to confirm their gendered preference in a highly gendered profession.

To combat this tendency, the researchers ran the study again but made a small change to the procedure. Participants reported what qualities were important to being a police chief before selecting a candidate instead of after. When participants pre-committed to a set of criteria, they were more consistent with the qualities they reported as important. Regardless of gender, people who prioritized being streetwise picked the streetwise candidate, and people who prioritized being educated picked the educated candidate. This finding shows that constraining decision-makers to more objective criteria can powerfully reduce bias.

Benefits and risks. The social situation can expand or contract the degree to which biases in individual minds can be expressed. When possible during important decisions, information about irrelevant group memberships should not be present as it can lead individual minds away from fairness and merit-based selection. Blinding does not require active effort from the decision maker. It simply removes distracting information to produce decisions that are more in line with one's conscious values. Once instituted, blinding procedures can become routine and require little effort to maintain.

Redesigning situations can also be effective because they have wide-reaching influence on unwanted biases. First, blinding or objective criteria can counter the effects of social knowledge well beyond the biases of which one is aware. For instance, blinding names in job applications may block evaluators from acting on race and gender biases, but also more subtle biases like those of age, social class, and implicit egotism (i.e., liking a name that shares the same first initial as yours; Pelham, Carvallo, & Jones, 2005; Polman, Pollmann, & Poehlman, 2013).

More broadly, these strategies can prevent general cognitive biases. For example, use of objective criteria can eliminate the possibility of biases that arise from subjective judgment such as confirmation bias (the tendency to interpret evidence as confirming of one's existing beliefs; Nickerson, 1988), framing effects (the tendency to react to information differently based on how it is presented; Tversky & Kahneman, 1985), or primacy and recency effects in memory (the tendency to remember the first and last things in a series better than the things in the middle; Murdock, 1962). These strategies add new tools to the toolbox for assuring fair treatment. Policy-makers can identify situations where bias might interfere, and introduce procedural changes to prevent bias (Thaler & Sunstein, 2009).

Despite their advantages, blinding and removing subjectivity from selection suffer from several limitations. As noted, blinding is simply not possible in the many situations that demand face-to-face interactions. Even if an individual actor wishes to de-bias herself by eliminating group membership information, such knowledge can intrude through the many channels that constitute the stream of information: voice, photos, and hearsay.

Implicit biases emerge from cues that are subtle and of which the perceiver is unaware. Having the foresight to anticipate potentially biasing behavior and implementing a fix can be a demanding requirement. Among these issues is the possibility that blinding increases failure to take into account potential at the time of decision making as opposed to pre-existing accomplishment. In the field experiment described above (Krause et al., 2012), some cases of blinding led to *increased* bias against women and/or migrants (see also Behaghel, Crépon, & Le Babanchon, 2014). In those cases, increased bias may result from removing information about demographics, when that prevents the possibility of enacting compensatory measures like affirmative action or undertaking forms of holistic review that take prior opportunities into account,in order to assess potential. Thankfully, this a tractable issue. If institutions seek to employ blinding but also desire to increase diversity within their institution, then blinding could be supplemented with counterweights that promote diversity-related outcomes. For example, a technology start-up that is composed of 80% men could choose to include applicant gender as a criterion in hiring when constructing objective hiring criteria. That would increase the probability of women being hired, all else equal.

4.2 Changing Implicit Associations

The human mind has acquired distinctive ways of processing information over the course of evolutionary time. The mind has developed specific ways of attending to the world and perceiving and forming inferences about the world based on the partial information provided by experience. Our effort is directed toward what we might do, given these features of the mind, in the context of the values that characterize democracies today. Based on the evidence we reviewed early in this chapter on the presence of implicit group-based associations that exist without conscious animus, an obvious location to conceive of change is in status of those associations themselves.

The logic goes something like this: We form good habits all the time. We didn't wear seat belts but now we do. We didn't pay attention to the threat from smoking but now we do. We learned to brush our teeth each morning and night even if we are incredibly tired. The list of human achievements in the small that keep us healthy are similar to the many changes we have made to our social lives in which we obey ordinary rules of conduct – such as who has right of way on the road, to courtesies we practice all the time – of standing in lines, giving people time to make an argument, and so on. The brains of humans today who routinely perform these actions are no different than their ancestors who did not, but our minds are unrecognizably different.

Psychology as a discipline placed the concept of "learning" at the center of the field early in its history and we have an enormous amount of knowledge about how we learn and change. If implicit associations are in part a reflection of the world in which we live and if these associations are also likely to be the cause of how we act, it is worth attempting to understand how and when changing implicit associations is possible. By reducing the very activation of particular implicit associations, the influence of bias might be reduced (Lai, Hoffman, & Nosek, 2013). Unlike explicit attitudes (e.g., stating that one doesn't like members of group X), implicit attitudes and beliefs have a unique pattern of learning that needs specific attention (Gawronski & Bodenhausen, 2006; 2011).

One such theory of learning suggests that implicit biases reflect the sheer accumulation of associations between stimuli in the environment (Rydell & McConnell, 2006). A direct method to change implicit biases, then, is exposure to stimuli that counter existing associations. Just as with learning that emerges from classical conditioning (Bouton, 2007; Pavlov, 1927; Staats & Staats, 1958), interventions present participants with dozens or hundreds of new pairings between a concept (like race or gender) and attributes (like good/bad, strong/weak; e.g., Bar-Anan, De Houwer, & Nosek, 2010; De Houwer, Thomas, & Baeyens, 2001; Karpinski & Hilton, 2001; Olson & Fazio, 2001). Interestingly, these conditioning trials often change implicit attitudes without affecting explicit attitudes and give credence to the view that implicit associations arise from distinct mental processes that operate by a unique sets of rules. For example, Olson and Fazio (2006) presented participants with positive images and words paired with Black faces and negative images and words paired with White faces. Exposure to these pairings reduced implicit racial attitudes immediately, and this effect persisted two days later. In contrast, explicit racial attitudes remained unchanged.

A more elaborate version of presenting new associations to change existing attitudes is to create encounters with people or media images that defy stereotypes. Mere exposure to well-liked Black people (e.g., Nelson Mandela, Michael Jordan) and disliked White people (e.g., Charles Manson, John Gotti) can reduce the typical pro-White/anti-Black bias (Dasgupta & Greenwald, 2001; Joy-Gaba & Nosek, 2010). Even imagining a counterstereotypical person can reduce implicit stereotyping, as was the case when college students were asked to imagine working for a woman as their boss even for a few minutes (Blair, Ma, & Lenton, 2001). Changing implicit associations through exposure to counterstereotypical information is one method of many. A recent meta-analysis (Forscher, Lai, et al., 2018) compared 492 experiments that study implicit bias change, of which 313 tested experimental manipulations on topics relevant to intergroup relations. Figure 4 shows a forest plot synthesizing the results of these 313 intergroup studies across 11 approaches.

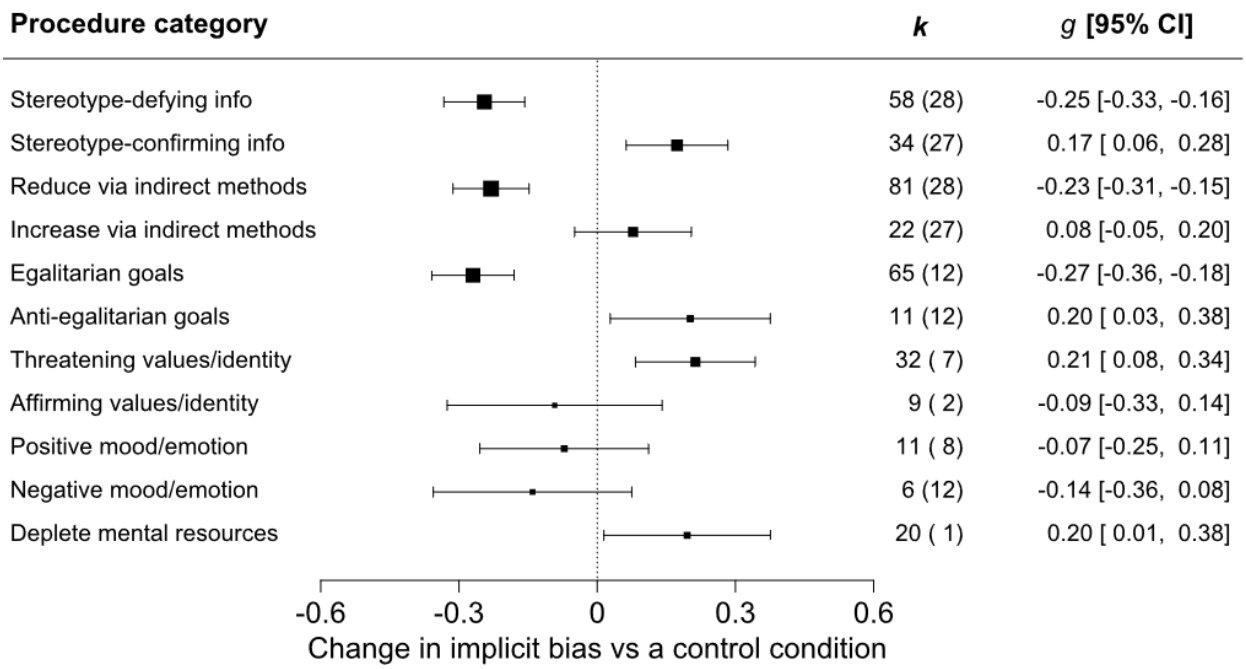

Figure 4. Forest plot of comparisons between each approach and a control condition on implicit intergroup bias. *k* gives the number of study samples that directly (or indirectly, listed in parentheses) compare the approach and a control condition. *g* gives the standardized mean difference between an approach and a control condition and its 95% confidence interval. Larger effect sizes reflect greater increases in implicit bias relative to a control condition. See Forscher, Lai, et al. 2018 for a complete description of the methods and analyses.

Overall, changes in implicit associations were small but robust. However, the magnitude of change depended heavily on the type of approach used to change implicit intergroup biases. An initial set of approaches involved experiences with stereotype-defying information (like the above) or stereotype-confirming information. These two approaches were effective at reducing and increasing implicit associations, respectively. A second class of approaches sought to change implicit associations indirectly by targeting higher-level mental states and beliefs that are not directly related to the implicit association. For example, taking the perspective of a member of another group or being led to feel powerful or high-status (Guinote, Willis, & Martellotta, 2010; Richeson & Ambady, 2001). Indirect approaches that sought to reduce implicit associations were effective at doing so, although indirect approaches that sought to increase implicit biases were not consistent at doing so. These results suggest that implicit biases are not merely the product of brute force associations. They can also be changed by higher-level mental states and beliefs.

A third class of approaches activated egalitarian or anti-egalitarian motivations or goals. Egalitarian motivations or goals were effective at reducing implicit biases and anti-egalitarian motivations or goals were effective at increasing them, suggesting that implicit biases are sensitive to the motivations or goals that individuals carry with them into social situations (e.g., Lun, Sinclair, Whitchurch, & Glenn, 2007; Stewart & Payne, 2008). A fourth class of approaches involved threatening or affirming ones' values or identities. Echoing decades of work on intergroup threat (Stephan & Stephan, 2000), threatening one's values or identity was effective at increasing implicit bias against members of other groups (e.g., Frantz et al., 2004). In sharp contrast, affirming one's values or identities did not have an overall effect on implicit intergroup biases (e.g., Rudman, Dohn, & Fairchild, 2007; Walton et al., 2015).

A final set of approaches looked at general mental states. Emotions and moods such as happiness or sadness did not have an overall effect in changing implicit intergroup biases (e.g., Kuppens et al., 2012). However, depleting mental resources did. In these studies, participants were led to feel mental fatigue or mental load and this mental state increased the expression of implicit biases. These studies suggest that tiredness, stress, or other mental states that constrain one's ability or motivation to think things through can enhance the effects of implicit bias.

These results are important. They teach us that implicit attitudes and beliefs are malleable. Even well-known and well-practiced habits of mind are able to stretch in opposite directions with brief and minimal interventions. It is illustrative that a few minutes of a variety of minor interventions can produce a reappraisal after a lifetime of learning that one group is superior to another.

Long-term change in implicit bias. Studies like these have raised the question of whether implicit attitudes are amenable to long-term change. Unfortunately, most studies use short-term interventions that examine change over the course of a single experimental session lasting an

hour or so. Only 18 out of the 313 experiments described above examined effects over a longer period, such as several days or weeks.

In a recent series of studies, we sought to address this lack with experiments that compared the impact of eighteen brief interventions to reduce implicit racial attitudes (Lai et al., 2014, 2016). Of these eighteen, nine were effective at reducing implicit attitudes immediately. The most effective intervention leveraged exposure to counterstereotypical exemplars to cut implicit preferences by half.

However, none of these nine interventions had persistent effects on implicit preferences after several days. Post intervention, participants returned to baseline levels of implicit preferences. These results are not necessarily surprising. Human minds have evolved to change in response to experiences that are potent, adaptive, or repeatedly practiced. The interventions used in the studies lacked many of these properties. They were mild, brief, and unlikely to be sustained in environments that routinely confirm existing biases. Being sensitive to context but able to snap back to default ways of thinking is exactly what an intelligent system should do. And this of course suggests just how difficult the project of changing attitudes and beliefs is. To understand how permanent changes in implicit biases can happen, one must investigate research that undertakes changes in implicit bias after subjecting it to robust intervention.

Long-term experiences with people outside of one's own group is one of the most-studied and effective approaches to reducing prejudice, both implicit and explicit (Allport, 1954; Pettigrew & Tropp, 2006). In one study, researchers took advantage of a natural experiment where White college freshmen were randomly assigned to a Black or White roommate and found that freshmen that lived with a Black roommate showed reduced implicit racial bias at the end of the first semester of living together (Shook & Fazio, 2008). Another study looked at gender stereotypes by studying how going to a women's single sex college related to gender stereotypes (Dasgupta & Asgari, 2004). They assessed female students during the beginning of their first and second years of college, half of whom attended a women's liberal arts college, and half of whom attended a coeducational liberal arts college. At the start of college, students in both colleges showed the same level of gender stereotypes associating men more with leadership than women. Over the course of their first year, the implicit gender-leader stereotypes of women at the women's college had dissipated, whereas the implicit stereotypes of women at the coeducational college had increased. Interestingly, this difference in stereotyping was explained not by the single sex versus coeducational structure of the institution, but rather by the extent to which students had been exposed to female faculty members – the greater the number of courses taken with female faculty the lower the gender-leader bias. Having greater exposure to female professors, an event more likely at the women's college, had shifted implicit gender-leader stereotypes away from the cultural norm.

Another way to assess the impact of long-term experiences is to track changes at the cultural level. A deep analysis of attitude IATs at the Project Implicit website was recently undertaken (Charlesworth & Banaji, 2019) using 4.4 million tests of implicit and explicit attitudes that were measured continuously from 2007-2016. Comparative analysis using time series models examined whether the U.S. has changed in attitudes toward six social groups: age, disability, race, skin-tone, sexual orientation, and body weight. Within that decade, all explicit attitudes show change toward neutrality. Implicit attitudes have also changed toward neutrality in three domains: implicit preferences for straight vs. gay people (33%), implicit preferences for

White vs. Black people (17%), and implicit preferences for light-skinned vs. dark-skinned people (15%). In contrast, other implicit attitudes remain unchanged (age, disability) or even moved away from neutrality (body weight). These data show that implicit attitude change is possible. We cannot isolate the causes of why sexuality, race, and skin tone attitudes have changed toward neutrality, but we speculate that their change may be related to the intense social dialogue that has surrounded questions of sexuality and race relative to other social categories in the U.S. in the early decades of the 21st century. Some initial evidence suggests that at least for race, reductions in implicit racial preferences were associated with the rise of the Black Lives Matter movement (Sawyer & Gampa, 2018). This suggests that active political discussion, disagreeable as the discussion may be, has had the effect of reducing implicit bias.

In this volume, Sethi also presents a case of a durable intervention resulting from a natural experiment—the increasing rates of the incarceration of women—and argues that it has made a durable difference for explicit attitudes about punishment.

These long-term studies of implicit attitudes provide insight into the kind of interventions that are likely to change and not change implicit attitudes. Although such comparisons do not provide a clear answer as to why some implicit attitudes change and others remain stable, they do allow us to hypothesize about the conditions that are necessary for change to occur.

Benefits and risks. If change in anti-gay bias is an example of the change that is possible even with robust negative attitudes, the power of media exposure, personal contact, and legal progress cannot be underestimated as vehicles of change. As psychologists, we are interested in change at the level of the individual mind. We are aware that one can choose to create the conditions of social life that will produce shifts in attitudes and beliefs. The path we walk everyday can be chosen rather than passively accepted. If the will exists, one can shape experiences in such a way that positive associations to otherwise negative social categories can be created through media exposure and personal experience. The power of counterstereotypes is not to be underestimated and if counterstereotypical encounters become typical, shift in attitudes and beliefs will follow. What has been learned, can be unlearned or learned in a new mold. But this is neither easy to implement nor something that is achieved by a few experiences. The stuff of life – where one lives, who one encounters and in what capacity, the ones who are the objects of our intimacy, and the representations of them we choose to present to our senses – all these phenomena contribute to the formation and re-formation of attitudes and beliefs and to some extent, they are within our own control.

Although possible, to expect that individuals will change their habits and their experiences at the level of minuscule threads of daily experience and to do so on a large scale is unreasonably optimistic. There are few incentives to do so, there are few "manuals" that show the way, and there are many daily experiences that oppose change. Existing group memberships, such as religious, political, and ethnic identities, are viewed around the world as sacred in nature and where explicit negative attitudes towards those from another group are common. This being the case, it is important to consider the role of mechanisms other than individuals volunteering to change individual minds. In many social sciences that are not as focused on the individual as psychology is, the idea of measuring progress and creating change has always been considered at the more macro level of groups, institutions, laws, and governance. As we consider the evidence on implicit cognition, it is clear that if individuals are largely unaware of their own mental states

and of the fact that their behavior is not in line with their own intentions and values, that levers of change other than individual minds must be activated.

Institutions in democracies have the power to shape environments to elicit the attitudes and beliefs their own citizens aspire to but fall short on. Institutions that control housing, education, business, healthcare, and the law can take up the challenge of creating implicit bias-reducing experiences that a society willingly acknowledges to be in the public interest. This focus on institutions becomes a necessity when such an aspiration is combined with knowledge of the deep bounds of the rationality of individual decision making. If a society believes that all its citizens deserve equal opportunity, then institutions may fill in the gaps created by the limits of individual decision-making. Achieving difference without domination may require us to activate the representational power latent in our institutions themselves.

4.3 Self-Regulating Behavior

In the animal kingdom, certain species dominate. The gargantuan blue whale eats up to 2.2 tons of food each day. The nimble cheetah can run at 70 miles per hour. The sage tortoise can outlast the lives of humans many times over. And yet, humans are the species they are because we have the gift of conscious thought. Among the functions of conscious thought is the ability to exert control over behavior, even if the mind has an opposing preference. Humans are champions of holding their tongue, suppressing maladaptive emotions, and delaying gratification to seek a greater later reward. Our ability to self-regulate flexibly exceeds that of any other species. In the domain of bias reduction, the Self-Regulation of Prejudice (SRP) Model has been proposed (Monteith, 1993; Monteith et al., 2002; Monteith & Mark, 2005; 2009; Monteith, Parker, & Burns, 2016) to explore this phenomenon. The model elucidates several consequences of becoming aware that one did something biased. These consequences are affective: the experience of negative emotions such as guilt or shame; behavioral: the inhibition of the biased behavior; and cognitive: reflection on the situation and the mental states associated with biased thought or behavior. A combination of these is expected to lead to the development of mental control. Practice with such routines creates learning that leads to inhibition of biased behavior, prospective reflection on how to behave differently, and the achievement of non-biased behavior.

Of course, self-regulation is never fool-proof. Failures to control one's own biases can arise at any point. Social actors may fail to realize that they are acting in a biased way to begin with. Or, social actors may fail to prevent themselves from acting on their biases, fail to experience the negative emotions about being biased, or fail to reflect on why they may be biased. Social actors may also fail to establish mental control or fail to be remember to enact mental control in a future event where they are potentially biased. To counteract these many self-regulation failures, we propose two evidence-based strategies: looking into the bias blind spot and implementation intentions.

Looking into the bias blind spot. A near-necessary condition for bias self-regulation is awareness that one's own behavior is biased. This sounds simpler than it is. The singular psychological obstacle to awareness is the inability to introspect accurately about the origins of one's own thoughts and feelings (Nisbett & Wilson, 1977). This tendency to fail at seeing bias in one's own judgments despite being able to perceive bias in others' judgments is known as the *bias blind spot* (Pronin, Lin, & Ross, 2002). People can consciously search their own thoughts for biased processing, but very rarely find any indication of it (Pronin & Kugler, 2007).

Just as humans cannot introspect accurately about their level of blood pressure or a developing cancer (and rational individuals must rely on external methods to learn about the workings of their bodies), so too with hidden mental content. One can imagine that efforts to educate about the existence of implicit bias and how it influences behavior may translate into successful regulation. Of course, such efforts to educate rely on the assumption that many individuals in modern democracies will aspire to behavior change if they are made aware that it is unaligned with their values and aspirations.

Implicit bias is most likely to influence behavior when individuals lack the motivation or ability to deliberate (Fazio & Olson, 2014). The motivation to deliberate can arise from motivations to be accurate (e.g., Schuette & Fazio, 1995), being accountable for a behavior (e.g., Sanbonmatsu & Fazio, 1990), concerns about being viewed favorably (e.g., Fazio, Jackson, Dunton, & Williams, 1995), or motivations to act without prejudice (e.g., Dunton & Fazio, 1997). The ability to deliberate can be impeded by factors such as mental fatigue (Govorun & Payne, 2006), time pressure (Ranganath, Smith, & Nosek, 2008), alcohol use (Bartholow, Dickter, & Sestir, 2006) or age-based cognitive decline (Gonsalkorale, Sherman, & Klauer, 2009). Greater self-insight into when one is most likely to act on implicit bias can combat these influences. A major caveat to this recommendation is the dearth of evidence examining the effectiveness of raising awareness about unconscious bias. We are aware of only one experiment that has done so. It found that a 2.5 hour workshop for university faculty raising awareness about gender bias led to increased self-efficacy in combating gender bias (Carnes et al., 2015).

Implementation intentions. Awareness of biased behavior does not eliminate the biased behavior in question. One must also act to replace it with non-biased behavior. This can be difficult in practice. A surprisingly effective strategy to bridge this gap between values and action is implementation intentions (Fujita, 2011; Gollwitzer, 1999; Gollwitzer & Sheeran, 2006). Implementation intentions are "if-then" plans that link a situational cue with a behavioral response (e.g., "If event X happens, then I will do Y."). Setting an if-then plan creates a mental association between the cue and response, making effortful behavior more automatic and unconscious (Bayer, Achtziger, Gollwitzer, & Moskowitz, 2009; Gollwitzer & Brandstätter, 1997). Despite its simplicity, implementation intentions have improved outcomes across many types of self-regulation problems, including exercise and dieting (e.g., Luszczynska, Sobczyk, & Abraham, 2007), recycling (Holland, Aarts, & Langendam, 2006), and smoking cessation (Armitage, 2007). Importantly, they have also successfully reduced the expression of implicit biases (Lai et al., 2014, 2016; Mendoza, Gollwitzer, & Amodio, 2010, Stewart & Payne, 2008). Although it is difficult to imagine that the strategy would succeed, research has shown setting an "if-then" implementation intention to think a counter-stereotypical thought when they see a Black person reduced the expression of implicit bias for at least two months after the implementation intention was taught (Monteith et al., 2016).

Benefits and risks. Self-regulation strategies are flexible and can be tailored to many areas of bias. Compared to efforts to reduce implicit bias, efforts to self-regulate are more likely to result in effective behavior change because the behavior and self-control are more closely linked. Efforts to promote bias regulation give agency to individuals who seek to treat people of different groups fairly, although they may be avoided individuals who do not share such commitments (Kulik, Pepper, Roberson, & Parker, 2007). In the policy realm, education initiatives could be used to inform individuals about approaches for successful self-regulation.

Public campaigns, school curricula, and other social media can orient people to simple tips for controlling unwanted behavior such as implementation intentions.

Conclusion

The human mind is limited in its ability to apply the ideals of freedom, opportunity and fairness equally to all, or to treat all with dignity, as Wingo also selaborates in this volume. This limitation must be recognized in order to enter into any discussion of creating a better society. Evidence of implicit bias has raised the bar on the challenges faced by modern democracies consisting of a plurality of social groups with differing histories, power, and potential futures. Understanding that discrimination is possible without an intention to harm is difficult to grasp and even harder to solve given the presence of legal systems founded on the idea of intent as pivotal in determining justice. However, recent discoveries on the possibility of addressing the pernicious consequences of implicit bias show that what may seem to be inevitable effects of implicit bias need not be so. The research we have reviewed shows individual minds to be sensitive to change given the right inputs. We hope that this approach to securing positive social change can aid in the project of safeguarding diverse societies.

References

Allport, G. W. (1954). *The nature of prejudice.* Cambridge, MA: Addison-Wesley.

Armitage, C. J. (2007). Efficacy of a brief worksite intervention to reduce smoking: The roles of behavioral and implementation intentions. *Journal of Occupational Health Psychology, 12*, 376-390.

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*, 1423-1440.

Axt, J. R., Ebersole, C. R. & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science, 25*, 1804-1815.

Ayres, I., Vars, F. E., & Zakariya, N. (2005). To insure prejudice: Racial disparities in taxicab tipping. *Yale Law Journal, 114*, 1613-1674.

Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people.* New York, NY: Random House.

Bar-Anan, Y., De Houwer, J., & Nosek, B. A. (2010). Evaluative conditioning and conscious knowledge of contingencies: A correlational investigation with large samples. *The Quarterly Journal of Experimental Psychology, 63*, 2313-2335

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven implicit measures of social cognition. *Behavior Research Methods, 46*, 668-688.

Bartholow, B. D., Dickter, C. L., & Sestir, M. A. (2006). Stereotype activation and control of race bias: cognitive control of inhibition and its impairment by alcohol. *Journal of Personality and Social Psychology, 90*, 272-287.

Bayer, U. C., Achtziger, A., Gollwitzer, P. M. & Moskowitz, G. (2009). Responding to subliminal cues: Do if-then plans facilitate action preparation and initiation without conscious intent? *Social Cognition, 27*, 183-201.

Behaghel, L., Crépon, B., & Barbanchon, T. L. (2015). Unintended effects of anonymous resumes. *American Economic Journal: Applied Economics*, *7*, 1-27.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review*, *94*, 991-1013.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology, 81*, 828-841.

Bouton, M. E. (2007). *Learning and behavior: A contemporary synthesis.* Sunderland, MA: Sinauer Associates Inc.

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

Carlsson, R., & Agerström, J. (2016). A closer look at the discrimination outcomes in the IAT literature. *Scandinavian Journal of Psychology*, 57, 278-287.

Carnes, M., Devine, P. G., Manwell, L. B., Byars-Winston, A., Fine, E., Ford, C. E., ... & Palta, M. (2015). Effect of an intervention to break the gender bias habit for faculty at one institution: A cluster randomized, controlled trial. *Academic Medicine, 90*, 221-230.

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science.*

Consolidated Brief of Lt. Gen. Julius W. Becton Jr. et al. as Amici Curiae in Support of Respondents, in *Grutter v. Bollinger*, et al., *Gratz & Hamacher v. Bollinger*, et al., Nos. 02-241, 02-516 February 19, 2003.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*, 1314-1329.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2007). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology, 37*, 1102-1117.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S. & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92*, 1006-1023.

Cunningham, W. A., Nezlek, J. B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin, 30,* 1332-1346.

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

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology, 40,* 642-658.

Dasgupta, N., & Greenwald, A.G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81,* 800-814.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*, 853-869.

Dennis, J. (1988). Political independence in America, Part I: On being an independent partisan supporter. *British Journal of Political Science, 18*, 77-109.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5-18.

Donders, F. C. (1868). On the speed of mental processes. Translated by W. G. Koster, 1969. *Acta Psychologica, 30*, 412-431.

Donders, N. C., Correll, J., & Wittenbrink, B. (2008). Danger stereotypes predict racially biased attentional allocation. *Journal of Experimental Social Psychology, 44*, 1328-1333.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*, 316-326.

Edelman, B. G., Luca, M., and Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*.

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. (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*. New York, NY: Guilford.

Forscher, P. S., Lai, C. K.**,** Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2018). A meta-analysis of change in implicit bias. Unpublished manuscript.

Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the computer: The race Implicit Association Test as a stereotype threat experience. *Personality and Social Psychology Bulletin, 30*, 1611-1624.

Fujita, K. (2011). On conceptualizing self-control as more than the effortful inhibition of impulses. *Personality and Social Psychology Review, 15*, 352-366.

Galdi, S., Gawronski, B., Arcuri, L., & Friese, M. (2012). Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious beliefs. *Personality and Social Psychology Bulletin, 38*, 559-569.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692-731.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44*, 59-127.

Gawronski, & B.K. Payne (Eds.). (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, NY: Guilford Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gilbert, G. M. (1951). Stereotype persistence and change among college students. *The Journal of Abnormal and Social Psychology, 46*, 245-254.

Glover, D., Pallais, A., & Pariente, W. (2016). *Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores*. Unpublished manuscript.

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

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*, 493-503.

Gollwitzer, P. M., & Brandstaetter, V. (1997). Implementation intentions and effective goal pursuit. *Journal of Personality and Social Psychology, 73*, 186-199.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology, 38*, 69-119.

Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2009). Aging and prejudice: Diminished regulation of automatic race bias among older adults. *Journal of Experimental Social Psychology, 45*, 410-414.

Govorun, O., & Payne, B. K. (2006). Ego-depletion and prejudice: separating automatic and controlled components. *Social Cognition, 24*, 111-136.

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., 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., Poehlman, T. A., Uhlmann, E. L., & 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.

Guinote, A., Willis, G. B., & Martellotta, C. (2010). Social power increases implicit prejudice. *Journal of Experimental Social Psychology, 46*, 299-307.

Hawkins, C. B., & Nosek, B. A. (2012). Motivated independence? Implicit party identity predicts political judgments among self-proclaimed independents. *Personality and Social Psychology Bulletin, 38,* 1441-1455.

Iceland, J., Weinberg, D. H., & Steinmetz, E. (2002). *Racial and ethnic segregation in the United States, 1980 –2000*. Washington, DC: U.S. Government Printing Office, U.S. Census Bureau.

Holland, R. W., Aarts, H., & Langendam, D. (2006). Breaking and creating habits on the working floor: A field-experiment on the power of implementation intentions. *Journal of Experimental Social Psychology, 42,* 776-783.

James, W. (1904). Does `Consciousness' exist? *The Journal of Philosophy, Psychology and Scientific Methods, 1,* 477-491.

James, L., Vila, B., & Daratha, K. (2013). Results from experimental trials testing participant responses to White, Hispanic and Black suspects in high-fidelity deadly force judgment and decision-making simulations. *Journal of Experimental Criminology, 9,* 189-212.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology, 33,* 1-27.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology, 25,* 881-919.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41,* 137-146.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 774-788.

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology, 28,* 280-290.

Keith, B. E., Magleby, D. B., Nelson, C. J., Orr, E. A., Westlye, M. C., & Wolfinger, R. E. (1992). *The myth of the independent voter*. Berkeley, CA: Univ of California Press.

Krause, A., Rinne, U., & Zimmermann, K. F. (2012). Anonymous job applications of fresh Ph. D. economists. *Economics Letters, 117,* 441-444.

Kugelmass, H. (2016). "Sorry, I'm not accepting new patients": An audit study of access to mental health care. *Journal of Health and Social Behavior, 57,* 168-183.

Kulik, C. T., Pepper, M. B., Roberson, L., & Parker, S. K. (2007). The rich get richer: Predicting participation in voluntary diversity training. *Journal of Organizational Behavior, 28,* 753-769.Kuppens, T., Pollet, T. V., Teixeira, C. P., Demoulin, S., Craig Roberts, S. & Little, A. C. (2012). Emotions in context: Anger causes ethnic bias but not gender bias in men but not women. *European Journal of Social Psychology, 42,* 432-441.

Ladd, H. F. (1998). Evidence on discrimination in mortgage lending. *The Journal of Economic Perspectives, 12,* 41-62.

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass, 7,* 315-330.

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.

Lai, C. K.**,** Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143,* 1765-1785.

Lun, J., Sinclair, S., Whitchurch, E. R., & Glenn, C. (2007). (Why) do I think what you think? Epistemic social tuning and implicit prejudice. *Journal of Personality and Social Psychology, 93*, 957-972.

LaVeist, T. A. (2003). Racial segregation and longevity among African Americans: An individual-level analysis. *Health Services Research, 38*, 1719-1734.

Loury, G. C. (2009). *The anatomy of racial inequality*. Cambridge, MA: Harvard University Press.

Luszczynska, A., Sobczyk, A., & Abraham, C. (2007). Planning to lose weight: Randomized controlled trial of an implementation intention prompt to enhance weight reduction among overweight and obese women. *Health Psychology*, *26*, 507-512.

Ma, D. S., Correll, J., Wittenbrink, B., Bar-Anan, Y., Sriram, N., & Nosek, B. A. (2013). When fatigue turns deadly: The association between fatigue and racial bias in the decision to shoot. *Basic and Applied Social Psychology*, *35*, 515-524.

Madon, S., Guyll, M., Aboufadel, K., Montiel, E., Smith, A., Palumbo, P., & Jussim, L. (2001). Ethnic and national stereotypes: The Princeton trilogy revisited and revised. *Personality and Social Psychology Bulletin, 27*, 996-1010.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*, 93-120.

Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin, 36*, 512-523.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice reduction efforts. *Journal of Personality and Social Psychology, 65*, 469-485.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology, 83*, 1029-1050.

Monteith, M. J., & Mark, A. Y. (2009). Self-regulation and prejudice reduction. In T. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 507–520). New York: Psychology Press.

Monteith, M.J., Parker, L.R., & Burns, M.D. (2016). The self-regulation of prejudice. In T.D. Nelson (Ed.), *Handbook of Stereotyping, Prejudice, and Discrimination* (pp. 409-432). New York: Psychology Press.

Monteith, M.J., & Mark, A.Y. (2005). Changing one's prejudice ways: Awareness, affect, and self-regulation. *European Review of Social Psychology, 16*, 113-154.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*, 482– 488.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*, 231-259.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics*, *6*, 101-115.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*, 166-180.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007a). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes* (pp. 265-292). New York: Psychology Press.

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.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*, 413-417.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*, 421-433.

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.

Pavlov, I. P. (Ed.). (1927). *Conditional reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press, London.

Pelham, B. W., Carvallo, M., & Jones, J. T. (2005). Implicit egotism. *Current Directions in Psychological Science*, *14*, 106-110.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751-783.

Polman, E., Pollmann, M. M., & Poehlman, T. A. (2013). The name-letter-effect in groups: sharing initials with group members increases the quality of group work. *PloS ONE*, *8:* e79039.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369–381.

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*, 565–578.

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology, 44*, 386-396.

Richeson, J. A. & Ambady, N. (2003). Effects of situational power on automatic racial prejudice. *Journal of Experimental Social Psychology*, *39*, 177-183.

Rivera, L. A., & Tilcsik, A. (2016). Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market. *American Sociological Review*, *81*, 1097-1131.

Rooth, D. O. (2007). Implicit discrimination in hiring: Real world evidence. *IZA Discussion Paper*, 2764.

Rudman, L. A., Dohn, M. C., & Fairchild, K. (2007). Implicit self-esteem compensation: Automatic threat defense. *Journal of Personality and Social Psychology*, *93*, 798-813.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*, 995-1008.

Sanbonmatsu, D. M., & Fazio, R. H. (1990). The role of attitudes in memory-based decision making. *Journal of Personality and Social Psychology*, *59*, 614-622.

Sawyer, J., & Gampa, A. (2018). Implicit and explicit racial attitudes changed during Black Lives Matter. *Personality and Social Psychology Bulletin, 44*, 1039-1059.

Schuette, R. A., & Fazio, R. H. (1995). Attitude accessibility and motivation as determinants of biased processing: A test of the MODE model. *Personality and Social Psychology Bulletin, 21*, 704-710.

Schuman, H., Steeh, C., Bobo, L., & Krysan, M. (1997). *Racial attitudes in America*. Rev. ed. Cambridge, MA: Harvard University Press.

Sethi, R. (2019). Crime and punishment in divided societies. In D. Allen & R. Somanathan (Eds.), *Difference without Domination: Pursuing Justice in Diverse Democracies*. Chicago, IL: University of Chicago Press.

Sim, J. J., Correll, J., & Sadler, M. S. (2013). Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot. *Personality & Social Psychology Bulletin, 39*, 291-304.

Shook, N. J., & Fazio, R. H. (2008). Interracial roommate relationships an experimental field test of the contact hypothesis. *Psychological Science, 19*, 717-723.

Staats, A. W., & Staats, C. K. (1958). Attitudes established by classical conditioning. *The Journal of Abnormal and Social Psychology, 57*, 37-40.

Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In S. Oskamp (Ed.), *Reduce prejudice and discrimination* (pp. 23–46). Mahwah, NJ: Lawrence Erlbaum Associates.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as an efficient means of thought control. *Personality and Social Psychology Bulletin, 34*, 1332-1345.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin, 34*, 1332-1345.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven: Yale University Press.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458.

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

U. S. Bureau of Labor Statistics. (2016). *Labor Force Statistics from the Current Population Survey*. Retrieved from http://www.bls.gov/cps/cpsaat11.htm

U. S. Census. (2015). *Homeownership rates by race and ethnicity of householder*. Retrieved from http://www.census.gov/housing/hvs/data/ann15ind.html

U. S. Department of Education, National Center for Education Statistics. (2016). *The condition of education 2016*. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016144

Xu, K., Nosek, B. & Greenwald, A.G., (2014). Psychology data from the Race Implicit Association Test on the Project Implicit Demo website. *Journal of Open Psychology Data, 2*, p. e3.

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology, 107*, 468-485.

Westgate, E., Riskind, R. G., & Nosek, B. A. (2015). Implicit preferences for straight people over lesbian women and gay men weakened from 2006 to 2013. *Collabra, 1*, 1-10.