The Moral Psychology Handbook

John M. Doris and the Moral Psychology Research Group

Fiery Cushman, John M. Doris, Joshua D. Greene, Gilbert Harman, Daniel Kelly, Joshua Knobe, Edouard Machery, Ron Mallon, Kelby Mason, Victoria McGeer, Maria W. Merritt, Shaun Nichols, Joseph M. Paxton, Alexandra Plakias, Jesse J. Prinz, Erica Roedder, Adina L. Roskies, Timothy Schroeder, Walter Sinnott-Armstrong, Chandra Sekhar Sripada, Stephen Stich, Valerie Tiberius, Liane Young

OXFORD
UNIVERSITY PRESS

13

Race and Racial Cognition

DANIEL KELLY, EDOUARD MACHERY, AND RON MALLON[1]

A core question of contemporary social morality concerns how we ought to handle racial categorization. By this we mean, for instance, classifying or thinking of a person as *black*, *Korean*, *Latino*, *white*, etc. While it is widely agreed that racial categorization played a crucial role in past racial oppression, there remains disagreement among philosophers and social theorists about the ideal role for racial categorization in future endeavors. At one extreme of this disagreement are short-term eliminativists who want to do away with racial categorization relatively quickly (e.g. Appiah, 1995; D'Souza, 1996; Muir, 1993; Wasserstrom, 2001/1980; Webster, 1992; Zack, 1993, 2002), typically because they view it as mistaken and oppressive. At the opposite end of the spectrum, long-term conservationists hold that racial identities and communities are beneficial, and that racial categorization—suitably reformed—is essential to fostering them (e.g. Outlaw, 1990, 1995, 1996). While extreme forms of conservationism have fewer proponents in academia than the most radical eliminativist positions, many theorists advocate more moderate positions. In between the two poles, there are many who believe that racial categorization is valuable (and perhaps necessary) given the continued existence of racial inequality and the lingering effects of past racism (e.g. Haslanger, 2000; Mills, 1998; Root, 2000; Shelby, 2002, 2005; Sundstrom, 2002; Taylor, 2004; Young, 1989). Such authors agree on the short-term need for racial categorization in at least some domains, but they often differ with regard to its long-term value.

[1] We are grateful to the Moral Psychology Research Group for several useful discussions of this material, and are particularly thankful to John Doris, Tim Schroeder, and Erica Roedder for their many insightful comments on earlier drafts of this chapter. We would also like to thank Luc Faucher for his feedback on a previous version. Remaining mistakes are ours. Finally, we would like to thank Project Implicit (http://www.projectimplicit.net/) for permission to use their stimulus materials in this chapter.

3. Racial Evaluation and Implicit Social Cognition

Racial categorization looks to raise problems both for eliminativists and conservationists. One might be tempted, however, to think those results weigh especially heavily against eliminativism, and tilt the balance of considerations toward conservationism. In this section, we suggest that the conservationist goal of reducing negative racial evaluation has problems of its own—problems that the disregard of psychology has kept from being addressed.

In social psychology, recent advances in experimental measurement techniques have allowed psychologists to explore the contours of our capacities for racial evaluation with great precision, and a set of unsettling results has emerged. Most relevant of these is a particular phenomenon that has been confirmed repeatedly: people who genuinely profess themselves to be tolerant, unbiased, and free of racial prejudice nonetheless often display signs of implicit racial bias on indirect experimental measures. These methods were designed to bypass one's explicitly held views, i.e. those available via introspection and self-report, and instead systematically probe the less transparent workings of attitudes, associations, and processes linked to categorization and evaluation. After reviewing the relevant findings, we shall go on to assess their implications for the normative debate between eliminativism and conservationism.

3.1. *Indirect Measures and Implicit Cognition*

Consider how you could find out about someone else's mathematical prowess, or their ability to distinguish the subtleties of red wines. Perhaps the most obvious way would be to simply *ask* that person outright, "How good are you at math? Can you integrate a multi-variable equation?" or "How educated is your wine palate? Can you appreciate the difference between a California merlot and a Chilean cabernet sauvignon?" Alternatively, you might take a more circuitous route, and proceed by giving the person a set of math problems or a wine taste test, and infer their mathematical abilities or wine sophistication from their performance on the respective tests. The first type of strategy depends for its reliability on the sincerity of the person's self-report, the absence of self-deception in their self-assessment, and their ability to introspectively access the relevant information. The second type, though less direct in some ways, has the advantage of bypassing all three of these obstacles.

For similar reasons, indirect strategies have become trusted instruments for investigating many cognitive capacities, and research on implicit social

cognition is no exception. We shall call meausures that rely on such strategies *indirect measures*.[18] According to Nosek et al. (2007), most indirect measures are:

> [M]easurement methods that avoid requiring introspective access, decrease the mental control available to produce the response, reduce the role of conscious intention, and reduce the role of self-reflective, deliberative processes. (2007: 267)[19]

This description isn't definitive, but it gets across the flavor of indirect measures, the most prominent of which will be described in more detail below.

First, though, some terminological stipulations will lend clarity to the discussion. The term "implicit" is a source of potential confusion in this literature, as it is often applied to both the cognitive processes as well as the experimental measures used to probe them, and is treated as loosely synonymous with "automatic," "unconscious," and various other terms (Greenwald and Banaji, 1995; Greenwald et al., 1998; Cunningham et al., 2001; Eberhardt, 2005; Nosek et al., 2007). In what follows, we shall use "indirect" to describe measurement techniques, namely those that do not rely on introspection or self report, and reserve "implicit" only for mental entities being measured. Moreover, we will follow Banaji et al. (2001) and use 'implicit' to describe those processes or mechanisms operating outside the subject's conscious awareness, and "automatic" to denote those that operate without the subject's conscious control.

The Implicit Association Test (IAT) The IAT has been the most widely used indirect measure, and has been consequently subjected to the most scrutiny.[20] It was initially conceived of as "a method for indirectly measuring the strengths of associations," designed to help "reveal associative information that people were either unwilling or unable to report" (Nosek et al. 2007: 269). At its heart, the IAT is a sorting task. Most instances of the IAT involve four distinct categories, usually divided into two pairs of dichotomous categories. For instance, an IAT might involve the category pairs *black* and *white* (called "target concepts"), on the one hand, and *good* and *bad* (called "attribute dimensions") on the other. In one common case, the exemplars of th

[18] Phelps et al. (2000) and Phelps et al. (2003) use this term to distinguish indirect from "direct measures that use techniques like interviews or questionnaires that rely on verbal and written self-repo

[19] Thus characterized, indirect testing is not a particularly recent development to psychology (se e.g., Stroop, 1935).

[20] The first presentation of the test itelf, along with the initial results gathered using it, can be fou in Greenwald et al. (1998). Greenwald & Nosek (2001) and Nosek et al. (2007) both present mo recent reviews of research using IATs, as well as assessments of the methodological issues generated use of the test and interpretation of results. It should also be noted that there are several variants of t basic paradigm (e.g. Cunningham et al., 2001).

categories *black* and *white* are pictures of black and white faces, while exemplars of the other two categories are individual words, such as "wonderful," "glorious," and "joy," for *good*, "terrible," "horrible," and "nasty," for *bad*. During trials, exemplars are displayed one at a time, in random order, in the middle of a computer screen, and participants must sort them as fast as they can.

Crucial to the logic of the test is the fact that participants are required to sort the exemplars from the *four* categories using only *two* response options. For instance, they are told to press "e" when presented with any exemplar of *good* or any exemplar of *black*, and press "i" when presented with any exemplar of *bad* or any exemplar of *white*. Equally crucial to the logic of IATs is that they are *multi-stage* tests (often comprising five stages), and the response options (the "e" and "i" keys) are assigned to different categories in different stages. So one stage might require the participant to respond to exemplars of *good* or *black* with the "e" response option and exemplars of *bad* or *white* with the "i" response option, while the next stage assigns *bad* or *black* to the "e" response option and *good* or *white* to the "i" response option. Paired categories such as *good* and *bad*, or *black* and *white*, however, never get assigned to the same response options (each response option is assigned one "target concept" and one "attribute dimension"). When a participant makes a sorting error, it must be corrected as quickly as possible before he or she is allowed to move on to the next exemplar. Precise reaction times are measured by the computer on which the test is being taken, as is correction time and number of errors.[21]

Coarse-grained interpretation of performance is fairly straightforward. Generally speaking, the "logic of the IAT is that this sorting task should be easier when the two concepts that share a response are strongly associated than when they are weakly associated." More specifically, "ease of sorting can be indexed

[21] See the citations in previous footnote for a much more detailed and technically precise discussion of this technique. In order to get the feel of the test, however, one is much better off simply taking one; different versions of it are available at https://implicit.harvard.edu/implicit/demo/.

both by the speed of responding (faster indicating stronger associations) and the frequency of errors (fewer errors indicating stronger association)" (Nosek et al., 2007: 270). The idea can be illustrated with our example case. If a participant is able to sort exemplars faster and more accurately when *good* and *white* share a response option than when *good* and *black* share a response option, this fact is interpreted as an indirect measure of a stronger association between the two categories *good* and *white*, and hence an implicit preference for white, or, conversely, an implicit bias against black. This is called the IAT effect. The size of the relative preference or bias is indicated by the disparity between the speed and accuracy of responses to the same stimuli using different response option pairings. Finally, the associations thus revealed are taken to be indicative of processes that function implicitly and automatically, because the responses must be made quickly, and thus without benefit of introspection or the potentially moderating influence of deliberation and conscious intention. While the details of the method can seem Byzantine, the basic idea behind the test remains rather simple: stronger associations between items will allow them to be grouped together more quickly and accurately; the sophisticated set up and computerization just allow fine-grained measurement of that speed and accuracy.

Modern Racism Scale (MRS) By way of contrast with indirect measures like the IAT, the MRS is a direct measure of racial attitudes, one that is often used in conjunction with the indirect measures. This is a standard self-report questionnaire that was designed to probe for racial biases and prejudices (McConahay, 1986). It poses statements explicitly about racial issues (e.g. "Over the past few years, Blacks have gotten more economically than they deserve"; "It is easy to understand the anger of Black people in America"; "Blacks are getting too demanding in their push for equal rights"), and allows participants to react to each statement by selecting, at their leisure, one of the responses, which range from Strongly Disagree to Strongly Agree.

The use of direct measures *together* with indirect measures is important because it is the conjunction of the two that supports the inference to not just automatic but *implicit* processes and biases in the sense discussed earlier. Recall that implicit processes operate outside the introspective access and awareness of participants, while automatic processes are those that operate beyond conscious control. There is much overlap, but these two terms are not completely coextensive; disgust responses, for example, may be automatic, but they are rarely implicit. That participants can exhibit biases on indirect measures, despite the fact that they report having no such biases when asked directly, lends support to the conclusion that what manifests in the indirect

tests is indeed the result of processes that are unavailable to introspection and self-report.

3.2. *Evidence of Biases and their Effects*

3.2.1. *Implicit Racial Bias* These types of indirect measures have been used to probe and reveal a wide variety of implicit biases, including age biases (e.g. Levy & Banaji, 2002), gender biases (e.g. Lemm & Banaji, 1999), sexuality biases (e.g. Banse et al., 2001), weight biases (e.g. Schwartz et al., 2006), as well as religious and disability biases (see Lane et al., 2007 for a review). Some of the first and most consistently confirmed findings yielded by these tests, however, center on racial biases.[22] Participants who profess tolerant or anti-racist views on direct tests often reveal racial biases on indirect tests. This result is quite robust; similar dissociations have been found using a wide variety of other indirect measures, including evaluative priming (Cunningham et al., 2001; Devine et al., 2002), the startle eyeblink test (Phelps et al., 2000; Amodio et al., 2003), and EMG measures (Vanman et al., 1997). In other words, it is psychologically possible to be, and many Americans actually are, *explicitly racially unbiased while being implicitly racially biased.*[23] Moreover, not only is it possible for two sets of opposing racial evaluations to coexist within a single agent, but, as we shall see, when it comes to altering and controlling them, the different types of biases may be responsive to quite different methods.

3.2.2. *Implicit Racial Bias and Behavior* Perhaps a natural question to ask before going any farther is whether or not the biases revealed by indirect measurement techniques have any influence on judgments or ever lead to any actual prejudicial behavior, especially in real-world situations. Obviously, the question is important for a variety of reasons, not least of which is assessing

[22] The first paper to showcase the IAT included the results from three separate experiments, one of which was a test for implicit racial biases in white American undergraduates (Greenwald et al., 1998). Results exhibited a now-familiar, but still disturbing, pattern: while most (19 of 26) of the participants explicitly endorsed an egalitarian, or even pro-black, position on the direct measures (including the MRS), all but one exhibited an IAT effect indicating implicit white preference. This was the first study using the IAT to investigate this phenomenon, but previous work using less sophisticated methods had revealed similar results (e.g. Devine, 1989; Greenwald & Banaji, 1995; Fazio et al., 1995). Since the initial 1998 paper, similar results from IATs have been reported so often and found so reliably that they have become a commonplace (Kim & Greenwald, 1998; Banaji, 2001; Ottaway et al., 2001).

[23] While the fact that implicit and explicit racial biases can be dissociated is no longer a subject of much controversy, the relationship between the two is still very much in question. While early discussions stressed the complete independence of subjects' performances on direct and indirect tasks (Greenwald et al., 1998), follow-up work has shown that the two can be involved in complicated correlations (Greenwald et al., 2003; Nosek et al., 2007).

the feasibility of revisionist proposals offered by philosophers of race. Racial theorists (and others) skeptical of the relevance of this psychological literature might be inclined to simply dismiss it on the grounds that tests like the IAT measure mere linguistic associations or inert mental representations that people neither endorse nor act upon in real-world scenarios (see, e.g., Gehring et al., 2003). Others, who grant that the results of indirect tests (which usually turn on differences that are a matter of milliseconds) are of legitimate theoretic interest to psychologists,[24] might still remain skeptical that implicit biases, whatever they turn out to be, are powerful enough to make any practical difference in day-to-day human affairs.

We do not think that such skepticism is justified. First, we are impressed by mounting evidence that race and racial bias can still have measurable and important effects in real-world situations. In a field study by Bertrand and Mullainathan (2003), researchers responded to help-wanted ads in Boston and Chicago newspapers with a variety of fabricated résumés. Each résumé was constructed around either a very black-sounding name (e.g. "Lakisha Washington" or "Jamal Jones") or a very white-sounding name (e.g. "Emily Walsh" or "Greg Baker"). When the résumés were sent out to potential employers, those bearing white names received an astonishing 50% more callbacks for interviews. Moreover, those résumés with both white names and more qualified credential received 30% more callbacks, whereas those highly qualified black résumés received a much smaller increase. The numbers involved are impressive, and the amount of discrimination was fairly consistent across occupations and industries; in Bertrand and Mullainathan's own words:

> In total, we respond to over 1300 employment ads in the sales, administrative support, clerical and customer services job categories and send nearly 5000 resumes. The ads we respond to cover a large spectrum of job quality, from cashier work at retail establishments and clerical work in a mailroom to office and sales management positions. (3)

Interestingly, employers who explicitly listed "Equal Opportunity Employer" in their ad were found to discriminate as much as other employers.

Similar evidence of race and racial bias influencing real-world situations comes from a recent statistical analysis of officiating in NBA (National Basketball Association) games, which claims to find evidence of an "opposite race bias" (Price & Wolfers, ms). The study, which took into account data from the 12 seasons from 1991–2003, found evidence that white referees called slightly

[24] For instance, some psychologists see problems with the quick inference from IAT results to the attribution of implicit prejudice (Blanton & Jaccard, 2008; Arkes & Tetlock, 2004).

but significantly more fouls on black players than white players, as well as evidence of the converse: black referees called slightly but significantly more fouls on white players than on black players.

The racial composition of teams and refereeing crews was revealed to have slight but systematic influence on other statistics as well, including players' scoring, assists, steals, and turnovers. The study found that players experience a decrease in scoring, assists and steals, and an increase in turnovers when playing before officiating crews primarily composed of members of the opposite race. (For example, a black player's performance will fall off slightly when at least two of the three referees are white. For the purposes of the study all referees and players were classified as either black or not black.) These findings are especially surprising considering the fact that referees are subject to constant and intense scrutiny by the NBA itself, so much so that they have repeatedly been called "the most ranked, rated, reviewed, statistically analyzed and mentored group of employees of any company in any place in the world" by commissioner David Stern (Schwartz & Rashbaum, 2007).

While neither the IAT, nor any other indirect, controlled experimental technique was given to participants in either the NBA or the résumé studies, explanations that invoke implicit biases look increasingly plausible in both cases. Indeed, the sorts of real-world findings coming from these sorts of statistical analyses and field studies, on the one hand, and the types of automatic and implicit mental processes revealed by the likes of the IAT, on the other, appear to complement each other quite nicely. Explicit racism on the part of NBA referees or the employees responsible for surveying resumes and deciding whom to contact for job interviews may account for some fraction of the results, but given the conditions in which the respective groups perform their jobs, we are skeptical that appeal to explicit racism alone can explain all of the results. Especially in the heat of an NBA game, referees must make split-second judgments in high-pressure situations. These are exactly the type of situations where people's behaviors are likely to be influenced by automatic processes.

Moreover, researchers have begun to push beyond such plausible speculation and explicitly link indirect measures with behavior in controlled settings. These studies further confirm that when participants have to make instantaneous decisions and take quick action, racial biases affect what they do. Payne (2006) reviews a large body of evidence concerning participants who are asked to make snap discriminations between guns and a variety of harmless objects. Participants, both white and black, are more apt to misidentify a harmless object as a gun if they are first shown a picture of a black, rather than a picture of a white. This effect has become known as the "weapon bias." Similar results are found with participants who explicitly try to avoid racial biases.

Moreover, presence of a weapon bias correlates with performance on the racial IAT (Payne, 2005). This suggests that implicit racial biases may indeed lie behind the weapon bias. (For more discussion and a wider range of cases that link implicit biases of all sorts to behavior, see Greenwald et al., 2009.)

The real-world relevance of such findings is increasingly difficult to deny. It could help explain familiar anecdotes of sincerely egalitarian people who are surprised when they are called out for racist behavior or biased decision-making, especially when such accusations turn out to be legitimate. Another, more concrete example is provided by the highly publicized death of Amadou Diallo in 1999. He was shot and killed by New York police officers who thought he was drawing a gun, when in actuality he was just reaching for his wallet.

3.2.3. *Mitigating the Effects of Implicit Racial Bias* In addition to its direct real-world relevance, this body of psychological research has implications relevant to normative racial theorists. Before discussing those implications, however, we wish to call attention to a relevant offshoot of this literature that investigates whether and how implicit biases can be brought under control, and whether their expression in behavior and judgment can be mitigated.[25] Preliminary evidence suggests that implicit biases and the downstream effects they typically give rise to can indeed be manipulated. Research is beginning to shed some light on the effectiveness, and lack thereof, of different methods for bringing them under control. We consider three different methods of mitigating the effects of implicit biases: manipulating the immediate environment, self-control, and blocking the development or acquisition of implicit bias.

First, some of these studies suggest that while implicit biases operate beyond the direct conscious control of the participants themselves, they can be rather dramatically influenced by manipulating aspects of a person's immediate *environment*, often their social environment. Dasgupta and Greenwald (2001) showed participants pictures of admired and disliked black and white celebrities (Denzel Washington, Tom Hanks, Mike Tyson, Jeffrey Dahmer) and found that exposure to admired blacks and disliked whites weakened the pro-white IAT effect. They also found that the weakening of the implicit bias measured immediately after exposure to the pictures was still present 24 hours later, while the subjects' explicit attitudes remained unaffected. Lowery et al. (2001) found that the implicit biases of white Americans (as measured by the IAT) could be lessened merely by having the participants interact with a black

[25] See the special issue of *Journal of Personality and Social Psychology* (vol. 81, issue 5, 2001), for an introductory overview and collection of articles devoted to this topic.

experimenter rather than a white experimenter. Richeson and Ambady (2003) showed situational differences can affect implicit biases: when white female participants were told they were going to engage in a role-playing scenario, either as a superior or a subordinate, immediately after they completed an IAT, those anticipating playing a subordinate role to a black in a superior role showed fewer traces of implicit racial bias than those anticipating play a superior role to a black in a subordinate role.

Other studies investigated the extent to which a participant can obliquely influence their own implicit biases by some form of *self-control*, either by actively suppressing their expression or indirectly affecting the implicit processes themselves. For instance, Blair et al. (2001) found that participants who generate and focus on counter-stereotypic mental imagery of the relevant exemplars can weaken their IAT effects. Richeson et al. (2003) present further brain-imaging and behavioral data suggesting that while so-called "executive" functions (in the right dorsolateral prefrontal cortex) can serve to partially inhibit the expression of racial biases on indirect tests, the act of suppressing them requires effort and (or perhaps in the form of) attention.

A different way to eliminate the pernicious effects of implicit biases might be to nip the problem in the bud, so to speak, and to keep people (young children, for instance) from acquiring or developing them in the first place. Research raises difficulties for this possibility, however. Preliminary evidence suggests that implicit biases are easier to acquire than their explicit counterparts. The same evidence suggests implicit biases are harder to alter once acquired, and are difficult to eliminate. This is given a rather striking experimental demonstration by Gregg et al. (2006). Participants in this study were told about two imaginary groups of people, the second of which was cast in a negative light in order to induce biases against its members. After they had been given this initial information, however, participants were told that the damning description of the second group was incorrect, the mistaken result of a computer error. Gregg and his colleagues then gave participants both direct and indirect tests, and found that while their explicit biases had disappeared, their implicit biases, as measured by an IAT, remained. Work on acquisition and the development of the capacity for implicit social cognition in general is still in its infancy, but initial forays into the area suggest that the development of the capacity for implicit bias is rapid, independent of explicit teaching, and distinct from the development of explicit biases (see Dunham et al., 2008).

These findings make up the beginning of a promising research program centered not only on implicit racial cognition itself, but on how the unwanted influence of implicit biases on judgment and behavior can be mitigated or brought under control. On the currently available evidence, it is not yet clear

whether the most effective strategies act on the implicit biases themselves, or on ancillary processes that underlie their expression in behavior or judgments. The bulk of this work does suggest that, at the very least, the expression of implicit biases is not impossible to alter. Indeed, while they are inaccessible via direct introspection and appear not to require—indeed, can even *defy*—deliberation or conscious intention, these studies suggest that implicit biases can be affected by changes in the social environment and less direct forms of self-control. While blocking their development or acquisition may be an uphill battle, their expression can be restrained via strategic alterations of the social environment and specific forms of self-control.

3.3. *Consequences for the Debate between Eliminativism and Conservationism*

While it is fascinating in its own right, this body of work in social psychology is clearly relevant to a variety of philosophical issues concerning race.[26] To be forthright, the psychological story is still far from complete, and in a number of ways:

(a) the extent to which many of the results reported can be generalized from one culture to the next remains uncertain, as does the manner in which those results might be generalized;
(b) whether and which results can be generalized to racial groups beyond blacks and whites within a single culture (to include other putative racial groups such as Hispanics, Indians, Asians, etc.) is also uncertain (but see Devos et al., 2007);
(c) there is little systematic data concerning the ontogenesis of implicit racial biases (but see Baron & Banaji, 2006, Dunham et al., 2008);
(d) a more detailed account of the cognitive architecture underlying these implicit biases is needed, preferably one that can shed light on the admittedly live issue of how and how often the evaluations measured by the indirect tests are also involved in causal processes that lead to actual judgment and action;
(e) it is currently far from clear whether implicit biases of different types, for instance implicit racial biases, gender biases, age biases, disability biases, etc., all reflect the workings of the same set of cognitive mechanisms;
(f) more fine-grained and theoretically motivated distinctions are needed, since the term "group" used to interpret much of the data is probably too ambiguous to be of much serious use—as alluded to in Section 2,

[26] For an initial attempt to wrestle with the ethical implications of implicit racial biases, see Kelly & Roedder (2008), Faucher & Machery (forthcoming).

but significantly more fouls on black players than white players, as well as evidence of the converse: black referees called slightly but significantly more fouls on white players than on black players.

The racial composition of teams and refereeing crews was revealed to have slight but systematic influence on other statistics as well, including players' scoring, assists, steals, and turnovers. The study found that players experience a decrease in scoring, assists and steals, and an increase in turnovers when playing before officiating crews primarily composed of members of the opposite race. (For example, a black player's performance will fall off slightly when at least two of the three referees are white. For the purposes of the study all referees and players were classified as either black or not black.) These findings are especially surprising considering the fact that referees are subject to constant and intense scrutiny by the NBA itself, so much so that they have repeatedly been called "the most ranked, rated, reviewed, statistically analyzed and mentored group of employees of any company in any place in the world" by commissioner David Stern (Schwartz & Rashbaum, 2007).

While neither the IAT, nor any other indirect, controlled experimental technique was given to participants in either the NBA or the résumé studies, explanations that invoke implicit biases look increasingly plausible in both cases. Indeed, the sorts of real-world findings coming from these sorts of statistical analyses and field studies, on the one hand, and the types of automatic and implicit mental processes revealed by the likes of the IAT, on the other, appear to complement each other quite nicely. Explicit racism on the part of NBA referees or the employees responsible for surveying resumes and deciding whom to contact for job interviews may account for some fraction of the results, but given the conditions in which the respective groups perform their jobs, we are skeptical that appeal to explicit racism alone can explain all of the results. Especially in the heat of an NBA game, referees must make split-second judgments in high-pressure situations. These are exactly the type of situations where people's behaviors are likely to be influenced by automatic processes.

Moreover, researchers have begun to push beyond such plausible speculation and explicitly link indirect measures with behavior in controlled settings. These studies further confirm that when participants have to make instantaneous decisions and take quick action, racial biases affect what they do. Payne (2006) reviews a large body of evidence concerning participants who are asked to make snap discriminations between guns and a variety of harmless objects. Participants, both white and black, are more apt to misidentify a harmless object as a gun if they are first shown a picture of a black, rather than a picture of a white. This effect has become known as the "weapon bias." Similar results are found with participants who explicitly try to avoid racial biases.

Moreover, presence of a weapon bias correlates with performance on the racial IAT (Payne, 2005). This suggests that implicit racial biases may indeed lie behind the weapon bias. (For more discussion and a wider range of cases that link implicit biases of all sorts to behavior, see Greenwald et al., 2009.)

The real-world relevance of such findings is increasingly difficult to deny. It could help explain familiar anecdotes of sincerely egalitarian people who are surprised when they are called out for racist behavior or biased decision-making, especially when such accusations turn out to be legitimate. Another, more concrete example is provided by the highly publicized death of Amadou Diallo in 1999. He was shot and killed by New York police officers who thought he was drawing a gun, when in actuality he was just reaching for his wallet.

3.2.3. *Mitigating the Effects of Implicit Racial Bias* In addition to its direct real-world relevance, this body of psychological research has implications relevant to normative racial theorists. Before discussing those implications, however, we wish to call attention to a relevant offshoot of this literature that investigates whether and how implicit biases can be brought under control, and whether their expression in behavior and judgment can be mitigated.[25] Preliminary evidence suggests that implicit biases and the downstream effects they typically give rise to can indeed be manipulated. Research is beginning to shed some light on the effectiveness, and lack thereof, of different methods for bringing them under control. We consider three different methods of mitigating the effects of implicit biases: manipulating the immediate environment, self-control, and blocking the development or acquisition of implicit bias.

First, some of these studies suggest that while implicit biases operate beyond the direct conscious control of the participants themselves, they can be rather dramatically influenced by manipulating aspects of a person's immediate *environment*, often their social environment. Dasgupta and Greenwald (2001) showed participants pictures of admired and disliked black and white celebrities (Denzel Washington, Tom Hanks, Mike Tyson, Jeffrey Dahmer) and found that exposure to admired blacks and disliked whites weakened the pro-white IAT effect. They also found that the weakening of the implicit bias measured immediately after exposure to the pictures was still present 24 hours later, while the subjects' explicit attitudes remained unaffected. Lowery et al. (2001) found that the implicit biases of white Americans (as measured by the IAT) could be lessened merely by having the participants interact with a black

[25] See the special issue of *Journal of Personality and Social Psychology* (vol. 81, issue 5, 2001), for an introductory overview and collection of articles devoted to this topic.