from the book...

DISPUTED MORAL ISSUES

A Reader Fifth Edition

Mark Timmons
University of Arizona

New York Oxford
OXFORD UNIVERSITY PRESS

DANIEL KELLY AND ERICA ROEDDER

Racial Cognition and the Ethics of Implicit Bias

Implicit bias is a phenomenon that has attracted much attention from scientists of various fields (including cognitive science) as well as philosophers, particularly as it bears on racial bias. In their article, Daniel Kelly and Erica Roedder summarize the empirical research on implicit racial bias as a basis for raising a number of normative ethical questions about behaviors and judgments "linked" to such biases, including: (1) Is such bias normatively problematic? (2) If so, how? (3) Should each person believe they are implicitly racially biased? (4) If so, does one have reason to compensate in some way to offset this bias?

1. INTRODUCTION

. . . The aim of this paper is two-fold. Our first goal is to call philosophical attention to some of the most provocative empirical work on racial cognition. Accordingly, the first half of the paper will discuss one portion of this large literature: work regarding implicit racial biases. Our second goal is to raise a number of philosophical questions about the proper normative assessment of behaviors and judgments linked to those implicit biases. In the second half of

From *Philosophy Compass* 3 (2008): 522–540.

the paper, then, we will assume these implicit racial biases are roughly as current research depicts them to be, and go on to sketch a few of the most promising avenues of philosophical research that we believe are opened up by the psychological complexities revealed in this work on racial cognition.

2. IMPLICIT RACIAL BIAS

Rather ingenious strategies have uncovered subtle forms of racial discrimination that still exist in real world settings. One recent study investigated the effect of race on hiring practices in two U.S. cities. Researchers sent out fabricated resumes to Help Wanted ads appearing in major newspapers in Boston and Chicago. Half of the resumes were headed by a very Black sounding name (e.g., Lakisha and Jamal), while the other half were headed by a very White sounding name (e.g., Emily and Greg).[1] The results were remarkable: overall, resumes bearing White names received an astonishing 50 percent more callbacks for interviews than their Black counterparts. Furthermore, an interesting pattern emerged for highly qualified resumes. For White sounding names, resumes with highly qualified credentials received 30 percent more callbacks than their less qualified White counterparts; in contrast, employers did not differentiate nearly as much between highly qualified Black resumes and their less qualified Black counterparts. The amount of discrimination was fairly consistent across occupations and industries. Of particular interest was the fact that employers who explicitly listed 'Equal Opportunity Employer' in their ad discriminated just as much as other employers (Bertrand and Mullainathan).

Another recent study found evidence of subtle forms of bias in the officiating of NBA basketball games. Despite the fact that referees are subject to constant and intense scrutiny by the NBA itself (Commissioner David Stern has called them 'the most ranked, rated, reviewed, statistically analyzed and mentored group of employees of any company in any place in the world'; Price and Wolfers, ms, p. 3), statistical analysis of data taken over a 12-year period found evidence of a slight 'opposite race bias'. This mainly manifested in the fact that White referees called slightly (but statistically significantly) more fouls on Black players than they called on White players, while Black referees called slightly (but again statistically significantly) more fouls on White than Black players. The racial composition of teams and refereeing crews was also found to have similar subtle effects on other statistics as well, including players' scoring, blocks, steals, and turnovers (Price and Wolfers). . . .

The resume and NBA studies are obviously suggestive, but other methods are needed to more directly address questions about the cognitive mechanisms that produce the patterns of behavior documented by those real world studies. And indeed, such methods exist. One of the most sophisticated and widely used windows into racial cognition is an experimental measurement technique called the Implicit Association Test, or IAT for short. More than any other technique, the IAT has been used to establish the existence and shed light on the character of implicit racial biases. In short, the IAT has been used to show that a great many people, including those who genuinely profess themselves to be racially impartial and explicitly disavow any form of racial prejudice, display subtle signs of racial bias in controlled experimental settings. Understanding how the IAT works will help make this clearer.

The Implicit Association Test (IAT)

The IAT was designed by psychologists to probe aspects of thought that are not easily accessible or immediately available to introspection.[2] Rather than provide a technical description of how the test works, it will be more useful to convey its flavor. Suppose you have to sort words from the following list as quickly as possible, putting every good adjective and Black name in column A, and every bad adjective and White name in column B.

Lakisha
Delicious
Sad
Jamal
Greg

Death
Happy
Unhappy

Suppose now that you are asked to do another iteration with a similar list of words, but with a crucial difference. This time, you must place the good adjectives and White names in column A and bad adjectives and Black names in column B. Again, you should go as fast as you can without making any mistakes.

Most likely, you found it easier to sort the words when the good adjectives were paired with the White names (delicious, Greg) and the bad adjectives were paired with Black names (sad, Lakisha). This simple exercise is similar to an IAT in a number of relevant ways. First of all, it involves items (in this case words) that obviously fall into one of four categories: White, Black, good, and bad. Second, it asks you to sort those items into one of two groups, column A or column B, which are specified disjunctively: for instance, in the first iteration, column A gets the items that are White or bad, column B gets the items that are Black or good. Third, the groupings are switched in various iterations: Black and good are grouped together in the first iteration, while Black and bad are grouped together in the second. Finally, speed and accuracy are of the essence in both.[3]

IATs are performed on a computer, and so differences in accuracy, as well as minute differences in speed of sorting, can easily be recorded and compared across iterations. The core idea behind both our toy sorting exercise and actual IATs is that stronger associations between items will allow them to be grouped together more quickly and easily. For instance, faster and more accurate performance on iterations when good and White items are to be grouped together than on iterations when good and Black are to be grouped together indicates a stronger association between good and White. Stronger associations between good and White, in turn, are taken to indicate an implicit bias towards Whites over Blacks. As should be evident, this test does not use self-report or explicitly ask subjects about their attitudes about race. Unlike those more direct tests that are based on self-report, and which are often used in conjunction with IATs (e.g. McConahay), the IAT requires subjects to make snap judgments that must be made quickly, and thus without moderating influence of introspection and deliberation and often without conscious intention. Biases revealed by an IAT are often thought to implicate relatively automatic processes.

IAT AND RACE

Indirect measurement techniques of this sort have been used to explore a wide variety of implicit biases, including those linked to age, gender, sexuality, disability, weight, and religion. Some of the first and most consistently confirmed findings, however, have centered on racial biases.[4] In using tools like the IAT in conjunction with more direct, self-report methods, researchers have further found that even those who sincerely profess tolerant or anti-racist views can nevertheless harbor implicit racial biases (often to their own surprise and chagrin).[5] Counterintuitive as it may seem, this robust pattern of results shows that a person's avowed views on race and racism are not a reliable guide to whether or not they are implicitly biased.

The dissociation between implicit and explicit racial attitudes is difficult to deny at this point, but some have remained skeptical of the significance of IAT results, suggesting that implicit biases have no influence on actual behavior. Rather, they hold out the possibility that tests like the IAT are simply measuring associations between otherwise inert mental representations (e.g. Gehring, Karpinski, and Hilton). While we respect a healthy sense of skepticism, we believe it is unjustified in this case. A recent meta-analysis of 103 IAT studies confirmed that performance on the IAT is predictive of many types of behavior and judgment. For instance, one study showed subjects harboring implicit biases against Blacks were more likely to interpret ambiguous actions made by a Black person negatively rather than neutrally (Rudman and Lee), while another documented subtle influences on the way subjects interacted with Black experimenters: when talking to a Black experimenter, subjects with implicit bias towards Blacks smiled less, talked less, and made more speaking errors versus when they interacted with a White experimenter (McConnell and Leibold). Recent work has even shown that implicit biases can influence which prescriptions doctors are likely to issue to Black versus White patients (Green et al. as cited in Lane et al.). Moreover, in research on intergroup discrimination (including racial discrimination), the IAT was found to be more predictive than self-report. Finally, the existence of the types of real world patterns discovered in the resume and NBA studies cries out for just the sort of explanation that implicit racial biases can provide. Recall that in both of those studies, evidence of racial bias was found despite the fact that those involved had obvious incentives and explicitly stated intentions to treat members of different races impartially and fairly.

We will conclude with a final example that speaks to both the influence of IAT results on behavior and real world relevance. Like those made by NBA referees, many important judgments must be made almost instantaneously and in high pressure situations. Such split second decisions have been shown to be sensitive to race in other ways as well. A number of studies have asked people to make snap decisions about whether a presented object is a gun or some other harmless object. Researchers found that when first shown a picture of a Black face, both White and Black Americans become more likely to misidentify a harmless object as a gun (Payne, 'Weapon Bias'). Not only is this 'weapon bias' found in people who explicitly try to avoid racial biases, but the weapon bias is highly correlated with the indirect measures of racial biases, including the IAT (Payne, 'Conceptualizing Control'). The relevance of such findings is difficult to deny, especially in light of tragedies such as the 1999 shooting of Amadou Diallo, who was shot 41 times by police officers who thought he was drawing a gun; he was actually just reaching for his wallet.

3. NORMATIVE QUESTIONS

So far, we have discussed the psychology of racial cognition, focusing on the implicit attitude test. Such findings introduce new and significant normative questions. In the rest of this article, we'll briefly survey some of the normative questions that we think are fruitful areas for future research on racial cognition, and consider attempts to answer questions similar to them. We'll focus on two questions. Stated as simply as possible, those questions are:

1. Is it morally problematic to harbor implicit racial biases, i.e., those measured by the IAT?
2. Given that implicit racial bias is, by definition, implicit, might I be racially biased and not know it? For instance, should I think that I am biased in my grading of Black student essays, and should that affect my grading of those essays?

Is It Morally Problematic to Harbor Implicit Racial Bias?

One major question is whether it is morally problematic, in and of itself, to have an implicit bias against members of a particular race. Obviously, implicit racial bias is problematic insofar as it leads to harmful or unfair consequences. For instance, suppose implicit bias forms part of the explanation of why an innocent Black man is shot by a police officer. In this case, implicit bias is clearly a bad thing: it partly caused a *harmful* consequence, i.e., the death of a young man. Similarly, implicit racial bias is clearly bad insofar as it leads to *unfair* consequences, e.g., the unequal promotion of White versus Black employees within a company.

Let us set aside such consequences for a moment and consider the question of the implicit attitude itself—is this attitude intrinsically a bad thing? Now, one might think that attitudes are not the sort of thing that are apt for normative evaluation. A consequentialist, for instance, might think that attitudes are bad only insofar as they lead to unfortunate consequences. But we think there is good reason to reject such a view.

To see this, consider an *explicitly* racist person. We might ask of him, is his explicitly racist attitude, in and of itself, a bad thing? Suppose, for instance, that a man were never to act on his explicit racial beliefs, keeping his racist thoughts and feelings to himself. Perhaps he secretly seethes with disgust after drinking from water fountains used by Blacks and often has thoughts like, 'It's so obvious that Black children aren't as smart as White children'.

Most Westerners, we suspect, would disapprove of such a person. Even if the man never acts on these racist thoughts and feelings, and even if he is morally upright in all the other aspects of his life (e.g., he goes to church, is faithful to his wife, etc.), there is still something morally problematic about his attitudes. While it's good that the man refrains from acting on these racist thoughts and feelings, it is unfortunate and morally condemnable that he has such attitudes at all.

Further support for the idea that racial attitudes can be reprehensible even when they don't manifest behaviorally can be garnered by considering non-racial attitudes. Intuitively, you can be ashamed of having ever *believed* your loving spouse was cheating on you, or ashamed of the competitive *emotions* you felt when playing basketball with your 6-year-old son, regardless of whether these mental states lead to more obviously problematic behavior.[6]

Finally, we should note that a number of philosophers have explicitly suggested that racist mental states, in and of themselves, can be morally problematic. For instance, Garcia writes, 'bad effects that actually occur are *not* necessary for some people and their and [*sic*] mental phenomena to be racist' (53, italics ours), where racism is understood to be always *prima facie* wrong; he then goes on to argue that accounts of racism that only apply to racist *behavior* are misguided. As another example, Blum writes that that 'false [stereotypical] beliefs can be bad even if they do not contribute harm to their target' (262).

These considerations are meant to show that *explicit* thoughts and feelings, apart from the behavioral consequences they might bring about, can be subject to moral evaluation. If this is right, can the same be said about *implicit* thoughts and feelings? For instance, what exactly *are* implicit attitudes? Are they akin to Freudian unconscious states, occupying some deep core of our psyche? Or are they more minimal and peripheral? After all, the implicit attitude test was originally developed to test the *association* between two ideas. Let us consider, for a moment, an extremely minimal construal of implicit attitudes suggested by this: an implicit attitude is simply a tendency to associate one concept with another, in the way that, for instance, the concept *salt* might prime the concept *pepper*. A high IAT score, on this understanding, means that a person strongly associates, e.g., Black faces with handguns. Assuming that this is an exhaustive description of the implicit attitude—a tendency to associate one concept with another—can a tendency to associate certain concepts, in and of itself, be morally problematic?

One way to approach this question is through the lens of rationality. While it is clear that explicitly racist beliefs are mostly irrational, in addition to being immoral (e.g., the thought that Black children are less smart than White children), there seems to be room to argue that some implicit racial associations are (to a *limited extent*) rational. Insofar as this is the case, does that make the attitudes less morally problematic?

To see why someone might argue that implicit attitudes are sometimes rational, let's first consider a different case, i.e., gender. IAT results suggest that most people strongly associate men with science, more so than they do women with science (see Nosek and Banaji). But if the implicit attitude really is *just* an association of concepts, might it be rational to make such associations? Women, as a matter of fact, are not as well-represented in the sciences. Indeed, the fact of unequal distribution is an empirical premise in arguments for affirmative action and other attempts to raise the number of women in science. With respect to the issue of rationality, our point is that if implicit attitudes are construed in this very minimal way—as indicating only that a person associates two concepts—it appears they can be rational in some sense (e.g., insofar the association between concepts accurately reflects a correlation or statistical regularity that holds among those referents of the concepts).

Let us now return to the racial example. Consider the tendency to associate the faces of young, Black men with handguns. Someone might analogously suggest that, were it true that young Black men carry guns at a higher rate than White men, then it would be rational to associate Black faces with handguns. This is important because, as we mentioned earlier, it might be thought that rationality and morality go hand-in-hand: insofar as one's attitude is rational, it can't be immoral. . . .

We suspect this is not the right way to think about rationality and implicit attitudes. First, we think that a rational attitude may still be an *immoral* one. Rationality and morality are different virtues, so it should be expected that a person can have the one without the other. For instance, let us suppose certain evidence (such as test results) suggest that Elisa, a 3rd grader, is not very smart, and let us assume this evidence is strong enough to justify a teacher's belief that *Elisa is dumb*. If this evidence is enough to justify a teacher's belief, it will be (in some cases) enough to justify her parents' belief that *our daughter is dumb*. Nevertheless, it would be unfortunate, and arguably immoral, for Elisa's parents to be persuaded by the same degree of evidence that persuades her teachers. Elisa's parents have a special relationship with their daughter, one that arguably places moral constraints on them. In particular, that relationship places moral constraints on what they ought to believe of their daughter; namely, they ought to be inclined to believe the best of her. Of course, this is not to say they should turn a blind eye; if the evidence is very persuasive, they ought to believe it all things considered. The idea is rather a parent should give his or her child the benefit of the doubt. Roughly, when multiple conclusions about his or her children are reasonable, a parent has a moral obligation to believe the conclusion that is most kind. Our point is that it can sometimes be unkind or uncompassionate to believe ill of a person, even if it is rational to do so. Thus it can sometimes be immoral to hold a belief that is, in fact, rational. . . .

Instead of focusing on matters of rationality, we think that philosophers would do well to take a different angle in determining whether and why implicit racial biases are immoral. We think the meat of the issue is really two-fold. First, what exactly is the nature of these implicit attitudes? Implicit racial attitudes raise a number of novel moral issues; getting a grip on them will require a better understanding of the character of the implicit attitudes themselves. As we pointed out earlier, they might be construed as Freudian unconscious states or as very minimal mental associations, and these options are far from exhaustive. Resolving this question will take both experimental and conceptual work.

The second question is: why is it that *explicitly* racist attitudes are problematic, and can the same story be told about implicit attitudes? That is, can current accounts of what makes racist attitudes wrong, accounts that usually focus on explicit and conscious attitudes, be extended to cover implicit attitudes as well? In the remainder of this section, we'll examine this second question by focusing on the work of two authors: Garcia's account of racism and Blum's account of stereotyping.

Garcia's analysis of racism stresses the intrinsic features of certain attitudes. He writes that someone is a racist when they have certain affective and volitional attitudes:

> what makes someone a racist is her disregard for, or even hostility to, those assigned to the targeted race . . . she is hostile to or cares nothing (or too little) about some people because of their racial classification . . . hate and callous indifference (like love) are principally matters of *will* and desire: what does one want, what would one choose, for those assigned to this or that race? (43)

Importantly, Garcia construes racism as a deformation of affect and the will, and this informs his account of why it is morally problematic: racist attitudes, in themselves, are 'inherently contrary to the moral virtues of benevolence and justice' (43). Such attitudes, he argues, are hateful and ill-willed, and are thus opposed to benevolence by their very nature. On Garcia's account, the question of whether it is wrong to harbor an *implicit* attitude will therefore boil down to whether the attitude is intrinsically opposed to benevolence, e.g., whether it is an attitude of hate or one of ill-will.

Determining whether implicit attitudes are intrinsically opposed to benevolence, however, will require progress on two fronts. First, there are issues tied to empirical work and how to interpret evidence provided by indirect tests. Implicit attitudes (or some implicit attitudes) may turn out to be *merely cognitive* associations, in which case they would be neither affective nor volitional. Such attitudes, on Garcia's account, would not be intrinsically opposed to benevolence, and so would not be morally problematic.[7] . . .

Let us turn now to Blum's account of racial stereotyping. Because he is mainly concerned with stereotyping, Blum focuses on cognitive rather than affective mental states. He is careful to distinguish stereotyping from prejudice: the former is a cognitive distortion (e.g., stereotyping all Asians as good at math), whereas the later may be affect-laden to various degrees.

In attempting to extend Blum's view into the realm of implicit bias, we encounter some of the same problems that beset a straightforward extension of Garcia's. For instance, Blum emphasizes that stereotypical content can be disrespectful: 'Respect for other persons, an appreciation of others' humanity and their full individuality is inconsistent with certain kinds of beliefs about them' (262). To apply this line of thought to implicit attitudes, one would need to determine whether, for instance, harboring a weapon bias is disrespectful or constitutes a failure to appreciate another's full humanity. As above, it remains less than clear whether or not this is the case.

There is another thread in Blum's account, however, that is more easily generalized to implicit attitudes. In much of his article, Blum analyzes what stereotypes *do*. Two of the most important features he describes are that they mask individuality (the stereotyper fails to be sensitive to an individual's quirks and characteristics) and that they lead to what he calls *moral distancing*. In moral distancing, the stereotyper sees a stereotypee as more 'other' than he or she really is, and this corrodes her sense of a common, shared humanity. Here, we think Blum's account can be usefully and straightforwardly generalized to implicit attitudes. One must simply ask: do implicit attitudes have these deleterious effects? Do implicit biases mask individuality and lead to moral distancing? These sound like clear-cut empirical questions. If implicit racial biases do lead thinkers to fail to appreciate the individuality of others or to morally distance themselves, then it follows from Blum's account that those implicit biases are morally reprehensible.

In the last few paragraphs, we've considered the prospects for extending two different accounts of racial bias so as to cover implicit racial attitudes. We hope to have shown that this project, while viable, also poses substantive philosophical and empirical issues.

As a final note, it seems to us that ethicists working on implicit racism might be well-served by making a distinction between what is wrong and what is morally blameworthy. Particularly in the case of implicit attitudes, it is salient that their acquisition may be rapid, automatic, and uncontrollable.[8] These features, it might be thought, are related to features that establish blameworthiness—such as identification (Frankfurt) or reasons-responsiveness (Fischer and Ravizza). For instance, it might be said that the implicitly racist person doesn't identify with his implicit attitude, or that the attitude isn't responsive to reasons; thus we cannot hold a person fully accountable for those implicit attitudes. If this is right, one might say that such attitudes are morally wrong—and condemnable—but that the person himself cannot be blamed for having them. We are reluctant to embrace this solution wholeheartedly—it may turn out, for instance, that narrow-mindedness partially explains the acquisition of implicit racism—but such a solution illustrates how the distinction between moral wrong and moral blame might be of use in thinking about implicit racism.

Might I Be Racially Biased?

One of the remarkable features of *implicit* bias is the possibility that individuals may not be aware of their own bias. Neither introspection nor honest self-report are reliable guides to the presence of such mental states, and one may harbor implicit biases that are diametrically opposed to one's explicitly stated and consciously avowed attitudes. Because of this, thinkers face a thorny, real-life epistemological problem: given that a large proportion of the population is implicitly racial biased, is it reasonable to conclude that I, myself, am racially biased? And if I believe I might be, how should that belief affect my deliberation and behavior?

The possibility that you, yourself, may harbor implicit biases has implications for your concrete beliefs about everyday matters. For instance, suppose you are a White professor grading a Black student's paper, and you are initially inclined to give the paper an 89/100. Does the possibility of implicit racial bias give you good reason to think the paper actually deserves slightly better, e.g., 90 or 91 points? Let's call this example *the savvy grader*, since the problem arises when a thinker is psychologically savvy and is thus aware of the prevalence of implicit racial bias (the example is discussed in Roedder).

An analogy will be helpful here. Suppose you learn of psychological research showing that most people are inclined to underestimate the size of circles when set across a hatched background. Suppose you are later asked to judge the size of a circle on a hatched background. In deciding the size of the circle, it is most rational to estimate it to be slightly larger than you are initially inclined to guess. In doing this, one's goal is simply to come up with the most accurate estimate possible, and it seems fairly obvious that doing so requires correcting for the known visual bias.

With this in mind, let us return to the case of the savvy grader. Assume for a moment that experiments uncovered a racial bias in the grading of student papers (if this is too hard to imagine, one might think of some other decision-making domain, such as the hiring of employees, where there is more psychological evidence). We maintain that by parity of reasoning, it would be wise to make a similar adjustment for the implicit bias in grading, just as you would correct for the visual bias in judging the size of a circle. In both cases, one is acting for purely *epistemic* reasons; in order to give the most accurate grade, i.e., in order to grade the paper based on its merits, it is reasonable for the savvy grader to correct for the effects of racial biases.

It is worth pointing out that the reasoning behind the savvy grader case is very different than that usually offered in justification of affirmative action, which is mainly driven and justified by *moral* considerations. In affirmative action, benefits are given to members of an under-represented minority, beyond what is warranted strictly by the merits of those individuals, in the interest of some moral or political aim such as promoting diversity. Indeed, it is the fact that it calls for benefits over and beyond what an individual strictly merits that is at the root of much of the resistance to affirmative action. In contrast, the savvy grader acts on purely epistemic reasons, and her aim in making an adjustment to the initial grade is to give the Black student *exactly the grade the essay deserves*. The situation of the savvy grader can be thought of as a rational impairment: if you harbor a racial bias, then you are not responding to reasons in the way that you ought to, and the most epistemically responsible thing to do is to make some sort of correction.

There is much more to say here. In particular, we might wonder how much evidence of implicit racial bias a savvy grader needs before it is reasonable for her to adjust how she assigns grades. Roedder argues that the epistemic requirements are strikingly low: it is enough if she knows that, ceteris paribus, the bias exists *on average*. Consider the visual analogy again. If one were told that, *on average*, people see the circle as 25% smaller than it really is, most of us would take that as a reason to increase our original estimate of its size by 25%. Here, too, the epistemic factors relevant to grading papers do not appear substantially different from those of the visual case. If one knows that a slight bias exists *on average*, the reasonable response in both cases is to make the appropriate adjustments, hence for the savvy grader to slightly increase the grades she assigns to her Black student's papers. (Indeed, one should be concerned that, insofar as one is reluctant to compensate for racial bias in grading, this reluctance might stem from a self-deceptive tendency to believe oneself to be better than average; see Mele for a lucid and eye opening discussion of the prevalence of self-deception; Kruger and Dunning for empirical work on the problems we face assessing our own abilities and competences).

Of course, we don't yet have evidence that directly bears on the question of whether or not normal thinkers are implicitly biased against their Black students when grading papers; to date there has not been a systematic effort to look for racial bias in essay grading at the college level. Studies have instead focused on racial bias in hiring, housing, and other domains. But here an interesting wrinkle arises. It might be argued that one does not need to have direct evidence of implicit biases influencing judgment in a specific domain in order to be rationally compelled to make epistemic adjustments for them in that very domain. Rather, it is enough if one believes that, *were* these studies run, they *would* show such a bias. That is, when certain conditions are met, it is ceteris paribus rational to compensate for bias (and irrational not to) even in the absence of evidence of their influence. Moreover, the relevant conditions are fairly lax: you should make corrective adjustments if, based on the evidence of implicit racial biases in other domains, you have a hunch that it is more likely than not that such implicit biases also influence the grading of papers. After all, if you believe it is more likely than not that grading is somewhat racially biased, how could you justify continuing to give uncorrected grades?

Thus the important question is this: knowing what you know now about implicit bias in other domains (perhaps from reading this very article!), and if you had to place a bet, would you bet that there *is* a racial bias in grading or that there *isn't*? If you find yourself inclined to think that (more likely than not) there is a racial bias in grading, and if the line of reasoning sketched here is correct, then merely having this empirical hunch is enough to rationally compel you to make some sort of compensatory adjustment in your Black students' grades. We, the authors, do not yet know what to make of such an argument—but it strikes us as a surprising and unexpectedly good one.

It bears mentioning that we use the case of grading because it hits so close to home. But the considerations raised here can be generalized along a number of dimensions, for instance to other contexts (such as resumes, interviews, police behavior, etc.) as well as to other sorts of implicit biases (such as gender bias, height bias, etc.). Indeed, we believe there is an even broader lesson that can be taken away from the discussion. Implicit racial bias is just one example where psychological science shows our *reasoning* capacities to be impaired, and where we have *no introspective access to our own impairment*. Whenever this is the case, and wherever thinkers are savvy enough to learn about the psychology of such biases, similar epistemological challenges concerning self-assessment and proper adjustment are likely to arise.

4. CONCLUSION

We had two goals for this paper: to review some of the most compelling empirical work on implicit racial bias, and to gesture at the sorts of normative questions it raises. In particular, we have looked at evidence indicating that implicit racial bias is widespread. There are two major and converging lines of evidence for this. First, there is laboratory evidence, primarily gathered with the IAT and similarly indirect measurement techniques. Second, there are studies that document statistical patterns of behavior in real-world situations, such as the resume and the NBA studies. Numerous other studies, which we exemplified with the work on the 'weapon bias,' have begun explicitly linking performance on the IAT to other activities that are likely to be influenced by implicit biases.

Given the character and prevalence of implicit racial bias, a number of novel normative issues arise. We focused on two of these. First, is implicit racial bias normatively problematic, and if so, how? Perhaps surprisingly, no simple answers to either of these questions are obviously correct or immediately convincing. After separating out moral assessment from issues centering on rationality, we describe some of the normative work that has been done on racism and stereotypes, respectively, and we pointed out where such work can be extended to address implicit racial biases—and where those extant views seem ill-equipped to deal with them. Second, ought each person to believe, of himself, that he is racially biased? Does one have epistemic reason to compensate for implicit racial bias when making more considered, deliberative judgments? On both of these accounts we suggested that—again, perhaps surprisingly—there are powerful arguments indicating that the answer is yes.

NOTES

1. Throughout, we will simplify the discussion by considering just two groups, and using the capitalized terms 'Black' and 'White' to refer those putative racial groups and their members. Other terminology, e.g., 'African-American', is less suitable for our purposes because it is overly restrictive. For example, it does not appear that implicit racial biases against Blacks apply only to Black *Americans*, or only to Americans of specifically *African* descent.
2. See Greenwald, McGhee, and Schwartz for the first presentation of the IAT itself, as well as the initial results obtained with it. Also see Greenwald and Nosek; Lane et al.; Nosek, Greenwald, and Banaji,

'Implicit Association Test' for more recent reviews of data gathered using IATs, and for useful discussions of the methodological issues surrounding the test.

3. For a more detailed and technically precise description of how the IAT works, see any of the papers cited in endnote 2. At the outset of their extensive survey of research based on the IAT (over 4.5 million tests have been taken on the Harvard Web site alone!). . . .
4. The very first study using the IAT found evidence of implicit racial biases in White American undergraduates (Greenwald, McGhee, and Schwartz). Since that initial paper, similar results have been found with disturbing frequency (Banaji; Ottaway, Hayden, and Oakes; see also Lane et al.).
5. Similar dissociations have been found using a wide variety of other indirect measures, including evaluative priming (Cunningham, Preacher, and Banaji; Devine et al.), the startle eyeblink test (Phelps et al.; Amodio, Harmon-Jones, and Devine), and EMG measures (Vanman et al.).
6. A nice discussion of non-voluntary attitudes, and how we can be responsible for them, can be found in Smith.
7. At least, they would not be morally problematic *in the way* that racist attitudes are problematic. Garcia offers an account of racism, not a complete moral theory.
8. See Gregg, Seibt, and Banaji. We have stated that it is more *salient* that implicit attitudes are uncontrollable. That's because, arguably, the acquisition of most *explicit* attitudes is uncontrollable as well; it's just not salient at first glance. One does not control one's acquisition of, for instance, one's beliefs about plants, one's attitudes towards pets, etc. So one will need to appeal to more complex or carefully delineated features—perhaps identification or reasons-responsiveness—if one wants to claim that implicit attitudes are not proper subjects of blame, but that explicit attitudes are.

WORKS CITED

Amadio, D., E. Harmon-Jones, and P. Devine. 'Individual Differences in the Activation and Control of Affective Race Bias as Assessed by Startle Eyeblink Response and Self-Report'. *Journal of Personality and Social Psychology* 84.4 (2003): 738–53.

Banaji, M. R. 'Implicit Attitudes Can Be Measured'. *The Nature of Remembering: Essays in Honor of Robert G. Crowder*. Eds. H. L. Roediger, III, J. S. Nairne, I. Neath, and A. Surprenant. Washington, DC: American Psychological Association, 2001. 117–50.

Bertrand, M. and S. Mullainathan. 'Are Emily and Greg More Employable Than Lakisha and Jamal?: A Field Experiment on Labor Market and Discrimination'. 2003. Poverty Action Lab Paper No. 3. Accessed 25 March 2008 from http://povertyactionlab.org/papers/bertrand_mullainathan.pdf.

Blum, Lawrence. 'Stereotypes and Stereotyping: A Moral Analysis'. *Philosophical Papers* 3 (2004): 251–89.

Cunningham, W., K. Preacher, and M. Banaji. 'Implicit Attitude Measures: Consistency, Stability, and Convergent Validity'. *Psychological Science* 12.2 (2001): 163–70.

Devine, P., et al. 'The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice'. *Journal of Personality and Social Psychology* 82.5 (2002): 835–48.

Fischer, J. and M. Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge UP, 1998.

Frankfurt, H. 'Freedom of the Will and the Concept of a Person'. *Journal of Philosophy* 68.1 (1971): 5–20.

Garcia, J. L. A. 'Three Sites for Racism: Social Structures, Valuings and Vice'. *Racism in Mind*. Eds. M. P. Levine and T. Pataki. Ithaca, NY: Cornell UP, 2004. 36–55.

Gehring, W. J., A. Karpinski, and J. I. Hilton. 'Thinking About Interracial Interactions'. *Nature Neuroscience* 6 (2003): 1242–3.

Green, A., et al. 'Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients'. *Journal of General Internal Medicine* 22.9(2007): 1231–38.

——, D. McGhee, and J. Schwartz. 'Measuring Individual Differences in Implicit Cognition: The Implicit Association Test'. *Journal of Personality and Social Psychology* 74.6 (1998): 1464–80.

——, and A. Nosek. 'Health of the Implicit Association Test at age 3'. *Zeitschrift fur Experimentelle Psychologie* 48 (2001): 85–93.

——, B. Nosek, and R. Banaji. 'Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm'. *Journal of Personality and Social Psychology* 85 (2003): 197–216.

Gregg, A. P., B. Seibt, and M.R. Banaji. 'Easier Done than Undone: Asymmetry in the Malleability of Implicit Preferences'. *Journal of Personality and Social Psychology* 90 (2006): 1–20.

Kruger, J. and D. Dunning, 'Unskilled and Unaware of it: How Difficulties in recognizing One's Own Incompetence

Lead to Inflated Self-Assessments'. *Journal of Personality and Social Psychology* 77.6 (1999): 1121–34.

Lane, K., et al. 'Understanding and Using the Implicit Association Test: IV'. *Implicit Measures of Attitudes*. Eds. B. Wittenbrink, and N. Schwarz. New York, NY: The Guilford Press, 2007. 59–102.

McConahay, J. 'Modern Racism, Ambivalence, and the Modern Racism Scale'. *Prejudice, Discrimination, and Racism*. Eds. J. F. Dovidio and S. L. Gaertner. Orlando, FL: Academic Press, 1986.

McConnell, A. R. and J. M. Leibold. 'Relations between the Implicit Association Test, Explicit Racial Attitudes, and Discriminatory Behavior'. *Journal of Experimental Social Psychology* 37 (2001): 435–42.

Mele, A. *Self-Deception Unmasked*. Princeton, NJ: Princeton UP, 2001.

Nosek, B. A., A. G. Greenwald, and M. R. Banaji. 'The Implicit Association Test at Age 7: A Methodological and Conceptual Review'. *Automatic Processes in Social Thinking and Behavior*. Ed. J. A. Bargh. Philadelphia, PA: Psychology Press, 2007.

Ottaway, S. A., D. Hayden, and M. Oakes. 'Implicit Attitudes and Racism: The Role of Word Familiarity and Frequency in the Implicit Association Test'. *Social Cognition* 18.2 (2001): 97–144.

Payne, B. K. 'Conceptualizing Control in Social Cognition: The Role of Automatic and Controlled Processes in Misperceiving a Weapon'. *Journal of Personality Social Psychology* 81 (2005): 181–92.

——. 'Weapon Bias: Split-Second Decisions and Unintended Stereotyping'. *Current Directions in Psychological Science* 15 (2006): 287–91.

Phelps, E., et al. Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation'. *Journal of Cognitive Neuroscience* 12.5 (2000): 729–38.

Price, J. and J. Wolfers. 'Racial Discrimination among NBA Referees'. Manuscript. Accessed 25 March 2008 from http://bpp.wharton.upenn.edu/jwolfers/Papers/NBARace(NBER).pdf.

Roedder, E. 'Savvy Thinking' and 'The Epistemology of Self-Correction for Implicit Bias'. *Beings Like Us: Deliberating in Light of Psychological Theory*. Ph.D. dissertation, New York University, Department of Philosophy. In preparation.

Rudman, L. and M. Lee. 'Implicit and Explicit Attitudes toward Female Authority'. *Group Processes and Intergroup Relations* 5 (2002): 483–94.

Smith, A. M. 'Responsibility for Attitudes: Activity and Passivity in Mental Life'. *Ethics* 115.2 (2005): 236–71.

Vanman, E. J., et al. 'The Modern Face of Prejudice and Structural Features that Moderate the Effect of Cooperation on Affect'. *Journal of Personality and Social Psychology* 73 (1997): 941–59.

READING QUESTIONS

1. Summarize briefly the laboratory evidence gathered using IAT (implicit attitude test).
2. What other empirical evidence do Kelly and Roedder summarize?
3. What evidence do the authors cite against those who might say that implicit biases have little or no influence on outward behavior?
4. How do the authors argue that racially biased attitudes and thoughts (apart from the behavior they might cause) are "in themselves" bad?
5. According to the authors, what is the main difference between reasons given for affirmative action and reasons that would justify a professor adjusting her students' grades in light of thinking that she is subject to implicit racial bias (the "savvy grader")?

DISCUSSION QUESTIONS

1. How might a professor guard against implicit racial bias other than adjusting upward the grades given to students whose race differs her own?
2. Should a person be held blameworthy for having bad attitudes they are completely unaware of? If not, why not? If so, why?
3. Suppose you become convinced by this article that you very likely have implicit racial biases that are bad to have. Are there steps you might take to rid yourself of them? What might those be?