

To appear in *Philosophy Compass*

# Racial Cognition and the Ethics of Implicit Bias

By

Daniel Kelly and Erica Roedder

## **Abstract:**

We first describe recent empirical research on racial cognition, particularly work on implicit racial biases that suggests they are widespread, that they can coexist with explicitly avowed anti-racist and tolerant attitudes, and that they influence behavior in a variety of subtle but troubling ways. We then consider a cluster of questions that the existence and character of implicit racial biases raise for moral theory. First, is it morally condemnable to harbor an implicit racial bias? Second, ought each of us to suspect ourselves of racial bias, and therefore correct for it in ordinary activity, such as grading student papers?

## **1. Introduction**

Questions about race arise in a number of different areas of philosophy, from debates about the metaphysics of race itself (e.g. Appiah 1995, Mallon 2004, 2006) to discussions in social morality about how to best deal with racial categories in our institutions and beyond (e.g. Outlaw 1996, Wasserstrom 2001). Independently, a literature on the psychology of race, which looks at how people intuitively conceptualize races and racial membership, has been flourishing in various areas of cognitive science, including developmental, evolutionary and social psychology. Much of this research has focused on the psychological underpinnings of *thought* about race, including racial classification and racial evaluation, as well as the effect these might have on behavior. While some have begun drawing out philosophical implications from the findings on racial cognition (Kelly et al forthcoming, Machery and Faucher 2005b), we believe there is much fertile ground that remains to be explored.

The aim of this paper is twofold. 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 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).<sup>1</sup> 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

---

<sup>1</sup> 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.

Opportunity Employer” in their ad discriminated just as much as other employers (Bertrand and Mullainathan 2003).

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”), 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 & Wolfers, ms).

Intriguing – and troubling – as they are, real world, behavior-based studies such as these make up only one of many windows onto our subject of interest, *racial cognition*.

Unfortunately, we do not have the space to provide an overview of the full breadth of empirical work being done on this subject matter. Instead, we will focus our discussion on implicit racial bias, but in doing so we are forced to leave to the side other important work by psychologists and anthropologists. Of particular philosophical interest is research done by psychologists who take an evolutionary perspective as their point of departure (Hirshfeld 1996, 2001; Kurzban et al. 2001; Gil-White 1999, 2001a, 2001b), and the ways in which philosophers have used this perspective to begin integrating social constructivist explanations with more psychologically grounded explanations of race (Machery and Faucher 2005a, 2005b, ms; also see Mallon 2004,

2006 for an philosophically sophisticated discussion of social construction and race) and to shed new light on more purely normative debates about how to best deal with racial categories and the evaluations that presuppose them (Kelly et al. forthcoming). Despite our selective focus in what follows, we think these are all thought provoking issues, and encourage the reader concerned with any of the aspects of race to look into these other exciting areas of research on racial cognition as well.

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.<sup>2</sup> Rather than provide a technical description of how the test works, it will be more useful to convey its flavor. Suppose you have

---

<sup>2</sup> See Greenwald et al. 1998 for the first presentation of the IAT itself, as well as the initial results obtained with it. Also see Greenwald & Nosek 2001, Lane et al. 2007 and Nosek et al. 2007 for more recent reviews of data gathered using IATs, and for useful discussions of the methodological issues surrounding the test.

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.<sup>3</sup>

---

<sup>3</sup> For a more detailed and technically precise description of how the IAT works, see any of the papers cited in footnote 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 website alone!), Lane et al. provide a more concise characterization:

“The IAT measures the relative strength of association between pairs of concepts, labeled for pedagogic purposes as *category* and *attribute*. When completing an IAT, participants rapidly classify individual

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.<sup>4</sup> 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 1986), 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

---

stimuli that represent category and attribute (in the form of words, symbols, or pictures) into one of four distinct categories with only two responses. The underlying assumption is that responses will be facilitated – and thus will be faster and more accurate – when categories that are closely associated share a response, as compared to when they do not.” (Lane et al. 2007, page 62)

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/selectatest.html>.

<sup>4</sup> More precisely: “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.” (Nosek et al. 2007, page 267).

biases.<sup>5</sup> 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).<sup>6</sup> 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 et al. 2003). 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 & Lee 2002), 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 & Leibold 2001). 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. ms, as cited in Lane et al. 2007). Moreover, in research on intergroup

---

<sup>5</sup> The very first study using the IAT found evidence of implicit racial biases in White American undergraduates (Greenwald et al. 1998). Since that initial paper, similar results have been found with disturbing frequency (Banaji 2001, Ottaway et. al. 2001, see also Lane et al. 2007).

<sup>6</sup> 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).

discrimination (including racial discrimination), the IAT was found to be more predictive than self-report.<sup>7</sup> 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 2006). 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 2005). 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,

---

<sup>7</sup> There is a growing literature on these issues. In particular, see Greenwald et al. ms for an overview of similar studies connecting implicit biases with behavior and judgment; also see Nosek et al. 2007, Lane et al. 2007 and Kang & Banaji 2006 for discussion of the significance of such findings.

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 essay?

*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.<sup>8</sup> 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.

---

<sup>8</sup> We know of no efforts to answer this question, although it is posed in (Jolls and Sunstein 2006).

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.”<sup>9</sup> 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.<sup>10</sup>

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 (2004) 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

---

<sup>9</sup> In stating these examples, we felt *extremely* uncomfortable, and we anticipate that our readers will feel the same way. However, we think concrete examples are needed in order to make salient our point: in general, explicitly racist attitudes—even if they are not acted upon—are morally damnable.

<sup>10</sup> A nice discussion of non-voluntary attitudes, and how we can be responsible for them, can be found in Smith (2005).

*behavior* are misguided. As another example, Blum (2004) writes that that “...false [stereotypical] beliefs can be bad even if they do not contribute harm to their target” (262).<sup>11</sup>

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 et al. 2005). But if the

---

<sup>11</sup> Both authors focus on mental states that are quite different from the implicit attitudes measured by the IAT. We discuss their views in more depth below.

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. For instance, Corlett (2003) writes:

“...epistemically speaking, one has a duty to eschew error and pursue truth. And one also has a moral duty to be epistemically responsible (in the dutiful sense). To the extent that racist beliefs are false representations of self and/or others, one's failures to at least earnestly attempt to rid them from one's belief system constitutes a failure to live up to one's epistemic and moral duty” (68)

Here, Corlett ties one's epistemic and moral duties closely together, leaving room for the view that *only* those attitudes stemming from epistemic failures are immoral.<sup>12</sup>

---

<sup>12</sup>We say he is “leaving room” because it is not clear whether Corlett endorses such a view or not. He is certainly saying that we sometimes have a moral duty to be rational, and that our attitudes can be morally wrong if they spring from irrational tendencies. But it is not clear if, on his view, epistemic dutifulness exhausts our moral duties with

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 3<sup>rd</sup> 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*.<sup>13</sup> 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.<sup>14</sup> 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.

---

respect to our attitudes themselves. Perhaps he thinks that, in addition, we have moral duties to have benevolent or respectful attitudes, as suggested by Garcia and Blum, respectively.

<sup>13</sup> Arguably, Elisa's parents have much more evidence than her teacher does, so the epistemic conditions of the teacher and parent are different. That's okay; we need only the point that *sometimes* evidence that is rationally persuasive for a teacher might also be rationally persuasive for the parent.

<sup>14</sup> In making this argument, our example assumes that one's evidence and background beliefs can sometimes support multiple rational conclusions. This makes it easier to stomach the thought that one might be morally compelled *not* to form a belief that would be rationally justified. After all, if there are multiple rational conclusions available, one can follow the moral compulsion not to form rational belief P, and instead form belief Q, *where Q is also rational*. But at least one of us (ER) thinks this can be taken one step further: moral considerations can sometimes compel us to believe things that are not rational. Considering the case of race, it seems intuitive that, even if it were rational to believe Black children were dumber than White children, it would still be morally repugnant to do so.

As a second point, suppose we were to grant, for the sake of argument, the suggestion that rational attitudes are moral and irrational ones are immoral. Even on this supposition, a case can still be made that implicit racial biases are morally problematic. We suspect that such associations (such as those found in studies on the weapon biases) almost always extend beyond what is rational, and there will almost always be a “remainder”: an implicit association that goes beyond what rationality endorses. If this is right, then even on the supposition that morality and rationality are tightly bound together, implicit attitudes will remain morally problematic to the extent that they outstrip what is rationally justifiable. Thus, the appeal to rationality would only partially mitigate the moral wrongness of having implicit racial attitudes.

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?<sup>15</sup> That is, can current accounts of what makes racist attitudes wrong, accounts that usually focus on explicit and conscious attitudes, be

---

<sup>15</sup> Of course, authors writing on racism have discussed unconscious racism, including Blum and Garcia. Garcia states explicitly that one’s racism might be unconscious, e.g. you may not know why you don’t take the elevator (43). Blum suggests that stereotyping may occur at a level below that of belief, e.g. when a woman unconsciously grabs her purse as a Black man passes by (266). But neither of these authors go into depth about unconscious attitudes, which leaves open the question of whether their accounts really generalize to unconscious or implicit attitudes in any straightforward way.

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 (2004) account of racism and Blum's (2004) 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<sup>16</sup>.

Suppose, on the other hand, some implicit attitudes are indeed affectively laden, as a growing body of empirical research suggests (e.g. Vanman et al. 1997, Phelps et al. 2000, Amadio et al. 2003, see also Payne et al. in press). This possibility raises a different kind of difficulty, which turns on whether such implicit attitudes should be thought of as "inherently contrary to the moral virtue of benevolence." While it is obvious that explicit, hate-filled racial rage is intrinsically opposed to benevolence, it is far less clear whether the more subtle attitudes measured by the IAT ought to be categorized in this way. One lesson to draw from this is that it is much easier to apply philosophy's normative categories (e.g., "intrinsically opposed to benevolence") to robust, explicit mental states (e.g., feelings of hostility, what we care about, etc.), than to implicit ones. As a result, there is real and important philosophical work to be done in figuring out whether and how these normative categories can be extended into the realm of implicit attitudes.

Let us turn now to Blum's (2004) 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 latter 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

---

<sup>16</sup> At least, they would not be morally problematic *in the way* that racist attitudes are problematic. Garcia (2004) offers an account of racism, not a complete moral theory.

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.<sup>17</sup> These features, it might be thought, are related to features that establish blameworthiness—such as identification (Frankfurt 1971) or reasons-responsiveness (Fischer and Ravizza 1998). 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?

---

<sup>17</sup> See Gregg et al. 2006. 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.

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 ms).

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.<sup>18</sup> 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

---

<sup>18</sup> There are, of course, many ways one might go about compensating for implicit racial bias. Most obviously, one might use conscious rules, e.g. "Bump up borderline grades of Black students." In addition, there are various psychological techniques which seem to mitigate implicit racial bias, such as entertaining counterstereotypic thoughts (e.g. imagining a positive Black role model, or a female scientist). Some of these are discussed in Faucher and Machery (ms).

paper based on its merits, it is reasonable for the savvy grader to correct for the effects of racial biases.<sup>19</sup>

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. One of us (Roedder ms) 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

---

<sup>19</sup> People sometimes question the idea that grades are apt to be “accurate” at all. In many ways, this is irrelevant—most of us want to avoid having the race of a student affect their grading. That desire is enough to motivate the problems we raise here: if one has this desire, it seems that one is rationally compelled to correct for possible influence of race in deliberation.

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 2001 for a lucid and eye opening discussion of the prevalence of self-deception, and Kruger and Dunning 1999 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.<sup>20</sup> 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.<sup>21</sup> 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:

---

<sup>20</sup> In an extensive search of two databases (PsychInfo and ERIC), and after consulting with several psychologists and one education researcher, we were able to locate only three small studies after the 1970s dealing directly with racial bias in the assessment of college level work (Ballantyne & Sparks 1991, Dorsey & Colliver 1995, and Amodio 2006). Of these, the most suggestive is Amodio 2006, which asked subjects to evaluate a fellow student on the basis of "his" essay (in truth, all essays were identical). Subjects were given demographic data about the essay's purported author and asked whether they thought the author of the essay was unintelligent, lazy, or had other stereotypical features. Amodio found that IAT scores predicated a racial bias in responding to these questions; students with high IAT scores were more likely to think that, if they were told the author was Black, he was unintelligent, lazy, etc.

<sup>21</sup> This point does not mitigate the importance of running such studies. In the absence of evidence, we argue, it is rational to act guided by your best-informed hunches as to how the studies would turn out. But this is a stop-gap measure, to be used until one can correct for the absence of evidence.

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*?<sup>22</sup> 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.<sup>23</sup> 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.

---

<sup>22</sup> To add to the case we are making, one can appeal to expert opinion. In that vein, we have discussed this issue with two members of Banaji's lab at Harvard, both of whom said they'd be "very surprised" if there wasn't implicit racial bias in the domain of grading.

<sup>23</sup> This version of epistemic argument is highly compressed, thus there isn't room to respond to a number of important objections. Roedder (ms) contains a much fuller exploration and defense of these claims.

#### **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 implicit attitude test (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 described 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.

## References

- Amadio, D., Harmon-Jones, E., and Devine, P. 2003. "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): 738-753.
- Amodio, Devine. 2006. "Stereotyping and Evaluation in Implicit Race Bias: Evidence for Independent Constructs and Unique Effects on Behavior." *Journal of Personality and Social Psychology*, Vol. 91, No. 4, 652-661.
- Appiah, K. A. 1995. The Uncompleted Argument: Du Bois and the Illusion of Race. In L. A. Bell and D. Blumenfeld (1995). pp. 59-78.
- Ballantyne, R. and Sparks, R. 1991. "Assessment Training in Geography Education." *Journal of Geography in Higher Education*, Vol. 15, No. 2.
- Banaji, M. R. 2001. "Implicit Attitudes Can Be Measured." In H. L. Roediger, III, J. S. Nairne, I. Neath, and A. Surprenant (eds.), *The Nature of Remembering: Essays in Honor of Robert G. Crowder*. Washington, DC: American Psychological Association. pp. 117-150.
- Bertrand, M. and Mullainathan, S. 2003. "Are Emily and Greg More Employable Than Lakisha and Jamal?: A Field Experiment on Labor Market and Discrimination." Poverty Action Lab Paper No. 3. [http://povertyactionlab.org/papers/bertrand\\_mullainathan.pdf](http://povertyactionlab.org/papers/bertrand_mullainathan.pdf)
- Blum, Lawrence. "Stereotypes and Stereotyping: A Moral Analysis." *Philosophical Papers*, 3: 251-289.
- Corlett, J. Angelo. 2003. *Race, Racism, and Reparations*. Ithaca: Cornell UP.
- Cunningham, W., Preacher, K and Banaji, M. 2001. "Implicit Attitude Measures: Consistency, Stability, and Convergent Validity." *Psychological Science*, 12(2): 163-170.
- Devine, P., Plant, E., Amodio, D., Harmon-Jones, E. and Vance, S. 2002. "The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice." *Journal of Personality and Social Psychology*, 82(5): 835-848.
- Dorsey, K. and Colliver, J. 1995. "Effect of anonymous test grading on passing rates as related to gender and race." *Academic Medicine*, 70(4): 321-323.
- Fischer, J., and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
- Faucher, L. and Machery, E. Untitled Manuscript.
- Frankfurt, H. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy*, 68(1): 5-20.
- Gehring, W. J., Karpinski, A., and Hilton, J. I. 2003. "Thinking About Interracial Interactions." *Nature Neuroscience*, 6: 1242-1243.
- Gil-White, F. 1999. "How Thick is Blood? The Plot Thickens ... : If Ethnic Actors are Primordialists, What Remains of the Circumstantialists/Primordialists Controversy?" *Ethnic and Racial Studies*, 22(5): 789-820.
- Gil-White, F. 2001a. "Are Ethnic Groups Biological 'Species' to the Human Brain?" *Current Anthropology*, 42(4): 515-554.
- Gil-White, F. 2001b. "Sorting is Not Categorization: A Critique of the Claim that Brazilians Have Fuzzy Racial Categories." *Cognition and Culture*, 1(3): 219-249.
- Green, A., Carney, D., Pallin, D., Iezzoni, L., & Banaji, M. Manuscript. "Physician's implicit biases predict differential treatment of Black versus White patients."
- Greenwald, A. and Banaji, M. 1995. "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes." *Psychological Review*, 102(1): 4-27.

- Greenwald, A., McGhee, D. and Schwartz, J. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology*, 74(6): 1464-1480.
- Greenwald, A., Nosek, B. and Banaji, R. 2003. "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm." *Journal of Personality and Social Psychology*, 85: 197-216.
- Greenwald, A., Poehlman, T., Uhlmann, E. and Banaji, M. Manuscript. "Understanding and Using the Implicit Association Test: III. Meta-analysis of Predictive Validity."
- Gregg, A. P., Seibt, B., and Banaji, M. R. 2006. "Easier Done than Undone: Asymmetry in the Malleability of Implicit Preferences." *Journal of Personality and Social Psychology*, 90: 1-20.
- Hirschfeld, L. W. 1996. *Race in Making: Cognition, Culture, and the Child's Construction of Human Kinds*. Cambridge, MA: MIT Press.
- Hirschfeld, L. W. 2001. "On a Folk Theory of Society: Children, Evolution, and Mental Representations of Social Groups." *Personality and Social Psychology Review*, 5(2): 107-117.
- Jolls, C., and Sunstein, C. 2006. "The Law of Implicit Bias." *California Law Review*, 94: 969-996.
- Kang, J., and Banaji, M. 2006. "Fair Measures: A Behavioral Realist Revision of 'Affirmative Action'." *California Law Review*, 94: 1063-1118.
- Kelly, D., Machery, E., and Mallon, R. Forthcoming. "Racial Cognition and Normative Racial Theory." In Doris, J., Mallon, R., Nichols, S. and S. Stich, (eds.), *The Oxford Handbook of Moral Psychology*. Cambridge: Oxford University Press.
- Kruger, J. & Dunning, D. 1999. "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): 1121-1134.
- Kurzban, R., J. Tooby and L. Cosmides. 2001. "Can Race Be Erased? Coalitional Computation and Social Categorization." *Proceeding of the National Academy of Science*, 98(26): 15387-15392.
- Lane, K., Banaji, M., Nosek, B., and Greenwald, A. 2007. "Understanding and Using the Implicit Association Test: IV." In Wittenbrink, B., and N Schwarz, (eds.), *Implicit Measures of Attitudes*. New York: The Guilford Press.
- Machery, E., and Faucher, L. 2005a. "Why do we Think Racially?" In H. Cohen and C. Lefebvre (eds.), *Handbook of Categorization in Cognitive Science*. Orlando, FL, Elsevier.
- Machery, E., and Faucher, L. 2005b. "Social Construction and the Concept of Race." *Philosophy of Science*, 72: 1208-1219.
- Machery, E., and Faucher, L. Ms. "Concepts of Races are Biological: A Cross-Cultural Study."
- Mallon, R. 2004. "Passing, Traveling, and Reality: Social Construction and the Metaphysics of Race." *Noûs*, 38(4): 644-673
- Mallon, R. 2006. "'Race': Normative, Not Metaphysical or Semantic." *Ethics*, 116(3): 525-551.
- McConahay, J. 1986. "Modern Racism, Ambivalence, and the Modern Racism Scale." In J. F. Dovidio and S. L. Gaertner (eds.), *Prejudice, Discrimination, and Racism*, Orlando, FL: Academic Press.
- McConnell, A. R., Leibold, J. M. (2001) "Relations between the Implicit Association Test, explicit racial attitudes, and discriminatory behavior." *Journal of Experimental Social Psychology* 37: 435-442.

- Mele, A. 2001. *Self-Deception Unmasked*. Princeton, NJ: Princeton University Press.
- Nosek, Greenwald, and Banaji. 2005. Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity. *Personality and Social Psychology Bulletin*, Vol. 31 No. 2, 166-180.
- Nosek, B. A., Greenwald, A. G., and Banaji, M. R. 2007. "The Implicit Association Test at Age 7: A Methodological and Conceptual Review." In J. A. Bargh (ed.), *Automatic Processes in Social Thinking and Behavior*. Philadelphia, PA: Psychology Press.
- Ottaway, S. A., Hayden, D. and Oakes, M. 2001. "Implicit Attitudes and Racism: The Role of Word Familiarity and Frequency in the Implicit Association Test." *Social Cognition*, 18(2): 97-144.
- Outlaw, L. 1996. *On Race and Philosophy*. New York: Routledge.
- Payne, B.K. 2005. "Conceptualizing Control in Social Cognition: The Role of Automatic and Controlled Processes in Misperceiving a Weapon." *Journal of Personality Social Psychology*, 81: 181-192.
- Payne, B.K. 2006. "Weapon Bias: Split-second Decisions and Unintended Stereotyping." *Current Directions in Psychological Science*, 15: 287-291.
- Payne, B.K., Cheng, C., Govorum, O. & Stewart B. In Press. "An inkblot for attitudes: Affect misattribution as implicit measurement." *Journal of Personality and Social Psychology*.
- Phelps, E., O'Connor, K., Cunningham, W., Funyama, S., Gatenby, C., Core, J. and Banaji, M. 2000. "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation." *Journal of Cognitive Neuroscience*, 12(5): 729-38.
- Price, J. and Wolfers, J. Manuscript. "Racial Discrimination Among NBA Referees."
- Roedder, E. Manuscript. "Savvy thinking" and "The epistemology of self-correction for implicit racial biases."
- Rudman, L. and Lee, M. 2002. "Implicit and explicit attitudes toward female authority." *Group Processes and Intergroup Relations*, 5: 483-494.
- Smith, A. M. 2005. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics*, 115, 2: 236-271.
- Vanman, E.J., Paul, B.Y., Ito, T.A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology*, 73, 941-959.
- Wasserstrom, R. 2001. "Racism and Sexism." *Philosophy and Social Issues: Five Studies*. Notre Dame, IN: Univ of Notre Dame Press. Reprinted in (ed. B. Boxill) *Race and Racism*. New York. Oxford University Press. Pp. 307-343.