Am I A Racist? Implicit Bias and the Ascription of Racism

Abstract: There is good evidence that many people harbor attitudes that conflict with those they endorse. In the language of social psychology, they seem to have implicit attitudes that conflict with their explicit beliefs. There has been a great deal of attention paid to the question whether agents like this are responsible for actions caused by their implicit attitudes, but much less to the question whether they can rightly be described as (say) racist in virtue of harboring them. In this paper, I attempt to answer this question using three different standards, providing by the three dominant kinds of accounts of racism (doxastic, behavioral and affective). I argue that on none of these accounts should agents like this be described as racists. However, it would be misleading to say, without qualification, that they are not racists. On none of these accounts are agents like this entirely off the hook.

Over the past fifty years, *explicit* racism has declined significantly in most Western societies. Since 1972, the General Social Survey at the University of Chicago has polled Americans' attitudes on many topics, race included. On the measures used by the GSS, racism has fallen quite dramatically. For example, in 1972 31% of white Southerners supported segregated schools; by 1985 the question had been removed since there were so few people willing to express support for the policy. In 1972, 48% of white Southerners reported that they would not be willing to vote for a Black candidate for the presidency. In 2010, that number had fallen to 6% (Barry-Jester 2015).

Despite this fall, racial injustice very obviously persists. Of course, injustice is far from entirely dependent on attitudes for its persistence: it has an inertia because it is inscribed into the structure of institutions and into practices. Nevertheless, there is evidence for racial discrimination, in addition to institutionally mediated racism. For instance, when otherwise identical CVs of minority and majority applicants are submitted to potential employers (with apparent group membership manipulated by using stereotypical names), the minority candidates get fewer callbacks from potential employers and fewer invitations to interviews (Dovidio & Gaertner 2000; Oreopoulos & Dechief 2011). Since only the names varied across the CVs – and therefore the candidates had the same level of education, age, work experience, and so on – this looks like direct evidence of racially motivated discrimination.

What explains the apparent disparity between self-reported attitudes and behavior? No doubt, part of the explanation is that people are sometimes unwilling to report their real attitudes to interviewers. People may be more racist than they are willing to admit. But the very fact that they feel it is unacceptable to report racist opinions is itself a signal of an important shift in attitudes: it is implausible that the unacceptability of expressing racist attitudes has come to be established and persist while a very large minority of people remain deeply explicitly racist. While self-presentation effects undoubtedly play a role, it is very likely that what psychologists call *implicit* attitudes explain some of the disparity between reported attitudes and behavior.[1]

Implicit attitudes belong to (or perhaps constitute) the set of content-bearing states that influence cognition and thereby behavior in ways that escape direct personal-level control, and without our being able to have direct access to their contents. We might best characterize them within the popular dual-process framework of psychological explanations (Evans 2008). According to this framework, explicit attitudes are attitudes that have contents that are introspectible, and which control personal-level cognition. They may, for instance, serve as premises in explicit reasoning. Implicit attitudes, in contrast, are not able to be deployed at the personal-level; rather, they influence cognition in ways that escape conscious control. Nor can their contents be introspected: though there is some evidence that agents (at least sometimes) come to be aware of the contents of their implicit attitudes (Nier 2005; Ranganath, Smith & Nosek 2008), access may be by inference rather than introspection.[2] They have a disproportionate effect on behavior when we do not, or cannot (because we lack cognitive resources: we are tired, stressed, or under cognitive load, or because we are required to respond too quickly for effortful processing) exercise personal-level control.

There are a variety of ways of probing implicit attitudes. The best known (by far) is the *implicit association test*, which uses speeded response to probe implicit attitudes (Greenwald, McGhee & Schwartz 1998; Greenwald et al. 2009). The IAT requires subjects to sort rapidly

presented stimuli (for instance photographs of Black and White people and positively and negatively valenced words) into assigned categories. For instance, on block one of an IAT, participants may be instructed to press one key when a Black face or a negatively valenced word (“death”; “ruin”; “fear”; “cancer”) is presented and another key when a White face or a positively valenced word “joy”; “flower”; “puppy”; “sunshine”) is presented. On subsequent blocks, the pairings shift: Black/positive and White/negative are paired. The much replicated finding is that the majority of White Americans are quicker to respond when the pairing is White/positive and Black/negative than when the pairing is White/negative and Black/positive. This has been taken as evidence that they *associate* Black people with negatively valenced properties and White with positive.

There are a variety of other ways to measure implicit attitudes. For example, sequential priming tasks may be used, like the affect misattribution procedure. In the AMP, the prime is (once again) a photograph of a Black or White person. Subsequent to very brief (usually around 75 milliseconds) presentation of the prime, a picture that has previously been rated by independent raters as neutral is presented (Chinese pictograms are standardly used), and the participant is asked to rate the picture’s attractiveness. White Americans typically rate the pictures as less pleasant after the presentation of a Black face than a White face, apparently as a result of misattributing feelings (or nonconscious correlates of feelings) aroused by the faces to the pictures (Payne et al. 2005).

IAT and AMP responses are a partially controlled behavior (see Huebner 2016). Speeded response, misdirection and cognitive load are all measures that are designed to reduce the influence of explicit processes relative to automatic processes, but almost all of our waking behavior is produced by some mix of both. We almost always exercise some degree of personal-level control over behavior that is influenced by our implicit attitudes. Nevertheless, their effects may be substantial. They may, for instance, cause us to confabulate what seem good reasons for our behavior, such that we see no reason to oppose their influence (see Levy

2014 for discussion). For this reason, IAT and AMP performance correlate with behavior. For instance, controlling for explicit prejudice, performance on the AMP predicted a greater likelihood of not voting for Obama in the 2008 election (Payne et al. 2010). It also predicts the future voting behavior of presently undecided voters (Lundberg & Payne 2014). The fact that these attitudes correlate with, and plausibly have a role in causing, behavior makes understanding them a pressing task. Latterly, concern with their nature and role has extended beyond psychology and into philosophy.

Philosophers have concentrated especially on the nature of implicit attitudes, and on our responsibility for the actions they cause (see, in particular, Brownstein & Saul 2016a; 2016b). But the question I want to focus on here has received very much less attention. It is this: are those agents who exhibit a bias against Black people on the IAT or the AMP appropriately described as racists, in virtue of this fact (that is, regardless of their explicit attitudes)? While a number of writers have touched on the topic, few have paid it sustained attention.[3] Here racism stands in for a larger class of prejudiced views: sexism, homophobia, transphobia, and so on. Since we have implicit biases against members of other minority or stigmatized groups, the answer to the question whether it is appropriate to describe people with this set of implicit biases as racists should generalize to these other normatively significant attitudes. I focus on racism because it is obviously extremely important, because there is a large amount of experimental and philosophical work on it, and for a personal reason: because *I* exhibit moderate bias on the IAT. The question is therefore an urgent one for me, but also of broad interest.

I will argue that the answer to the question "am I racist" is not simple. In virtue of my explicit attitudes, there are grounds for saying that I am emphatically *not* a racist. But I don't get off the hook entirely: in virtue of my implicit attitudes, there are strong grounds for a very unemphatic "yes" to the question. It would be very misleading to describe me as a racist, but – insofar as the assessment of my attitudes alone is concerned – I warrant some of the

opprobrium that attaches to the label. That fact gives us – gives me – a reason to attempt to change my attitudes, somewhat independent of worries about how they might influence my reasoning and my behavior.

I will attempt to answer the question "am I racist" by reference to the standard provided by existing accounts of racism. Accordingly, I will devote the rest of this section to a very brief outline of central features of the main competing models. My sketch of these models elides many of their important features and differences between members of each class. I am concerned only to highlight the features that they hold to be central, inasmuch as these features may provide the standard we need to answer our question.

Accounts of the nature of racism have focused on three kinds of features: (1) cognitive attitudes (beliefs), (2) noncognitive attitudes like hatred or contempt, and (3) behavior (Schmid 1996; Faucher & Machery 2009; Atkin 2012).[4] While most authors recognize that none of these features is sufficient to account for all instances of racism by itself, different authors emphasize different features. For instance, at least in his earlier work on the topic, Appiah identifies racism with a set of beliefs, together with a disposition to act on these beliefs. To be a racist is, inter alia, to believe that races exist and that they are morally significant, combined with an irrational resistance to evidence against these beliefs (Appiah 1990; see D'Souza 1995 for another doxastic account). In an influential set of writings, in contrast, Garcia has emphasized certain affects, in particular malevolence or callous indifference toward members of certain racial groups (Garcia 1996; 2011). Finally, some writers defend behavioral accounts of racism. For Fredrickson (2002), racism exists "only when one ethnic group or historical collectivity dominates, excludes or seeks to eliminate another", on the basis of what it perceives to be racial differences.

Almost all prima facie plausible accounts of racism are to some extent hybrid.[5] Appiah, for instance, requires not just belief but also an irrational resistance to belief revision, thereby

allowing noncognitive properties to be partially constitutive of racism; Fredrickson's behavioral model, requires the relevant behaviors to be directed at groups believed to possess certain unalterable characteristics while Blum (2002) present a disjunctive theory, on which racism requires either inferiorization of racialized groups through behavior or the direction of antipathy toward them (or both). Nevertheless, different accounts clearly make different properties central to characterizing racism. Of course, defenders of each account put forward arguments for the superiority of their view over rivals. For the most part, I will ignore these arguments. Instead, I will separately examine each component they emphasize. Are implicit attitudes the kind of mental state that could play the role that beliefs play in accounts of racism? Do they cause or motivate behaviors such that those who harbor them might count as racists on accounts which emphasize overt action? Do they cause, or constitute, affects or noncognitive attitudes with the kinds of content which make affective accounts of racism plausible? I will address these questions one by one.[6]

A final remark on the virtue of approaching the question "am I racist" by focusing on belief, behavior and affect in turn. It is noteworthy that different measures of implicit attitudes do not correlate well with one another (Fazio & Olson 2003; Bar-Anan & Nosek 2014). Further, there is evidence that even when a single measure is used, scores for stereotypical associations may dissociate from scores for affect in one and the same individual (Amodio and Devine 2006). This evidence has sometimes been taken to suggest that it may be at best misleading to speak, as I have above, of implicit biases, as if they were unitary phenomena (Holroyd & Sweetman, 2016). In the face of this evidence, we might adopt a fine-grained account of implicit attitudes, according to which racial implicit attitudes (for instance) divide into multiple states and processes, with different targets and different functional profiles (alternatively, we could follow Machery (2016) and deny that it is appropriate to talk of implicit attitudes at all, on the grounds that the behavior measured in psychology labs is evidence for multitrack dispositions to behave and cognize: traits, not attitudes (on this picture, the relevant traits are constituted by fine-grained states and processes of the kind

identified by Holroyd and Sweetman)). As a matter of fact, I am not persuaded by the case for the heterogeneity in kind of implicit attitudes. Madva and Brownstein (forthcoming) argue convincingly that the dissociations are better explained by differences in content and the relative strength of the affective and semantic components of implicit attitudes than by differences in the kind of attitude involved. However, though (as Holroyd and Sweetman demonstrate), the apparent heterogeneity of implicit attitudes has important implications for how we approach questions of blame and proposals for ameliorating bias, little turns on the question in this context. Because I focus on belief, behavior and affect separately, my approach is sensitive to possible differences between different attitudes. It would be a mistake, in face of the evidence, to ask whether agents are disposed to feel in ways that might be characterized as racist on the basis of evidence that suggests that they harbor stereotypes about members of other races that are not affective. But my approach involves going directly to evidence in the domain in question.[7]

I. The Doxastic Model of Racism.

On the basis of the behavior those who harbor implicit biases evince, it seems reasonable to attribute racist beliefs to them. Someone who is disposed to discount the qualifications of a Black person in virtue of a mental state they possess may be someone who believes *that Black people are less able or less motivated to perform demanding jobs than members of other races*. Someone who is disposed to think that an object in the hands of a Black person is a weapon (Payne 2006) may be someone who believes *that Black people are violent*. These are beliefs in virtue of which it seems plausible to characterize those who harbor them as racist.

As Kelly and Roedder (2008) emphasize, however, ascertaining whether these intuitively plausible claims are in fact true requires detailed investigation of the nature of implicit attitudes. Are such attitudes beliefs? Answering this question is important because proponents of doxastic accounts of racism are surely correct in holding that beliefs play an important

functional role in human behavior; a role that gives them normative significance. Beliefs are states that are responsive to evidence. We should avoid idealizing beliefs: there are many states that seem to qualify as beliefs that fail to respond to evidence, sometimes in striking ways (Bortolotti 2009). The more attached we are to a claim – the more central it is to our identity – the less responsive we are to evidence against it (Ditto & Lopez 1992). At the same time, though, we should be impressed by how evidence responsive most of our beliefs actually are. Our credence in a proposition may drop to zero instantly, upon presentation of decisive evidence against it. Even when we are motivated to believe a proposition, we often come to accept it is false (almost everyone accepts that their bodies are mortal and a substantial number believe that this entails that they will cease to exist entirely at some time in the near-term future). Those thinkers who emphasize our irrationality are (rightly) impressed by how resistant those who are motivated to reject claims are to scientific evidence (with regard to evolution, for example), but even with regard to these prime exhibits, many people do come to accept the evidence despite being motivated to reject it. Beliefs are also apt for inference. Again, we should not idealize beliefs: there may be beliefs that are relatively encapsulated from evidence such that they fail to be apt for certain classes of inferences. But, again, we should be impressed by the aptness for inference that most of them exhibit. They are *inferentially promiscuous*, as Stich (1978) puts it.

If the standard account of the nature of implicit attitudes were correct, we could quickly conclude that they are not beliefs and (therefore) not poised to play the role in behavior that gives belief their normative significance. On the standard account, implicit attitudes are *associations*. Associations are not beliefs. They are not evidence sensitive. They respond to counterconditioning and extinguish slowly when they are not reinforced, whereas beliefs typically respond very rapidly to evidence (Mandelbaum 2016). Moreover, associations are not apt to serve as premises in inferences or to guide behavior in the manner of beliefs. Nothing follows from my association of “salt” and “pepper”, not even when conjoined with a desire for pepper or a belief that excessive salt consumption raises blood pressure.

But implicit attitudes are not mere associations (De Houwer 2014; Mandelbaum 2016). Surveying a wide range of experimental data, Mandelbaum in particular has shown that they do feature in inference. Take (for illustration) the phenomenon of celebrity contagion (Newman, Diesendruck and Bloom 2011).[8] Participants were willing to pay a higher price for a sweater that allegedly has been worn by a celebrity than for an identical item that had not been worn by him. This behavior is neatly explained by an associative account of implicit attitudes: celebrity stardust rubs off on the sweater. But an associative account does not explain why the price participants are willing to pay falls if they are told that the sweater has been laundered since the celebrity wore it. Implicit attitudes to laundering are positive, so if anything an associative account predicts that they ought to be willing to pay more. The kind of reversal of valence seems to require that implicit attitudes be propositionally structured, such that they can feature in inference (Mandelbaum 2016).

Mandelbaum's case for the claim that implicit attitudes are not mere associations is to my mind overwhelming. But it falls short of showing that they are "propositionally structured beliefs," as he claims. Implicit attitudes have some of the features of beliefs, but to a limited extent. For instance, they seem apt to feature in inference or to respond to evidence only in limited domains, in a way that ensures that they fall short of being beliefs. Genuine beliefs are always imperfectly responsive to evidence or apt for inference, but the degree to which implicit attitudes fail in these respects is too extensive for them to count as beliefs.

For instance, implicit attitudes may fail to respond to evidence, or respond perversely, to propositions that are not evidence. For an example of a failure to respond to evidence, consider Gregg, Seibt and Banaji (2006). They gave their subjects information about the members of two novel groups. Members of one group did mainly positive things while members of the other did mainly negative things: this is known to be an effective procedure for inducing implicit attitudes in participants. In one condition, participants were then told

that there had been an error in the induction: the behavior attributed to the first group was actually performed by the second and vice-versa. This manipulation was sufficient to reverse participants' explicit attitudes about the groups, but not their implicit. That's a failure of attitudes to respond to evidence. For an example of perverse responsiveness, consider Han et al. (2006). They had children learn facts about a Pokemon character and then had them watch a video in which other children expressed beliefs about the character that were inconsistent with what they had learned. The subjects rejected the opinions expressed by the children in the video, but the knowledge that these opinions were false did not prevent them from altering the children's implicit attitudes.

More persuasively still, consider Rozin, Markwith & Ross (1990), which is actually cited by Mandelbaum as evidence in favor of his view. In this experiment, subjects preferred drinks sweetened with sugar from a jar labeled "sucrose, table sugar" to those sweetened with sugar from a jar labeled "not sodium cyanide, not poison". The most plausible explanation of this preference turns on a failure of the relevant mechanisms to respond to the negation in the label "not poison". There is independent evidence that implicit processes are blind to negation (Wegner 1984; Deutsch, Gawronski & Strack 2006; Hasson & Glucksberg 2006). Systematic insensitivity to negations is a dramatic restriction of aptness for inference.

The best explanation of the full range of experimental data is that implicit attitudes are neither mere associations nor beliefs: they are apt to feature in some inferences under some conditions, but they lack the kind of rich and systematic propositional structure that entails evidence sensitivity and inferential promiscuity. They occupy a territory that is midway between beliefs and associations (Levy 2015). They are therefore poised to play some of the roles in cognition of beliefs, but not all of them.[9]

We are now in a position to draw some conclusions about whether agents who exhibit implicit biases thereby have the kind of doxastic attitudes which make doxastic accounts of

racism plausible. Suppose such a person is an explicit egalitarian, sincerely affirming the equality of races in relevant respects (to avoid controversy, let's stipulate that the egalitarian sincerely affirms the moral, intellectual and cultural equality of all races; the stipulation can be adjusted if there are good reasons to do so, without affecting the argument). In virtue of this explicit belief, it seems appropriate to answer the question "are they racist" with a resounding "no". It would be misleading to answer this question "yes" in virtue of their implicit bias: after all, that bias is not constituted by a belief. But it would also be misleading to answer with an unqualified "no;" implicit attitudes have some of the features of beliefs, and in virtue of those features are apt to play some of the roles in cognition that make beliefs normatively significant. They guide behavior in ways that are sometimes responsive to evidence, or what is taken to be evidence; they may play a role in constituting agency. Assuming that racism is attributable to agents in virtue of their beliefs, the answer to the question "am I racist" seems to be a resounding "no" in virtue of my explicit beliefs and a highly qualified and tentative "yes" in virtue of my implicit bias.

II. The Behavioral Model of Racism.

Behavioral models are attractive and plausible for obvious reasons. Much (if not all) of the harm of racism stems from the discriminatory and disrespectful behavior it causes. Racism is normatively significant, in very important part, because it causes such behavior. There is a strong case for thinking that someone who is disposed to behave in discriminatory and disrespectful ways is a racist.

Of course, the simplest way to assess whether someone is racist, given that the behavioral model is correct, is to measure their behavior. Behavior is typically easier to measure than mental states, especially implicit mental states, since it is overt (indeed, measures of implicit mental states are usually mediated *by* behavior: response on the IAT, for instance). Our question, though, is not whether implicit bias is a good (say, useful for many purposes)

measure of whether someone is racist, but whether someone is racist in virtue of their implicit bias. Moreover, it is possible to know what someone's implicit attitudes are without knowing the extent to which they have engaged in racist behavior, (my hiring decisions might have been racially biased without my realizing it, for instance). Further, measuring implicit bias might give us a basis for assessing an agent's disposition to engage in future behavior, in a way that may well be more reliable than a measure of past behavior.

As we have seen, implicit bias causes correlative behavior. This fact is a central part of the reason why we are interested in it. It plays a role in voting behavior, hiring behavior, and behavior in many other spheres too. But does this entail that those who harbor such attitudes are racists, assuming the truth of a behavioral model?

Implicit attitudes play a proportionately greater role in driving behavior under certain conditions: when we lack the capacity, the time or the motivation to engage in controlled processing. This occurs when we are tired, stressed, under cognitive load (multitasking, for instance) or when we are required to respond too quickly for controlled processing. Methods probing implicit attitudes attempt to bypass controlled processes, by requiring rapid response (as with the IAT), by misdirection (as with the AMP, which works because the prime influences cognition in ways that are opaque to introspection) and similar manipulations. When we have time and the resources for controlled processing, and we are motivated to engage these resources (as explicit egalitarians often are), our behavior will be influenced to a much greater degree by our explicit attitudes. But we should not conclude that implicit attitudes influence our behavior very rarely. While psychologists carefully design their methods to reduce the capacity for controlled processing, there is every reason to expect that these kinds of conditions are replicated outside the laboratory. We are sometimes required to make (consequential) decisions while tired, stressed, under cognitive load or very rapidly (the decision by a police officer deciding whether to fire a gun might be under conditions which have all these features).

Even when we have time and the capacity to reflect, we may be influenced by implicit attitudes, in ways that escape our awareness. However, controlled processes remain influential in these kinds of cases. Consider, for illustration, the behavior of those voters higher in implicit prejudice but low in explicit prejudice in the 2008 presidential election. Their implicit bias influenced their behavior, but their explicit attitudes remained influential. Rather than vote for McCain, as individuals high in implicit *and* explicit prejudice did, they tended to refrain from voting altogether or to vote for a third-party candidate. Implicit bias influences behavior but in interaction with explicit attitudes which remain powerful determinants of behavior. When the agent has the time and resources to deploy controlled processing, it is only when matters are relatively finely balanced or a *prima facie* plausible rationalization of the behavior can be given that they cause behavior which is strongly at variance with the agent's explicit attitudes.

In order to answer the question whether agents like me satisfy behavioral criteria for the attribution of racism, it would be extremely useful to be able to identify the proportion of behavior that is significantly influenced by implicit bias (such that it has a character that is significantly different from the character it would have had were behavior controlled by explicit attitudes alone). Here we can turn to the literature on the predictive power of implicit attitudes. Oswald et al. 2013 report that the IAT is a 'poor' predictor of discriminatory behavior. They report a predictive validity correlation of 0.148. IAT scores account for just 2.2% of variance in behavior. A rival meta-analysis, using different inclusion and exclusion criteria, reported a correlation of .236 (Greenwald et al. 2009). While the difference between the figures generated by each meta-analysis is significant, on both explicit attitudes explain very much more of the variance in behavior and are far better predictors than implicit.

This not to conclude that implicit attitudes are trivial. As Greenwald, Banaji & Nosek (2015) emphasize, statistically small effects can produce very significant outcomes. For a large

population (say the black population of the United States; around 38 million) a statistically small effect can be expected to produce many thousands of acts of discrimination a day. Moreover, small effects can have large results in individual cases: the difference between being hired and not being hired for a particular job may be an important one (depending on the state of the economy, of course, and therefore the availability of jobs). Some effects may be more significant still: the differences in being charged with a crime or being convicted are ramifying differences: they tend to affect later stages of the life (especially since Black ex-offenders are significantly less likely to be hired on release than white offenders; Prager 2003). Very consequential behaviors are driven by implicit biases. But just as consequential behaviors are driven by explicit, and the evidence suggests that explicit attitudes are much better predictors of actions than implicit. This is unsurprising, given the finding that implicit attitudes are midway between beliefs and mere associations, and therefore apt for inference in only a limited range of cases. Given this fact, we should expect that explicit attitudes would cause and explain a greater range of behavior than implicit attitudes.

Given that implicit attitudes explain and predict a much smaller range of consequential behavior than implicit, our conclusions about whether I am a racist, insofar as behavior is central to justified ascription of the term, should be precisely the same as it was when the doxastic model provided our standard. My explicit egalitarian beliefs justify a resounding "no!" to the question. But my implicit attitudes drive some of my behavior, and in some contexts (if I do not take care to avoid them, or if I am simply unlucky), that behavior will be consequential. It would be therefore highly misleading to leave it at that. A resounding "no" in virtue of my explicit attitudes should be conjoined to a weak and tentative "yes" in virtue of my implicit bias.

III. The Affective Model of Racism.

On behavioral and doxastic models of racism, I have argued, people like me who are explicitly egalitarian but implicitly biased are weakly racist. We are only weakly racist because our implicit biases explain and predict very much less of our consequential behavior than our explicit attitudes. Since so much of our behavior is driven by controlled processes, and beliefs are the kind of attitudes which controlled processes engage, on these models we are better – more strongly – identified with our beliefs than our implicit attitudes.

Prima facie at least, matters look very different with regard to the affective model. The reason is this: while controlled processing appears to predict most of our behavior by engaging with explicit attitudes, affects seems to belong to the sphere of automatic processing. Affect cannot be directly controlled. It is automatically triggered, given the right stimulus. It is encapsulated against domain-general information. It is anticorrelated with the availability of the resources on which controlled processing relies. Thus, if being a racist is primarily or essentially a question of affect, it seems that there is a strong case for identifying a racist with her implicit biases, regardless of her explicit beliefs.

While it is true that affect is strongly resistant to direct control, however, it can certainly be indirectly controlled to some degree. We may control affect, for instance, by controlling attention, or by exposure therapy, or mental exercises. In fact, we routinely and pervasively exercise indirect control over our affects by controlling the inputs into cognition and by effortful framing of these inputs. Our affects in the domain of racial injustice illustrate the phenomenon as well as any other domain. People like me are disposed to feel characteristic and strong emotions when we perceive racial injustice (like the viral videos that played an important role in precipitating the Black Lives Matter movement). The fact that affect is automatic and mandatory given the right stimuli doesn't prevent our explicit attitudes from framing information such that it plays the role of triggering correlative affects, like anger, indignation and outrage.

With regard to behavior, there were data available which provided an estimate of the proportion of behavior predicted and explained by implicit and explicit attitudes. I know of no such data which would allow us to estimate the proportion of affects that have contents that are aligned with implicit biases and those that are instead aligned with explicit attitudes. In the absence of such data, it is very difficult to assess to what degree people like me count as racist on affective models. We can make some progress, however, by looking to the content of the affects predicted and explained by implicit biases and explicit beliefs. While there is evidence that implicit bias causes negative affect with regard to its targets (Amodio & Devine 2006), these affects seem to be relatively subtle. In fact, methods like the AMP *depend* on the relevant affects being subtle: it is only because the affects aroused by looking at racially coded faces are quite subtle that people may misattribute them to unrelated stimuli. Were the emotions strong, they would be rendered salient, and no misattribution would occur.[10]

That is not to say that the affects are trivial. On the contrary: they have a content that renders them normatively suspect. We get clues to their content by the way in which incidental emotions heighten implicit bias. Induction of incidental anger heightens implicit bias toward Arabs but not towards homosexuals, whereas induction of incidental disgust heightens implicit bias toward homosexuals but not toward Arabs (Dasgupta et al. 2009). That suggests that Arabs induce implicit anger in (American) participants, while homosexuals induce disgust. Similarly, presentation of Black faces increases the magnitude of the startle reflex in many subjects (Amodio, Harmon-Jones & Devine 2003), suggesting that Blacks induce implicit fear or anxiety in these participants, an interpretation bolstered by the finding of increased amygdala activation following subliminal presentation of Black faces (Cunningham et al. 2004). These responses are certainly normatively problematic: egalitarians should want themselves not to experience unwarranted disgust or fear, even mildly, in response to members of minority groups. But their mildness and their content gives us reason to think that these morally criticizable attitudes nevertheless fall short of the kinds of attitudes that warrant

describing those that harbor them as racist, even given the truth of an affective account of racism.

Extant affective accounts of racism are explicit in maintaining that the affects in question should have a breadth that disgust and anxiety do not have. For Schmid (1996), for instance, the racist is someone desires to dominate other races, to put them down. In his influential work, Garcia has identified the affective core with "hatred, ill-will, directed against a person or persons on account of their assigned race" (1996: 7), though he accepts that there are less central (and less vicious) forms of racism that consist in indifference toward others on account of their race. All of these affects are global states of a person: they define their core and their general attitude toward members of another race. Disgust and anxiety, in contrast, can coexist with other affective responses, at least when they are mild enough not to crowd them out. I can feel anxious in your presence *and* desire your well-being; I can feel mildly disgusted by you *and* nevertheless wish you well. It seems much harder to hate you *and* also wish you well. My affective responses are morally criticizable, it seems to me, but fall short of underwriting the accusation that I am a racist.[11]

Because they are morally criticizable, though, and they are partially constitutive of me, I don't get off the hook completely. In virtue of certain of my affective responses (my intrinsic desire for the welfare of people regardless of their race, perhaps), I can confidently proclaim that I am not a racist. But in virtue of these affective responses, I need to accept that this is far from the whole story about me. On an affective account (too), the answer to the question "am I a racist" is "perhaps, a little."

Conclusion

Doxastic, behavioral and affective accounts of racism are all *prima facie* plausible. Proponents of each (or rather, of models that emphasize one or another of these features) have

powerful arguments as to why we should regard one or another of our dispositions as central to the characterization of racism (see Atkin 2012 for an overview). In this paper, I have been agnostic with regard to these debates. On most accounts, the verdict would be the same.[12] Explicit egalitarians like me are clearly not racists in any full-blown sense: our explicit attitudes, our behaviors and our affects do not underwrite the ascription of that heavily freighted label to us. But we do not get off the hook entirely. Far from it: it would be very misleading to say that we are not racists without qualification. Rather, we have affects and mental states with a representational content which dispose us toward behavior; in virtue of these mental states and our dispositions, we might appropriately be said to be a little bit racist.

To call someone racist, or even a little bit racist, is to charge them with with something morally serious. It is an accusation we should not level without good reason.[13] In this paper, I have argued that there is good reason to describe people like me – explicit egalitarians who nevertheless harbor implicit biases – as somewhat racist (of course racism is here only an example: we may also be a little bit sexist, homophobic, and so on). Some of the sting of the accusation is taken out of it by the acknowledgment that we are also, and more strongly, *not* racist. Of course, coming to a full assessment of our character will depend on other facts about us (were active with regard to the acquisition of these attitudes? Did we acquire them at time when we could not reasonably have been expected to understand their import? Is it reasonable to expect us to attempt to change them, or to prevent their expression, and have we taken steps toward this end?) We should acknowledge the great gulf between us and between explicit racists, who are wholehearted in their racism. But we should also acknowledge our moral faults. Perhaps the gulf between me and the wholehearted egalitarian is smaller (perhaps much smaller) than that between me and the Klan member, but it is real and significant.[14]

NOTE

[1] I use the phrase "implicit attitudes" to refer to the broad class of attitudes that are opaque to introspection and escape direct control and "implicit bias" to refer to a subset of such attitudes, distinguished from the rest by their prejudicial content. Many implicit attitudes have a normatively innocuous content; some may be epistemically justified.

[2] It may be (as Carruthers 2011 suggests) that we lack introspective access to *any* of our attitudes Rather, we self-attribute attitudes on the basis of interpretive processes. In that case, the relative difficulty of accessing the content of our implicit biases would arise from a relative paucity of data, rather than a difference in kind.

[3] Kelly & Roedder devote several pages to the question "Is it morally problematic to harbor implicit racial bias?" (2008: 527ff). They do not answer the question, but instead outline the kinds of research (into the nature and content of implicit biases) they hold is needed to answer it. Faucher & Machery (2009) argue that implicit biases are responsible for what they call "racial ills", and on that basis conclude that they serve as counterexamples to Garcia's account of racism. As Garcia (2011) himself points out, however, the fact that implicit biases cause "racial ills" may not entail that those who harbor them are racists. Garcia explicitly denies that those who are implicitly biased are racist, on the grounds that these attitudes constitute "neither a cognitive nor a noncognitive stance against a race" (251). But Garcia bases this conclusion on the claim that an implicit bias "consists in mere association of concepts" (256); as we shall see, this claim is probably false. Finally, Holroyd & Kelly (forthcoming) have broached a closely related question: are implicit attitudes part of an agent's character? They argue that while we do not exercise direct control over our implicit attitudes, we do possess 'ecological control' – a kind of indirect, world-involving, control – over them, and that they count as part of our character on that basis. Holroyd & Kelly cite evidence that we can intervene to prevent their expression: for instance Dasgupta and Greenwald's (2001) finding that we can using counterstereotypical priming to counteract the effects of implicit attitudes. While it is apparent that agents possess *some* degree of control over the expression of their implicit attitudes in *some* contexts, however, I think we possess much less control than they suggest (more recent work on counterstereotypical priming, for instance, has found it to be a weak effect; Joy-Gaba & Nosek 2010). Further, interventions to alter implicit attitudes rarely have persisting effects (Lai et al. 2016). Control comes in degrees; it is very plausible that we have too little control over our implicit attitudes for them to count as part of our character on the criterion Holroyd & Kelly urge.

[4] That is, accounts of racism as a property of individuals have focused on these properties. Accounts of *institutional* racism (e.g. Haslanger 2004) focus on quite different, structural and institutional, properties, which are not located in individuals. Institutional racism is extremely significant: it might play a bigger role in explaining injustice, for instance, than the racism of

individuals. Nevertheless, it is not my focus here: I am concerned with what makes individuals racist, and accounts of individual racism plausibly highlight the features mentioned.

[5] Glasgow (2009) is an important exception (I thank a referee for drawing my attention to his paper). Glasgow argues that all instances of racism can be understood in terms of disrespect. A major motivation for his account is that there seem to be instances of racist behaviour from individuals who do not harbor racist attitudes, and from institutions that (plausibly) do not harbour attitudes at all. Since I am concerned with what makes agents racist, though, these concerns are orthogonal to mine (if racist acts do not require racist individuals, then answers to the question "what makes an act racist" may not enable us to identify racist individuals). It is plausible that if agents count as racist in virtue of racialized disrespect, this disrespect will depend on their behaviour, beliefs or affects, so probing whether agents are racist in the way I propose may enable us to answer whether agents satisfy Glasgow's condition.

[6] As a referee for this journal pointed out to me, there is another way to approach the question whether agents might be racist in virtue of their implicit attitudes. Rather than asking whether agents satisfy existing conceptions of racism in virtue of their implicit biases, we might revise our account of racism in the light of our growing knowledge of implicit bias. I think that the question I ask here needs to be answered prior to assessing the prospects for revision. First, 'racism' is first and foremost a concept that has its home in ordinary usage; plausibly, it is in light of folk usage that the special opprobrium that attaches to it is appropriate. That fact makes it important to assess the extent to which agents qualify as racist by folk lights. Second, if we discover that implicit bias is sufficiently like ordinary racism by asking the questions I probe here, there seems no need to revise the ordinary concept.

[7] Of course, it might turn out that even within a domain there is a diversity of processes and representations, such that agents who are disposed to certain racist affects may not be disposed to have others. That would greatly complicate the picture. It is worth pointing out that it would be a problem for everyone, not just me: it would seem to give us strong grounds not merely for pluralism about racism, but for identifying a diversity of racisms.

[8] In citing this experiment as evidence for the structure of implicit attitudes, I follow Mandelbaum in assuming that the same kind of attitude is involved across a number of paradigms in which behavior is driven by representational states that are automatically activated and opaque to introspection. If Holroyd and Sweetman are right that implicit attitudes are heterogeneous, *and* this heterogeneity entails differences in structure, then this assumption may need to be revisited.

[9] The evidence cites by Holroyd and Sweetman (2016), which appears to show that implicit attitudes are heterogeneous, might provide further evidence against Mandelbaum's view. On

his picture, implicit attitudes are representational states, with much the same functional profile as beliefs. If Holroyd and Sweetman are right, however, they are in fact a motely of different processes and states with quite different functional profiles (see Madva and Brownstein (forthcoming) for discussion of the extent to which Mandelbaum's model is committed to denying that implicit attitudes are heterogeneous in kind).

[10] It should be acknowledged that there is evidence that the affects are sometimes strong enough to be detectable. Ranganath, Smith & Nosek (2008) asked participants to rate their "gut reactions" and their "actual feelings" toward gay people. People reported "gut reactions" that were more negative than their "actual feelings"; moreover, their gut reactions correlated significantly with their implicit attitudes. Of course, these participants had had multiple opportunities to observe their affective responses and draw conclusions from their observations (and from their behavior as well); this evidence is therefore compatible with the claim that the affects are relatively mild.

[11] Brownstein (forthcoming) may take issue with the claim that the person should not be identified with their racist affects. Brownstein argues that implicit biases entail corresponding 'cares', where cares are tightly linked to dispositions to feel emotion, and such cares belong to the agent's deep self. As Brownstein recognizes, his position commits him to denying my claim that the deep self should be understood as a relatively integrated whole (Levy, forthcoming). Obviously, I cannot hope to settle that debate here, but I will briefly mention two reasons to think that we should not be moved by Brownstein's claims, at least in this context. First, as Brownstein emphasises, cares have a broad dispositional profile. But implicit 'cares' may lack such a profile, in virtue of their lack of integration with other states. The person who is implicitly, but not explicitly, biased against women is disposed to denigrate *and* celebrate women's achievements. Second, even if we accept that we should think of implicit cares as belonging to the agent for the purposes of attributing responsibility (Brownstein's target), it may nevertheless be appropriate to accept an integration standard for the purpose of ascribing racism to them.

[12] There are accounts which might render a different version. If a conjunctive theory like Appiah's (1990) is correct, and in addition a fine-grained account of implicit attitudes like that defended by Holroyd and Sweetman (2016) is also correct, agents might harbor problematic implicit biases without qualifying as racist at all. I might satisfy one arm of Appiah's conjunctive test in virtue of my implicit attitudes without satisfying the other, for instance. Even if a conjunctive account of racism and a fine-grained account of implicit attitudes is correct, however, we will discover whether agents like me are racists by asking the same questions and examining the same evidence I have discussed here. I thank a referee for this journal for forcing me to think through this complication.

[13] Even when it is epistemically warranted, there may be pragmatic reasons why we should nevertheless refrain from the accusation: as Saul (2012) notes, the accusation may give rise to hostility and defensiveness. It is an empirical question whether the very qualified ascription of racism to the explicit egalitarian I have suggested is warranted would also provoke hostility and defensiveness (I can attest that hedged though it is, the description invokes discomfort in me, even when I am the one applying the label to myself).

[14] I am grateful to two referee for *Philosophical Quarterly* for uncommonly helpful comments. This paper is significantly better for their objections and thoughts. I have also benefitted from discussion with audiences at Otago University, The University of Adelaide and the University of Sydney, as well as from correspondence with Michael Brownstein.

References

Amodio, D. M., Harmon-Jones, E. & Devine, P. G. 2003. Individual differences in the activation and control of affective race bias as assessed by startle eyeblink responses and self-report. *Journal of Personality and Social Psychology* 84: 738–753.

Amodio, D. M. & Devine, P. G. 2006. Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology* 91: 652-661

Appiah, K.A. 1990. Racism. In D.T. Goldberg (ed), *Anatomy of Racism*. Minneapolis: University of Minnesota Press.

Atkin, A. 2012. *Philosophy of Race*. Durham: Acumen

Bar-Anan, Y., & Nosek, B. A. 2014. A comparative investigation of seven indirect attitude measures. *Behavior Research Methods* 46: 668-88.

Barry-Jester, A.M. 2015. Attitudes Toward Racism And Inequality Are Shifting. *FiveThiryEight*, June 23.
< http://fivethirtyeight.com/datalab/attitudes-toward-racism-and-inequality-are-shifting/ >

Blum, L. 2002. *I'm Not a Racist, But...The Moral Quandary of Race.* Ithaca: Cornell University Press.

Bortolotti, L. 2009. *Delusions and Other Irrational Beliefs*, Oxford: Oxford University Press.

Brownstein, M. & Saul, J. (eds) 2016a. *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*. Oxford: Oxford University Press.

Brownstein, M. & Saul, J. (eds) 2016a. *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

Brownstein, M. Forthcoming. Attributionism and Moral Responsibility for Implicit Bias. *Review of Philosophy and Psychology*.

Carruthers, P. 2011. *The Opacity of Mind*, Oxford: Oxford University Press.

Cunningham, W., Johnson, M., Raye, C., Gatenby, J., Gore, J., & Banaji, M. 2004. Separable neural components in the processing of black and white faces. *Psychological Science* 15: 806-813.

Dasgupta, N. & Greenwald, A. G. 2001. On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology* 81: 800–14.

Dasgupta, N., DeSteno, D. A., Williams, L., & Hunsinger, M. 2009. Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice. *Emotion* 9: 585–591.

De Houwer, J. 2014. A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass* 8: 342-353.

Deutsch, R., Gawronski, B., & Strack, F. 2006. At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology* 91: 385–405.

Ditto, P.H. & Lopez D.F. 2012. Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology* 63: 568– 584.

Dovidio, J. F., & Gaertner, S. L. 2000. Aversive racism and selection decisions: 1989 and 1999. *Psychological Science* 11: 319-323.

D'Souza, D. 1995. *The End of Racism*. New York: Free Press.

Evans, Jonathan S. B. T. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59: 255–78.

Faucher, L. & Machery. E. 2009. Racism: Against Jorge Garcia's moral and psychological monism. *Philosophy of the Social Sciences* 39: 41-62.

Fazio, R. and Olson, M. 2003. Implicit measure in social cognition research: Their meaning and use. *Annual Review of Psychology* 54: 297–327.

Fredrickson, G. 2002. *Racism: A Short History*. Princeton: Princeton University Press.

Garcia, J.L.A. 1996. The heart of racism. *Journal of Social Philosophy* 27: 5-46.

Garcia, J.L.A. 2011. Racism, Psychology and Morality: Dialogue with Faucher and Machery. *Philosophy of the Social Sciences* 41: 250-68.

Glasgow, J. 2009. Racism as Disrespect. *Ethics* 120: 64-93.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. 1998. Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74: 1464-1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97: 17–41.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. 2015. Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology* 108: 553–561.

Gregg AP, Seibt B, Banaji MR. 2006. Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology* 90: 1-20.

Han, H. A., Olson, M. A., & Fazio, R. H. 2006. The influence of experimentally-created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology* 42: 259-272.

Haslanger, S. 2004. Oppressions: Racial and Otherwise. In M. Levine & T. Pataki (eds.), *Racism in Mind.* Ithaca: Cornell University Press.

Hasson, U., & S. Glucksberg. 2006. Does negation entail affirmation? The case of negated metaphors. *Journal of Pragmatics* 38: 1015–32.

Holroyd, J. & D. Kelly, forthcoming. Implicit Bias, Character, and Control. In J. Webber and A. Masala (eds.) *From Personality to Virtue*, Oxford: Oxford University Press.

Holroyd & Sweetman, 2016. The Heterogeneity of Implicit Bias. In Michael Brownstein and Jennifer Saul (eds.) *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology* (pp. 80-103). Oxford: Oxford University Press.

Huebner, B. 2016. Implicit bias, reinforcement learning, and scaffolded moral cognition. In Michael Brownstein and Jennifer Saul (eds.) *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology* (pp. 47-79). Oxford: Oxford University Press

Joy-Gaba, J. A., & Nosek, B. A. 2010. The surprisingly limited malleability of implicit racial evaluations. *Social Psychology* 4: 137-146.

Kelly, D. & Roedder, E. 2008. Racial cognition and the ethics of implicit bias *Philosophy Compass* 3: 522–540.

Lai, C. K., Skinner, A. L., Cooley, E., et al. 2016. Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General* 145: 1001-1016.

Levy, N. 2014. *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Levy, N. 2015. Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs* 49: 800-823.

Levy, N. Forthcoming. Implicit Bias and Moral Responsibility: Probing the Data. *Philosophy and Phenomenological Research*.

Lundberg, K.B., & Payne, B. K. 2014. Decisions among the Undecided: Implicit Attitudes Predict Future Voting Behavior of Undecided Voters. *PLoS ONE,* 9: doi: 10.1371/journal.pone.0085680.

Machery, E. 2016. De-Freuding Implicit Attitudes. In Michael Brownstein and Jennifer Saul (Eds.) *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology* (pp. 104-129). Oxford: Oxford University Press.

Madva, A. & Brownstein, B. Forthcoming. Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind. *Noûs*.

Mandelbaum, E. 2016. Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs* 50: 629–658.

Newman, G., Diesendruck, G., & Bloom, P. 2011. Celebrity contagion and the value of objects. *The Journal of Consumer Research* 38: 215–228.

Nier, J. A. 2005. How dissociated are implicit and explicit racial attitudes?: A bogus pipeline approach. *Group Processes & Intergroup Relations* 8: 39–52.

Oreopoulos, P., & Dechief, D. 2011. *Why do some employers prefer to interview Matthew, but not Samir? New evidence from Toronto, Montreal, and Vancouver*. Metropolis British Columbia, Working Paper Series, N°11-13.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* 105: 171–192.

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. M., Govorun, O., & Stewart, B. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology* 89: 277-293.

Payne, B. K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. 2010. Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology* 46: 367-374.

Prager, D. 2003. The Mark of A Criminal Record. *American Journal of Sociology* 108: 937-975.

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

Rozin, P., Markwith, M. & Ross, B. 1990. The sympathetic magical law of similarity, nominal realism, and neglect of negatives in response to negative labels. *Psychological Science* 1: 383-384.

Saul, J. 2013. Implicit Bias, Stereotype Threat, and Women in Philosophy. In K. Hutchison & F. Jenkins (eds). *Women in Philosophy*: *What Needs to Change?* Oxford: Oxford University Press, pp. 39-60

Schmid, W.T. 1996 The Definition of Racism. *Journal of Applied Philosophy* 13: 31-40.

Stich, S. 1978. Beliefs and subdoxastic states. *Philosophy of Science* 45: 499-518.

Wegner, D. 1984. Innuendo and damage to reputation. *Advances in Consumer Research* 11: 694-96.