Implicit Bias, Moods, and Moral Responsibility

Alex Madva
alexmadva@gmail.com

Forthcoming in *Pacific Philosophical Quarterly*

Abstract

Are individuals morally responsible for their implicit biases? One reason to think not is that implicit biases are often advertised as unconscious, 'introspectively inaccessible' attitudes. However, recent empirical evidence consistently suggests that individuals are aware of their implicit biases, although often in partial and inarticulate ways. Here I explore the implications of this evidence of partial awareness for individuals' moral responsibility. First, I argue that responsibility comes in degrees. Second, I argue that individuals' partial awareness of their implicit biases makes them (partially) morally responsible for them. I argue by analogy to a close relative of implicit bias: moods.

1. Introduction

In an influential 2002 study, John Dovidio and colleagues found that participants, who were white college students, tended to have anti-racist explicit attitudes but racially biased implicit attitudes. They explicitly disavowed racism on a questionnaire but exhibited racial bias on an indirect, computer-based measure (the Implicit Association Test; IAT). Subsequently, they engaged in an unrelated conversation with one white and one black interlocutor. During this interaction, the participants' explicit anti-racist attitudes best predicted the friendliness of what they said to the black interlocutor, while their implicit biases best predicted the *un*friendliness of their nonverbal 'microbehaviors.' They made less eye contact, blinked more often, and sat farther away from the black interlocutor than the white interlocutor. Strikingly, the white participants generally formed positive impressions of the conversations, while their black interlocutors tended to believe that the participants were consciously prejudiced against them. 'Our society is really characterized by this lack of perspective,' Dovidio says. 'Understanding both implicit and explicit attitudes helps you understand how whites and blacks could look at the same thing and not understand how the other person saw it differently' (Carpenter, 2008, 36).

Widespread and frequent repetitions of such microbehaviors constitute 'microinequities,' which stack up over time to reinforce macro-level disparities between social groups (Valian 1998). For example, Lilia Cortina and colleagues (2011) found that women, and especially women of color, tended to report experiencing more interpersonal incivility in the workplace than do men. In many cases, the incivility did not consist in overt or intentional harassment, or involve explicit reference to gender or race; rather it consisted in generic forms of rudeness, such as speaking condescendingly or interrupting a colleague. Discourteous behavior of this sort is deeply ambiguous: since just about *everybody* interrupts and gets interrupted *sometimes*, it is difficult to identify any particular instance of interruption as expressive of bias, as opposed to, say, misplaced enthusiasm.[1] But because women, and especially women of color, were evidently treated in such uncivil ways more often, it should come as no surprise that they were, as Cortina also found, more likely to intend to quit. Cortina's findings contribute to broader patterns of evidence that suggest that experiencing a work environment as hostile leads one to quit (e.g., Sims et al. 2005).

While gender- and race-based inequities depend to a great extent on structural-institutional forces, and while structural change is necessary for redressing such inequities,[2] these findings suggest that individuals' subtle expressions of implicit bias aggregate to have significant negative consequences. What responsibility, if any, do individuals bear for these subtle expressions of bias? My answer is 'some,' and in this essay I try to make some headway toward defending this position. I focus on the backward-looking question whether individuals actually are responsible and blameworthy. I advance a forward-looking case for the strategic role that *holding* individuals responsible can play in combating discrimination and inequality elsewhere.[3]

Some maintain, however, that holding individuals responsible or blameworthy for expressing implicit bias would be morally unfair and strategically counterproductive. I will call these theorists Exonerators. For example, Charles R. Lawrence, III writes:

> Understanding the cultural source of our racism obviates the need for fault, as traditionally conceived, without denying our collective responsibility for racism's eradication. We cannot be individually blamed for unconsciously harboring attitudes that are inescapable in a culture permeated with racism. And without the necessity for blame, our resistance to accepting the need and responsibility for remedy will be lessened. (1987, 325-6)

Exonerators like Lawrence argue that holding individuals responsible would be unfair insofar as implicit racial biases are unconscious and inevitable byproducts of being brought up in a systemically racist world.[4] It might be strategically unwise to boot, if it leads individuals to become hostile rather than supportive of social change.

Jules Holroyd has convincingly undercut many of the Exonerators' claims, for instance by citing evidence that acquiring implicit biases is not inevitable after all (2012, §2.1). However, Holroyd's generally persuasive defense of individual responsibility for implicit bias effectively concedes two of the most intuitively powerful grounds for exoneration. These are, first, that individuals may not be aware of their implicit biases in the relevant sense, and, second, that even when made aware, individuals may not be able to control them in the right way. Awareness strikes many as a non-negotiable necessary condition for moral responsibility. Awareness is, moreover, plausibly necessary for control, and control strikes many as another necessary condition for responsibility.[5] While Holroyd admits that there are certain qualified senses in which individuals might be aware and in control of their implicit biases, she ultimately argues that awareness and control are 'unreasonably demanding,' too onerous to be necessary conditions for moral responsibility (2012, 294; 2015, §3.1). Accepting these conditions, she argues, would generate global skepticism about moral responsibility, because individuals simply cannot be aware and in control of all the factors influencing their actions. While I am sympathetic with the view that awareness and control are not always necessary for moral responsibility (Adams 1985; Smith 2004, 2005), I take it that awareness and control are (often) at least sufficient for moral responsibility. That is, if individuals *are* aware and in control of their implicit biases, so much the better for the claim that they are responsible.

Here I argue that individuals are—at least to an important degree—aware of their biases in a sense relevant to moral responsibility and blame. I also believe that individuals are sufficiently in control of their biases, but I focus on control elsewhere (2012, forthcoming, ms). I hope to contribute to the argument that individuals are morally responsible for their implicit biases without appealing to comparatively controversial or revisionary claims about the general requirements for moral responsibility. My focus will be whether individuals are responsible for the behavioral *expressions* of their biases—for *acting* in biased ways—rather than for simply acquiring or harboring biases at all.[6] I refer to such behavioral expressions as *implicit discrimination*. I review empirical evidence regarding the kinds and degrees of awareness individuals have of their implicit discrimination (§2). I situate this research in terms of a graded notion of responsibility, i.e., the view that responsibility comes in degrees, which is both independently plausible and useful for understanding responsibility for implicit discrimination in particular (§3). To argue that individuals' degree of awareness makes them responsible for implicit discrimination, I offer an argument by analogy to a close relative of implicit biases: moods (§4). I appeal to our ordinary practices and reactive attitudes to argue that individuals are sufficiently aware of their moods so as to be responsible for mood-influenced behavior. The awareness individuals have of implicit discrimination is relevantly similar (§5).

2. Awareness of Implicit Bias & Discrimination: Empirical Evidence

Tony Greenwald and Mahzarin Banaji (1995) introduced the term *implicit attitudes* to counter the prevailing social-psychological assumption that people could, as a rule, unproblematically report the contents of their beliefs and feelings about social groups. They developed the IAT to measure implicit attitudes, which they stipulatively defined as 'introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects' (8). But *are* implicit attitudes in-principle 'unavailable to self-report and introspection' (5), i.e., permanently concealed from our conscious minds; or are they merely 'unidentified (or inaccurately identified)' in particular cases, i.e., attitudes we may be conscious of but sometimes fail to notice, or attitudes we may notice but sometimes fail to interpret accurately?

The content of implicit attitudes was originally thought to be 'introspectively inaccessible' because participants did not report them, as in studies like Dovidio's. Participants might not have been aware that such biases exist at all, that they themselves harbored them, that they were then expressing them, or that they were able to do anything about them. Lacking awareness in any of these respects might suffice to excuse their implicit discrimination.

But over a decade of evidence suggests otherwise. Researchers have found numerous ways to reduce disparities between overt and indirect measurements of attitudes. If participants are asked to 'focus on their feelings,' they are more likely to report explicit attitudes in line with their implicit attitudes (Gawronski and LeBel 2008). When participants separately report *both* their initial 'gut reactions' and their considered 'actual feelings' about social groups, then their reported gut reactions correlate strongly with indirect measures (Ranganath and colleagues 2008). That is, individuals seem more willing to report prejudiced sentiments when they can clarify that they reflectively disavow them. Participants also seem more willing to report prejudiced attitudes when told that the IAT will be an 'accurate measure of racial attitudes,' and 'the closest thing to a lie detector that social psychologists can use to determine your true beliefs about race' (Nier 2005, 43). In fact, Hahn and colleagues (2014) dispensed with these pretexts and found that participants could predict their performance on an IAT with impressive accuracy, even when given virtually no explanation of what implicit attitudes are or how they were to be measured. Moreover, participants who *were* given such explanations were no better than naïve participants at predicting their performance. 'Raising awareness'—that is, increasing participants' theoretical knowledge of the nature and measurement of implicit attitudes—did not improve their ability to predict their own implicit biases.

Perhaps most striking, Erin Cooley and colleagues (2015) found that simply telling participants whether their 'gut feelings' did or did not reflect their 'genuine' views influenced self-reported attitudes. Some participants were told that the negative gut feeling they may have had while looking at photos of same-sex couples reflected their 'genuine attitude towards homosexuality.' Those who had stronger anti-gay implicit biases were, on a subsequent questionnaire, significantly more likely to oppose gay marriage and military enrollment. By contrast, other participants were told that their gut feeling did *not* reflect their genuine attitude; in this case, implicitly biased participants were likely to *support* gay marriage and military enrollment.[7] Cumulatively, these studies are difficult to interpret without supposing that participants tend to have some introspective awareness of their implicit biases.[8] Moreover, although my focus here is the subtler manifestations of bias, this research also exemplifies how easy it is for authority figures (in this case, scientists, but presumably also professors, parents, media personalities, and religious and political leaders) to *tap into and legitimize* these negative gut feelings, and effectively transform implicit bias into overtly endorsed discrimination. In fact, I argue (ms) that the recent resurgence of explicit bigotry and misogyny in North America and Europe—and the apparent collapse of the consensus that it is intolerable for political leaders to give full voice to prejudice—depend *in part* on the fact that implicit biases have pervasively persisted as introspectively accessible mental states, which are interpreted and acted-upon in various ways depending on structural-contextual factors such as power relations and perceived social norms. Note, for example, that Cooley's 'genuine attitude' manipulation only elicited prejudiced self-reports from individuals who displayed strong implicit biases. Those who lacked such negative gut feelings in the first place were simply impervious to this rhetorical maneuver.

Evidence for introspective awareness of these social gut feelings is not new. In 1991, before the term ‘implicit attitude’ was coined, Devine and colleagues found evidence for robust self-awareness of tacit prejudice among individuals who espoused anti-prejudiced ideals. Participants reported how they *would* versus *should* respond, according to their ‘own personal standards,’ in a variety of different situations. For example, they rated their agreement or disagreement with statements such as, ‘Imagine that a Black person boarded a bus and sat next to you. You *should* [*would*] feel uncomfortable that a Black is sitting next to you’ (1991, 819). Other situations included ‘feeling uncomfortable that a job interviewer is Black,’ and ‘seeing a Black woman with several small children and thinking “How typical.”’ Over 70% reported significant discrepancies between how they would and should respond in these situations. They tended to admit that they *would* feel uncomfortable sitting next to a black person on the bus, but also to insist that, by their own lights, they *should not* feel that way. Most who reported such discrepancies also reported feeling guilty or disappointed with themselves.[9] Such findings suggest not just mere awareness of the existence and content of biased gut reactions, but a relatively rich knowledge of both their moral significance and their effects on a range of thoughts, feelings, and actions.

Examples from film, literature, and journalism arguably provide further evidence for widespread awareness of implicit bias. In the film *Gentleman’s Agreement,* for example, a reporter named Phil goes undercover to study anti-Semitism, pretending to be Jewish for six weeks. In effect, Phil encounters repeated implicit discrimination. When he explains to Kathy, his romantic interest (played by Dorothy McGuire), that, ‘I’m going to let everybody know that

I'm Jewish, that's all,' she responds by saying, 'Jewish? But you're not, are you?' The scene had to be repeatedly re-shot because McGuire's original look of dismay was too overt; it had to be more subtle.[10] Kathy's tacit anti-Semitic attitudes are particularly salient in the film because *she* first came up with the idea for an exposé of anti-Semitism. Her commitment to root out concealed prejudice sets the film's plot in motion, making hers a case of sincere egalitarianism coupled with implicit anti-Semitism. In many ways, the film is structured around the transformation she undergoes in coming to face her own explicitly disavowed biases. The director bluntly described the film's message this way: 'You are an average American and you are anti-Semitic. Anti-Semitism is in you.'[11]

Gentleman's Agreement came out in 1947, seven years before social psychologist Gordon Allport's seminal work, *The Nature of Prejudice*. The film won the Academy Awards for Best Picture, Best Director, and Best Supporting Actress. *The New York Times* wrote that, 'To millions of people throughout the country, it should bring an ugly and disturbing issue to light.'[12] Although it was received by some as a revelation, it is implausible that *Gentleman's Agreement* could have achieved such notoriety if the phenomena it examined had been completely alien to the lived experience of American moviegoers (a common criticism of the film, then and now, is that the journalist is so *surprised* by what he learns about tacit anti-Semitism). *Gentleman's Agreement* was fictional, but it inspired white journalists Ray Sprigle (1949) and John Howard Griffin (1961) to artificially darken their skin, pretend to be black, and write memoirs of their experiences of discrimination. Both memoirs were widely publicized. Sprigle won the Pulitzer Prize. Griffin won the Davenport Catholic Interracial Council's *Pacem in Terris* Peace and Freedom Award, and his memoir was adapted into a 1964 film of the same name.

More recently, *Avenue Q*, winner of the 2004 Tony Award for Best Musical, includes a song entitled, 'Everybody's a Little Bit Racist:'

> Everyone's a little bit racist/ Sometimes./ Doesn't mean we go/ Around committing hate crimes./ Look around and you will find/ No one's really color blind./ Maybe it's a fact We all should face/ Everyone makes judgments/ Based on race…
> If we all could just admit/ That we are racist a little bit,/ Even though we all know/ That it's wrong,/ Maybe it would help us/ Get along.

With these works in mind, I take contemporary empirical research on implicit bias to be contributing to how we understand phenomena that have been part of our collective awareness for quite some time, and *ipso facto* part of the awareness of a great many individuals. Moreover, before the recent explosion of research on implicit attitudes, quite a few feminists, race theorists, and activists sought to draw attention to related issues—although their efforts were often less publicized and laureled than the writings of white men like Sprigle and Griffin. Exactly how our collective awareness of tacit prejudice and discrimination has progressed (or regressed, or been repeatedly repressed) over time is an important question in its own right (see, e.g., Mills, 1997), but not one I will explore further here.

So, for example, Washington and Kelly (2016, 23, original emphasis) are wrong to say, 'In 1980, *no one* knew the unsettling psychological facts about implicit biases; the psychological research had not yet been done, and so today's wealth of empirical evidence simply did not exist.' Many people knew, and the knowledge was (occasionally) broadcast widely.[13] Certainly the quality and depth of our knowledge has changed as more research has been done, and I grant that these developments make a difference to our moral responsibility. But, like so much else in this domain (§3), these are differences of degree. Instead of asserting that individuals were categorically *not* blameworthy in 1980 but *are* blameworthy now, a more natural response is that individuals were simply *less* blameworthy then than now, or so I'll argue. Conversely, to argue that these differences are of kind rather than degree, some account of what the thresholds are and when our community crossed them seems required (cf. Kelly and Washington 2016, 28-32).

Of course, then and now, many people have lacked sophisticated theoretical knowledge about implicit bias. But most also lack scientific knowledge about their moods, beliefs, and all other folk-psychological states. Is there a categorical and normatively relevant *difference* between implicit biases and other mental states on this score? If we assume (as I do) that ordinary, theoretically naïve agents are sometimes morally responsible, e.g., that self-avowed racists are responsible and blameworthy for overt discriminatory behavior, then it's frankly unclear why widespread theoretical knowledge of implicit bias should be relevant to individual attributions of responsibility and blame.[14]

But how do we square all this evidence for awareness with findings like Dovidio's, in which white participants seemed completely oblivious of their visible discomfort and unfriendly behaviors toward blacks? Their apparent lack of awareness is especially striking in light of Devine and colleagues' (1991) finding that participants tended to predict that they would feel inappropriately uncomfortable in interracial interactions much like this one. Nevertheless, I don't think we have to look very hard for a variety of plausible ways of fitting all the evidence together.[15] For example, perhaps Dovidio's participants would have been more willing to admit discomfort and admit to making less eye contact if they had been queried about hypothetical cases, rather than queried about the awkward interaction they had just fumbled through. Perhaps they would have admitted discomfort if given the opportunity to clarify that their gut feelings don't reflect their considered commitments or personal standards (or if an admired political leader had encouraged them to believe their discomfort was actually justified). Perhaps the stress of the interaction, or their anxiety about appearing racist, or simple wishful thinking prevented them from noticing or reporting their negative gut feelings and microbehaviors. In fact, while studies like Dovidio's find direct correlations between implicit bias and negative microbehaviors (e.g., McConnell and Leibold, 2001; Jacoby-Senghor et al., 2016), other studies find that implicitly biased participants can sometimes 'hold it together' and act friendly during intergroup interactions, but then show signs of cognitive fatigue afterward (Richeson and Shelton, 2007; Gonsalkorale et al., 2009). In these cases, participants seem to be actively suppressing their negative impulses, at least temporarily. Needless to say, it would be difficult to suppress an impulse of which one was completely unaware.

While many empirical questions remain unanswered, it seems clear that we cannot cast implicit biases into what popular authors such as Gladwell (2005) call 'the locked door of the unconscious.' Individuals are, or can be, introspectively aware of their biased gut reactions, and even aware of discrepancies between how they would and should (according to their considered judgments) respond in various situations. The working hypothesis should be that the affect-laden contents of implicit biases contribute, or are available, to conscious experience, although in many instances without being the object of explicit attention. The evidence is best understood in light of a familiar distinction between the content of one's phenomenology (i.e., that which is experienced) and the content of one's focal attention. That is, they may be *felt* without being *noticed*, just as a person can be in a grumpy or lighthearted mood without noticing as much. Different theories of consciousness will characterize (roughly) this distinction differently. In Block's (1995) terms, implicit attitudes are *phenomenally conscious*, if not always *access conscious*. According to a higher-order or attention-based theory of consciousness, implicit attitudes would typically be *potentially conscious*, i.e., *accessible* even if not often *accessed* (see also Madva 2016a).

What are the implications of this often partial and inarticulate awareness of implicit bias for moral responsibility? Are individuals who act on the basis of such biased gut feelings responsible and blameworthy? Among the 'folk,' there seems to be traction to the intuition that such individuals would be responsible. In Cameron, Payne, and Knobe (2010), participants assessed the responsibility and blameworthiness of hypothetical employers, renters, and exam graders influenced by implicit racial bias. Some participants read about an individual who 'thinks people should be treated equally, regardless of race' but 'has a sub-conscious dislike for African Americans' (2010, 276). He tries to promote people on merit alone, but 'because he is unaware of this sub-conscious dislike,' he 'sometimes unfairly denies African Americans promotions.' Other participants read about an individual who 'has a gut feeling of dislike toward African Americans'—which he is aware of, and sincerely rejects, but has 'difficulty controlling.' Participants tended to judge that the individual who was aware of his dislike, but struggled to control it, was significantly more responsible and blameworthy than the one who was unaware of it altogether.

I share the intuition that being partly aware of one's implicit biases increases the degree to which one is responsible for them. In what follows, I defend this intuition by drawing an analogy between implicit biases and moods. First I urge that awareness, control, responsibility, and blameworthiness all come in degrees (§3). I then argue that the degree of awareness individuals have of their moods suffices to make them morally responsible for mood-influenced behavior (§4), and that individuals' awareness of implicit bias is importantly similar (§5). I defend many of these claims by appeal to our ordinary practices and reactive attitudes. I recognize, however, that Exonerators tend not to share the intuition that individuals are responsible for implicit discrimination, and they might further argue that appeals to 'folk' intuition and ordinary practices are unreliable guides for moral reflection on implicit bias (see Levy 2017, 4-5). I will not offer a wholesale defense of the methodological value of appeals to intuition and practices (I certainly agree that intuitions and practices can be wrong). Instead, I'll offer a more concessive response on these issues: what follows can, to a significant extent, be construed as a partial explanation of the *sources* of the common intuition that individuals are responsible for implicit discrimination. That is, it represents an attempt to situate this intuition in a more familiar web of moral practices and attitudes.[16] Better understanding the intuition's sources is useful even if we ultimately decide that, for whatever reason, the intuition is mistaken. In fact, it could be useful for the purposes of *figuring out why* it's mistaken. To that extent, then, I think Exonerators can accept much of what follows.

3. Responsibility in Degrees

Awareness strikes many as a necessary condition for responsibility. However, research on implicit bias (and on consciousness more generally) suggests that awareness is not an all-or-nothing affair. It is possible for individuals to be partly but not perfectly aware of their mental states, for example, when a headache is felt but initially unnoticed, or when a stressful feeling of hunger is noticed but misinterpreted as anger at another person (Bushman et al. 2014). Awareness comes in degrees. In fact, it is increasingly clear that all the core notions in this family of concepts—awareness, control, responsibility, and blameworthiness—are graded (Björnsson and Persson 2013, Buchak 2014, Coates and Swenson 2012, Raz 2010, Sinnott-Armstrong 2013).

Although my primary focus in this essay is awareness, briefly consider control. Exonerators point out that even if individuals are aware of their implicit biases, they may not be able to control them properly. Sometimes trying to control implicit discrimination even exacerbates its harms (Norton et al. 2006). At issue here is what I call *local control*, the ability to directly control implicit discrimination in particular instances, which differs from *indirect control*, which involves taking steps in advance to block discrimination (e.g., anonymous reviewing) and *long-term control*, which involves debiasing one's social habits through repeated practice (Madva 2012, 2016a,b, forthcoming; see also Holroyd 2012, §2.2; Levy 2017). A central consideration is whether implicit discrimination is altogether *uncontrollable* or merely *difficult to control*. The evidence suggests the latter: controlling implicit discrimination can indeed be difficult—taxing, demanding, depleting—but not impossible. Implicit discrimination truly exceeds individual control only in rarefied circumstances, such as during timed computerized tasks cleverly designed by psychologists for the express purpose of circumventing control. These tasks become trivially easy when participants can take all the time they need. Normatively speaking, there is a world of difference between being difficult and impossible to control. Controlling the expression of our bladders can sometimes be difficult, but, for healthy adults, conditions have to be extreme before control is altogether lost, and we readily incorporate such considerations (about individuals' health and specific circumstances, e.g., if a person is unwell, pregnant, or just overcome by intense laughter) into our normative assessments of breakdowns of bladder control.

Control is thus a matter of degree. Theorists have widely acknowledged the significance of this fact for assessing the responsibility of individuals in compromised circumstances, such as addicts (Holton and Berridge, 2013; Sinnott-Armstrong, 2013), children (Mele, 2006), and negligent agents (Raz, 2010). Such individuals are *somewhat* in control and therefore *somewhat* responsible. This seems a point of growing consensus especially among philosophers working at the intersection of normative theory and empirical and clinical research in psychology, neuroscience, and experimental philosophy. I agree with such analyses, and take implicit discrimination to be relevantly similar, falling somewhere between full control and the total lack thereof. Borrowing a term from Schwitzgebel's (2010) work on belief, we might say that phenomena like implicit bias and addiction are 'in-between' controllable—and infer that individuals are, therefore, in-between responsible for controlling them.

Exonerators typically opt for a different strategy, seeking out morally significant properties of implicit biases distinct from responsibility and blame. For example, Zheng (2016) argues that implicit biases are not always 'attributable' to agents' real selves (cf. Brownstein 2015; Levy 2011, 2017), but that individuals may nevertheless be 'accountable' for them.[17] A central motivation for drawing such distinctions between more and less 'stinging' forms of moral criticism is the pragmatic, forward-looking concern that laying blame will be counterproductive, which I address elsewhere (2012, 2016b, forthcoming, ms). But distinguishing attributability from accountability is compatible with a graded conception of responsibility; in fact, I think attributability and accountability are also graded.

An alternative view, defended perhaps by Fischer and Ravizza (1998), is that responsibility is binary (i.e., all-or-nothing: there exist necessary and sufficient conditions for responsibility that are simply satisfied or not), but that blameworthiness is a matter of degree. But it is mysterious why we should continue to believe that responsibility is binary if awareness, control, and blame (and other related phenomena, such as reason-responsiveness, attributability, duress, criminal and civil punishments, etc.) all come in degrees. It is difficult to imagine nonarbitrary ways of setting thresholds for the minimal degrees of awareness and control that would be necessary to pass from unalloyed innocence to unequivocal responsibility.[18]

Given that a graded notion of responsibility is independently plausible, why not apply it to the case of implicit bias? The reasoning here is straightforward: individuals are somewhat aware and somewhat in control of their implicit discrimination, and so they are somewhat responsible, and somewhat to blame. They are more responsible than they would be for purely unconscious, reflexive or pathological behavior, but they are less responsible than they would be for explicit discriminatory behavior. When individuals' awareness of their implicit biases is not fully comprehensive or articulate—as when an implicit bias is felt but not noticed, or noticed but misinterpreted, or when it is interpreted correctly but its causal influence on judgment or action is underestimated—these are mitigating factors. They make the individual in question less responsible but not completely off the hook. Determining the precise extent of an individual's responsibility and blameworthiness for a specific action or omission is a complex, context-sensitive affair, analogous in some respects to the nuances and challenges of determining criminal or civil liability. I will say more about what goes into this 'responsibility calculus' in §§4-5.

Next I argue that our ordinary practices already accommodate relevantly similar in-between cases of moral responsibility. Moods represent another type of mental state that occupies this moral middle ground.

4. Awareness, Reactive Attitudes, and Moods

Some philosophers claim that the awareness relevant to responsibility is a kind of reflective one, on display when an individual 'steps back' and considers, e.g., whether it would be appropriate to act on a particular desire (e.g., Korsgaard 1997). This reflective awareness clearly requires that the relevant desire is explicitly noticed, accessed, and attended-to. I take it that this sort of reflective awareness seems intimately tied to responsibility because the ability to step back and reflect is thought to be the best candidate for enabling an individual to *control* her automatic impulses. Intuitively, it is when an individual can take a moment to deliberate that she is best poised to resist her immediate inclinations.

The capacity for this sort of reflective awareness may be necessary for an individual to be a possible candidate for responsibility and blame in the first place. It is a plausible background condition essential to being the sort of entity to whom responsibility and blame could ever appropriately be assigned. But it is less plausible that, in particular cases, an individual has to be capable of stepping back and reflecting in order to be responsible for what she does. In our ordinary practices, we often take less robust forms of awareness to be sufficient for, or at least directly relevant to, responsibility and blame.

Consider how we commonsensically understand the effects of moods on behavior, and how this understanding in turn figures in our attributions of responsibility.[19] Suppose Gertie is in a grumpy mood. She might have no knowledge of the causal source of her mood; it could be due to stress, hunger, a headache, air pollution (Rotton et al., 1978), hitting one red light too many during the morning commute, or simply 'waking up on the wrong side of the bed.' Her mood might have all sorts of unknown effects on what she thinks and does. Still, it would be hasty to conclude that the mood itself (or its content, whatever it may be) is in any deep respect unconscious. Gertie may not even notice that she is in a grumpy mood, but she is in one just the same, and feeling it all the while. Perhaps her mood passes in and out of focal awareness, or perhaps it just hovers in the periphery.[20] These are empirical, if notoriously difficult to tackle, questions.

But the fact that she fails to notice her grumpy mood would not simply exonerate any rude behavior this grumpiness might cause—or so we intuitively think.[21] We routinely hold others and ourselves responsible for the things said and done because of bad moods. If Gertie's mood leads her to roll her eyes, adopt a cantankerous tone, or interrupt a friend or colleague (or even a stranger), the latter would intuitively be warranted in holding her responsible and blameworthy for doing so. If Mordy is in a good mood because he has just received exciting news, he might fail to react with sufficient empathy and concern when he discovers Gertie has received bad news, and Gertie might reasonably hold it against him.

It's true that being influenced by a mood can *make a difference* to responsibility and blame. Citing a bad mood as a (partial) explanation for inappropriate behavior can make the behavior appear less objectionable, somehow mitigating the severity of the offence (perhaps the behavior comes to seem less intentional or 'personal'). It could be that citing a bad mood leads us to judge that the behavior is less blameworthy, or it could be that citing a bad mood leads us to shift blame from the behavior itself to the failure to restrain the behavior. In the latter case, individuals might just be responsible for letting the mood get the best of them.[22]

Some nevertheless identify mood-related misbehavior as blameless. Levy (2011, 245) mentions in passing a case in which, 'George's shortness with his colleagues might be excused because of the stress he has been under recently,' insofar as George is not properly aware of the reasons for his acting that way.[23] I agree that stress can play a mitigating role, but Levy seems to overstate its exonerating force. If we seriously entertain being one of George's colleagues, would we ordinarily take his stress to *fully* exculpate his shortness? Would it transport him entirely from the realm of responsibility and blame, and lead us to switch completely from the ‘participant’ to the ‘objective’ stance described by Strawson? I think not (although it might depend to some extent on the details of the case, for example, on just how trying or traumatic the source of stress is). A graded conception of responsibility more naturally accommodates this sort of case. Learning that George has been under stress may help to make his shortness more intelligible to his colleagues, without thereby giving him free license to be uncivil. George’s being under stress might mitigate the severity of his offence, making him *less* responsible and blameworthy, without becoming completely off the hook for his rudeness. In short, moods mitigate, not exonerate.

The mitigating status of moods can be illuminated by considering how we offer and accept apologies for mood-influenced behavior. When George snaps at his colleague, and subsequently apologizes, he might say, ‘I’m sorry for being irritable. I’ve just been under a lot of pressure lately,’ or, ‘I just woke up on the wrong side of the bed today.’ How would his colleague respond in this case? Would the colleague say, ‘Come now, you have *nothing* to apologize for. You didn’t do anything wrong.’ More likely, the colleague would say something like, ‘It’s okay. Don’t worry about it. I know things have been stressful for you.’ The apology is not out of place here, I argue, because citing a bad mood does not completely absolve one of responsibility or blame. It often has the effect of putting the person on the receiving end of the rudeness in a position to *accept* the apology, or acknowledge it some way, rather than deny the need for it altogether. This sort of mitigating factor thus does not transport an individual out of the realm of responsibility and blame, but shifts her location within that space.

The mitigating role of moods is, of course, subject to a number of complicating factors. For one thing, it makes a difference what sort of behavior is supposed to be illuminated by reference to the bad mood. Being influenced by a mood can affect an individual in many ways beyond unfriendly microbehaviors, perhaps by making her a harsher grader or a less sympathetic interviewer. What if George's stress leads not just to shortness but to verbally abusive screams, or significant property damage, or violence? If a mood leads an individual to punch someone in the face, or to deny a parole application (Danziger et al., 2011), then citing it might do considerably less exculpatory work. Other things equal, the more serious the consequences of the behavior, the less mitigating we'll take a bad mood to be.[24] If we are ever justified in adjusting our attributions of responsibility and blame in light of the severity of consequences, then part of the justification might lie in our (more or less explicit) knowledge that people are often better able to control themselves when the stakes are raised. For example, someone in a bad mood might be much more able, or at least more likely, to restrain his rude impulses in the presence of an armed mugger than in the presence of a close friend, or—to take an example more pertinent to implicit bias—in the presence of his bosses than in the presence of subordinates. Suppose these external factors do affect how easy it is to control the influence of moods on behavior. The upshot is not that individuals ought to be exonerated for their mood-related misbehavior *when the stakes are low*, or when it is difficult to control, but that they are, at least to some degree, responsible for such behavior regardless of the presence or absence of these mitigating factors.

I cannot here capture all the nuances of our intuitions and practices surrounding mood-related misbehavior. Nevertheless, it seems that, in paradigmatic cases, when an agent is in a bad mood, and as a result acts in an unfriendly way, she is to a certain degree (held) responsible and blameworthy, even if she never introspectively noticed being in that mood. Being influenced by this unnoticed psychological state does not transport her out of the realm of responsibility for her actions.

There are for my purposes two sorts of pernicious effects that moods can have on behavior.[25] The first is an expressive harm: foul moods can lead individuals to express unwarranted negative affect toward others. Such expressive harms, and the moral importance of affective phenomenal experiences more generally, were a centerpiece of Strawson's (1962) influential essay. The reactive attitudes he cited as integral to our understanding of responsibility were not the cold, cognitive evaluations judges and jurors are asked to make in determining the scope of a defendant's criminal responsibility. Rather, they were affective and unreflective. We care about whether people bear us good or ill will, whether they appreciate or resent us, smile or frown. Expressions of good or ill will are as much, or sometimes even more so, a matter of our immediate tendencies to react to others as they are a matter of our reflective judgments about those reactions. We care about how others *actually feel* about us, and this is true to some extent independently from whether they would reflectively endorse (acting upon) those feelings. We take the automatic, affect-expressive behaviors of others to reflect their attitudes toward us, and to license certain affect-laden responses from us in turn, such as when I feel resentful toward you for being short with me. These affective reactions are first and foremost part of phenomenal awareness: qualitatively felt even if not explicitly noticed. Individuals *can* become reflectively aware of them, but they need not in order for those reactions to constitute tacit forms of approval or disapproval, and to be potential candidates of praise or blame. In this vein, mood-related rudeness is (ordinarily treated as) a *prima facie* expression of ill will that needs to be accounted for somehow. People deserve to be treated civilly, and feeling grumpy is not a license to shirk common courtesy.

Apart from this expressive harm, a second significant effect of moods is that they bias cognition, leading individuals to selectively attend to, ignore, misinterpret, and even misperceive features of their environment (Sizer 2000). A mood might lead individuals to misperceive a sincere smile as a smug smirk, or to dwell on the shortcomings of a résumé and overlook its merits. For example, Forgas (2011) manipulated participants' moods before they read a philosophy essay either written by 'a middle-aged bearded man in a suit with spectacles' or 'a young woman with frizzy hair wearing a t-shirt.' Participants in a good mood relied on their gut feelings, and evaluated the older man's essay, competence, and likeability much more highly than the young woman's. However, being in a bad mood induced more vigilant, attentive thinking, and reduced this age/gender bias to statistical insignificance. In such cases, the mood itself may be entirely unnoticed, but it leads to distortions of what we do notice.[26]

Suppose a tired or angry individual drives past a stop sign without slowing down. In many such cases, it is plausible that the driver saw the stop sign (one might say that 'it passed through her field of vision'), but did not explicitly notice it because of the bad mood. In these cases, the relevant feature of the situation is right there in front of us, and, insofar as we are generally competent at reading résumés and faces, or noticing traffic signals, we would (ceteris paribus, take ourselves to) be culpable if we later discovered that we had failed to respond to it properly.[27] The unreflective nature of the error might be a mitigating factor (again, the wrong might come to seem less personal), but we'd nevertheless have something to make good on, or apologize for. Part of our willingness to hold individuals responsible in these cases, I submit, is that the individuals in question genuinely perceive (at least to some degree) the relevant feature of the situation, although they fail to respond to it appropriately.

Again, different theories of consciousness and perception will characterize these phenomena in different terms. One might deny, for example, that the individual genuinely, consciously perceives the stop sign, and say that the individual is simply poised to perceive it. Even so, there is still room to say the individual *should have* been aware, should have accessed the information, etc. (and, of course, being poised to perceive will also be matter of degree). On this line, these cases would approximate a certain kind of negligence, when an agent detects something tacitly, and should respond to it (perhaps by becoming access-conscious or focally aware of it), but fails to. In such cases, obliviousness could be blameworthy. I will not here adjudicate between these different interpretations because, either way, these factors mitigate blameworthiness without exonerating one entirely. In these cases, it is either our tacit awareness or our potential awareness of the feature that makes us responsible for responding to it, while our mood distorts our capacity to notice, interpret, or respond to it appropriately. Tacit awareness puts us on the hook, but the mood mitigates the offence.

5. Awareness, Responsibility, and Implicit Bias

Now consider two cases that involve not (merely) bad moods but implicit biases. First, take George Yancy’s rich phenomenological analysis of stepping into an elevator:

> Well-dressed, I enter an elevator where a white woman waits to reach her floor. She ‘sees’ my Black body, though not the same one I have seen reflected back to me from the mirror on any number of occasions. Buying into the myth that one’s dress says something about the person, one might think that the markers of my dress (suit and tie) should ease her tension. What is it that makes the markers of

> my dress inoperative? She sees a Black male body ‘supersaturated with meaning, as they [Black bodies] have been relentlessly subjected to [negative] characterization by newspapers, newscasters, popular film, television programming, public officials, policy pundits and other agents of representation’. Her body language signifies, ‘Look, *the* Black!’ On this score, though short of a performative locution, her body language functions as an insult. Over and above how my body is clothed, she ‘sees’ a criminal, she sees me as a threat. Independently of any threatening action on my part, my Black body, my existence in Black, poses a threat.
>
> There is not anything as such that a Black body needs to do in order to be found blameworthy. As such, the woman on the elevator does not really see me, and she makes no effort to challenge how she sees me… she may come to judge her perception of the Black body as epistemologically false, but her racism may still have a hold on her lived body. I walk into the elevator and she feels apprehension. Her body shifts nervously and her heart beats more quickly as she clutches her purse more closely to her. She feels anxiety in the pit of her stomach. Her perception of time in the elevator may feel like an eternity… The point here is that deep-seated racist emotive responses may form part of the white bodily repertoire, which has become calcified through quotidian modes of bodily transaction in a racial and racist world... Despite how my harmless actions might be constructed within her white racialized framework of seeing the world, I remain capable of resisting the white gaze’s entry into my own self-vision. I am angered. Indeed, I find her gaze disconcerting and despicable. (2008, 846-7)

Drawing on Du Bois, Fanon, and Gooding-Williams, Yancy mines this scenario to make several points. I will highlight just a few. (1) Yancy is wearing a suit and tie. It is unlikely that a white man so adorned would be perceived as a similar threat,[28] but the woman’s racial attitudes bias her cognition, and prevent her from noticing or being moved by these standard markers of ‘respectability.’ (2) Yancy takes for granted that the woman’s distorted perception reflects the fact that she has been bombarded with stigmatizing representations of black men in ‘mass media.’ That is, it’s built into the case that her implicit discrimination (a range of automatic perceptual, cognitive, affective, and bodily responses) is a product of her immersion in a systemically biased world. (3) The woman might reflectively disavow her implicit discrimination, i.e., harbor sincere egalitarian commitments and ‘judge her perception of the Black body’ to be false. (4) But her implicit discrimination nevertheless constitutes, or is at least experienced as, an insult to Yancy—a tacit act of blame—which is despicable and elicits justified anger. Although Yancy is acutely aware of factors like (2) and (3), which plausibly mitigate her responsibility and blameworthiness, his moral resentment persists.

It could be that Yancy would feel less resentment and benefit psychologically if he actively concentrated on the mitigating factors, e.g., by reminding himself that the woman's biases are simply byproducts of an upbringing in an unjust social reality, but the presence of these factors does not obviously make her blameless. She might not be as blameworthy as she would be if she reflectively endorsed her implicit discrimination, but she is more blameworthy than she would be for a mere behavioral reflex, like blinking in response to a bright light. In fact, I think that if we seriously imagine ourselves in Yancy's shoes, it is quite difficult to insist that the woman is completely unconscious of these reactions, or completely free of blame. The suggestion that her affective-bodily responses are on a moral par with behavioral reflexes, or that it is Yancy's responsibility to exercise cognitive-therapeutic techniques to reduce his stress (rather than the woman's responsibility to not act that way) strike me as quite problematic. While we need not conclude of her, or each other more generally, that we are all bad prejudiced people, it is fair to conclude that she could be, in an important sense, better than she is, and that we could be better than we are.

The potential for implicit biases to influence what we notice and how we respond is also evident in the following interaction described by Virginia Valian:

> A storm has damaged a large tree in the back yard, and a tree surgeon has come to look at it… As I ask the tree expert various questions about the damage and what needs to be done, I feel there is something a little odd about his responses. Finally, I realize that I am looking at him when I ask my questions, but that he is looking at J when he answers them. For his part, J is mostly looking abstractly out into space, reflecting his lack of interest in the proceedings. For the entire consultation, in fact, J is silent. I continue with my questions, and the surgeon continues to direct his answers to J. Perhaps he is riveted by J's virtuosic

> ventriloquism. I got the information I wanted, but I don't know what modifications I might have made—speaking louder? asking longer questions? being more assertive?—to get the tree surgeon to talk to me instead of J. I can imagine the surgeon saying to his crew afterward, 'Did you see that woman? She didn't let that guy get a word in edgewise.' J himself has noticed nothing, because he has been thinking about something else the whole time. (1998, 146)

In this case, my intuition is that the tree surgeon is somewhat responsible and blameworthy for failing to attend to the relevant social cues, and for acting in an oblivious, uncivil way. Unless he suffers from a visual impairment, extreme social anxiety, etc., he knows who is asking the questions, and how to answer a person who asks a question. His social environment is giving him ample information to suggest that he should adjust his unreflective behavior, but he fails to absorb it. Valian would be warranted in resenting *him* for this behavior, rather than just, say, resenting American culture more broadly for leading the tree surgeon to develop these habits of selective attention (although she could reasonably resent American culture, too). In this case, the tree surgeon's bias may not lead him to express any sort of negative affect, as did the woman on the elevator, but it does lead him to discount or ignore what's right in front of him, and culpably so. Part of the explanation for this, I believe, is the implausibility of supposing that the tree surgeon is *completely unaware* of what he's doing. Suppose Valian had said something explicitly about his failure to make eye contact. He might have reacted with defensive hostility or denied that he meant any ill will, but would he really have had no clue what she was even *referring to*?

It is not just that we can trace things back to some prior moment in which the individual should have reflected upon things and decided to form better social habits; there is a kind of awareness operative *at the time*, as the conversation is unfolding, which puts the individual in contact with the relevant feature of the situation, and on the hook for acting appropriately. To spin this point in more forward-looking terms, one reason to make a lot of hay out of the sort of first-personal awareness that individuals seem to have of their implicit biases is that this awareness *presents an opportunity for intervention*. Implicit biases aren't just coloring our thoughts, perceptions, and actions from behind the locked door of the unconscious, but are themselves palpably present (or at least accessible) to awareness. This first-personal, in-the-moment awareness of our biased thoughts, feelings, and actions opens up a distinctive set of opportunities for us to do better.

6. Conclusion

Of course, it may or may not be productive for Valian or Yancy to say something accusatory in such situations. In this vein, Exonerators argue that holding individuals responsible for implicit discrimination would be not just unfair but also strategically ill-advised. I take up related forward-looking concerns elsewhere (2012, 2016a,b, forthcoming, ms). While I agree that we should not necessarily saddle individuals with '-ist' labels that portray them as horrible people for possessing and expressing implicit biases, it is a mistake to conflate sanctimonious name-calling with the view that implicit discrimination is often worthy of blame, broadly construed. Blame is not so blunt an instrument. We can acknowledge the failings of others and ourselves to live up to our commitments without calling the sincerity of those commitments into question. In many cases, we can insist that individuals bear a legitimate degree of responsibility and blame, even if they lack perfect awareness of what they do. If it is ever strategically unwise to lay blame, then the upshot is not to jettison implicit bias from the sphere of moral responsibility; the upshot is to take great care in locating it properly within that sphere. Needless to say, I am not suggesting that the most effective way to combat systemic discrimination and oppression is simply to stamp out individuals' biased microbehaviors. We should combat systemic ills by changing the system. In addition to thinking about how the system needs to change, we cannot forget who needs to change it.[29]

Department of Philosophy
California State Polytechnic University, Pomona

References

Adams, Robert M. (1985). 'Involuntary sins.' *The Philosophical Review*, *94*(1), 3-31.

Anderson, Elizabeth (2010). *The Imperative of Integration*. Princeton: Princeton University Press.

Beedie, Christopher, Terry, Peter, and Lane, Andrew. (2005). 'Distinctions between emotion and mood.' *Cognition & Emotion*, *19*(6), 847-878.

Björnsson, Gunnar, and Persson, Karl. (2013). 'A unified empirical account of responsibility judgments.' *Philosophy and Phenomenological Research*, *87*(3), 611-639.

Block, Ned. (1995). 'On a confusion about the function of consciousness.' *Behavioral and Brain Sciences* 18, 227-247.

Brownstein, Michael. (2015). 'Attributionism and Moral Responsibility for Implicit Bias.' *Review of Philosophy and Psychology*, 765-786.

Brownstein, Michael, Madva, Alex, and Gawronski, Bertram (Manuscript). 'Meta-Analyses and Predicting Behavior: In Defense of Implicit Attitude Measures.'

Buchak, Lara. (2014). 'Belief, credence, and norms.' *Philosophical Studies*,*169*(2), 285-311.

Bushman, Brad J., DeWall, C. Nathan, Pond, Richard S., & Hanus, Michael D. (2014). 'Low glucose relates to greater aggression in married couples.' *Proceedings of the National Academy of Sciences*, *111*(17), 6254-6257.

Calhoun, Cheshire. (1989). 'Responsibility and reproach.' *Ethics*, *99*(2), 389-406.

Cameron, C. Daryl, Payne, B. Keith, & Knobe, Joshua. (2010). 'Do Theories of Implicit Race Bias Change Moral Judgments?' *Social Justice Research*, *23*(4), 272-289.

Carpenter, Siri. (April/May 2008). 'Buried Prejudice.' *Scientific American Mind*, 35.

Chartrand, Tanya L., van Baaren, Rick B., and Bargh, John A. (2006). 'Linking automatic evaluation to mood and information processing style: consequences for experienced affect, impression formation, and stereotyping.' *Journal of Experimental Psychology: General*, *135*(1), 70.

Coates, D. Justin, and Swenson, Philip. (2013). Reasons-responsiveness and degrees of responsibility. *Philosophical Studies*, *165*(2), 629-645.

Cooley, Erin , Payne, B. Keith, Loersch, Chris , and Lei, Ryan. (2015). 'Who owns implicit attitudes? Testing a metacognitive perspective.' *Personality and Social Psychology Bulletin*,*41*(1), 103-115.

Cooley, Erin , Payne, B. Keith, and Phillips, K. Jean (2014). 'Implicit bias and the illusion of conscious ill will.' *Social Psychological and Personality Science*, *5*(4), 500-507.

Cortina, Lilia M., Kabat Farr, Dana, Leskinen, Emily, Huerta, Marisela and Magley, Vicki J. (2011). 'Selective incivility as modern discrimination in organizations: Evidence and impact.' *Journal of Management*.

Crowther, Bosley. (November 12 1947). '"Gentleman's Agreement," Study of Anti-Semitism, Is Feature at Mayfair—Gregory Peck Plays Writer Acting as Jew.' *The New York Times*. URL = < http://movies.nytimes.com/movie/review?res=9E0DE7DE113AE233A25751C1A9679D946693D6CF >

Danziger, Shai, Levav, Jonathan, and Avnaim-Pesso, Liora. (2011). 'Extraneous factors in judicial decisions.' *Proceedings of the National Academy of Sciences,108*(17), 6889-6892.

Devine, Patricia G., Monteith, Margo J., Zuwerink, Julia R., and Elliot, Andrew J. (1991). 'Prejudice with and without compunction.' *Journal of Personality and Social Psychology,60*(6), 817-830.

Dovidio, John F., Kawakami, Kerry, and Gaertner, Samuel L. (2002). 'Implicit and explicit prejudice and interracial interaction.' *Journal of Personality and Social Psychology* 82, 62-68.

Farinola, Michele, and Freedman, Mimi. (Directors). (October 21, 2001). *AMC Backstory: Gentleman's Agreement*. United States: Prometheus Entertainment.

Faucher, Luc. (2016). 'Revisionism and Moral Responsibility for Implicit Attitudes.' In Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume 2: Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

Fischer, John M., and Ravizza, Mark. (1998). *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.

Forgas, Joseph P. (2011). 'She just doesn't look like a philosopher…? Affective influences on the halo effect in impression formation.' *European Journal of Social Psychology, 41*(7), 812-817.

Fricker, Miranda. (2010). 'The relativism of blame and William's relativism of distance.' *Aristotelian Society Supplementary Volume* 84(1): 151–77.

Gawronski, Bertram, Brochu, Paula M., Sritharan, Raj, and Strack, Fritz. (2012). 'Cognitive consistency in prejudice-related belief systems: Integrating old-fashioned, modern, aversive and implicit forms of prejudice.' *Cognitive consistency: A fundamental principle in social cognition*, 369-389.

Gawronski, Bertram, and LeBel, Etienne P. (2008). 'Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not.' *Journal of Experimental Social Psychology*, 44, 1355–1361.

Gladwell, Malcolm. (2005). *Blink: The Power of Thinking Without Thinking.* New York: Little, Brown and Company.

Glasgow, Joshua. (2016). 'Alienation and Responsibility.' In Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume 2: Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

Gonsalkorale, Karen, von Hippel, William, Sherman, Jeffrey W., and Klauer, Karl C. (2009). 'Bias and regulation of bias in intergroup interactions: Implicit attitudes toward Muslims and interaction quality.' *Journal of Experimental Social Psychology,45*(1), 161-166.

Greenwald, Anthony G., and Banaji, Mahzarin R. (1995). 'Implicit social cognition: Attitudes, self-esteem, and stereotypes.' *Psychological Review* 102, 4-27.

Greenwald, Anthony G., Mahzarin R. Banaji, and Brian A. Nosek. (2015). 'Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects.' *Journal of Personality and Social Psychology* 108 (4): 553–61. doi:10.1037/pspa0000016.

Griffin, John H. (1961). *Black Like Me*. Boston: Houghton Mifflin.

Griffiths, Paul E. (1989). ‘Folk, Functional, and Neurochemical Aspects of Mood.’ *Philosophical Psychology* 2, 17-33.

Hahn, Adam, Judd, Charles M., Hirsh, Holen K., and Blair, Irene V. (2014). ‘Awareness of implicit attitudes.’ *Journal of Experimental Psychology: General*, *143*(3), 1369.

Holland, Rob W., de Vries, Marieke, Hermsen, Berlinda, and van Knippenberg, Ad. (2012). ‘Mood and the Attitude-Behavior Link: The Happy Act on Impulse, the Sad Think Twice.’ *Social Psychological and Personality Science, 3*(3), 356-364.

Holroyd, Jules. (2012). ‘Taking Responsibility for Bias.’ In Special Edition of *Journal of Social Philosophy*, edited by M. Crouch & L. Schwartzman, 43:3, 274-306.

Holroyd, Jules. (2015). ‘Implicit bias, awareness and imperfect cognitions.’ *Consciousness and Cognition*, *33*, 511–523. https://doi.org/10.1016/j.concog.2014.08.024

Holton, Richard, & Berridge, Kent. (2013). ‘Addiction between compulsion and choice.’ *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience», Oxford University Press, New York*, 239-268.

Jacoby-Senghor, Drew S., Sinclair, Stacey, and Shelton, J. Nicole (2016). ‘A lesson in bias: The relationship between implicit racial bias and performance in pedagogical contexts.’ *Journal of Experimental Social Psychology*, *63*, 50-55.

Kalev, Alexandra, Dobbin Frank, and Kelly Erin. (2006). ‘Best Practices or Best Guesses? Diversity Management and the Remediation of Inequality.’ *American Sociological Review* 71, 589-917.

Kazan, Elia. (Director). (1947). *Gentleman’s Agreement*. United States: Twentieth Century Fox.

King, Matt and Carruthers, Peter. (2012). ‘Moral responsibility and consciousness.’ *Journal of Moral Philosophy*, *9*(2), 200-228.

Kelly, Daniel and Roedder, Erica. (2008). ‘Racial Cognition and the Ethics of Implicit Bias.’ *Philosophy Compass*, 3 (3), 522-540, doi:10.1111/j.1747-9991.2008.00138.x.

Korsgaard, Christine M. (1997). ‘The Normativity of Instrumental Reason.’ In *Ethics and Practical Reason*, edited by G. Cullity and B. Gaut, 27-68. Oxford: Clarendon Press.

Lawrence III, Charles R. (1987). ‘The id, the ego, and equal protection: Reckoning with unconscious racism.’ *Stanford Law Review*, 317-388.

Levinson, Justin D., and Smith, Robert J. (Eds.). (2012). *Implicit Racial Bias across the Law*. Cambridge, UK: Cambridge University Press.

Levy, Emanuel. *Gentleman’s Agreement (1947)*. URL = <http://www.emanuellevy.com/review/gentlemans-agreement-1947-6/>

Levy, Neil. (2011). ‘Expressing who we are: Moral responsibility and awareness of our reasons for action.’ *Analytic Philosophy*, *52*(4), 243-261.

Levy, Neil. (2014a). ‘Consciousness, implicit attitudes and moral responsibility.’ *Noûs, 48*(1), 21-40.

Levy, Neil. (2014b). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Levy, Neil. (2015). ‘Neither fish nor fowl: Implicit attitudes as patchy endorsements.’ *Noûs*, 49 (4), 800-823.

Levy, Neil. (2017). ‘Implicit Bias and Moral Responsibility: Probing the Data.’ *Philosophy and Phenomenological Research*, *94*(1), 3–26. https://doi.org/10.1111/phpr.12352

Madva, Alex. (2012.) *The Hidden Mechanisms of Prejudice: Implicit Bias and Interpersonal Fluency*. PhD dissertation, New York: Columbia University.

Madva, Alex. (2016a). ‘Virtue, Social Knowledge, and Implicit Bias.’ In *Implicit Bias and Philosophy, Volume 1*, edited by Michael Brownstein and Jennifer Saul, 191–215. New York: Oxford University Press.

Madva, Alex. (2016b). ‘A Plea for Anti-Anti-Individualism: How Oversimple Psychology Misleads Social Policy.’ *Ergo* 3 (27): 701–28. doi:10.3998/ergo.12405314.0003.027.

Madva, Alex. (Forthcoming). ‘Biased against Debiasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle against Prejudice.’ *Ergo*.

Madva, Alex. (Manuscript). ‘Social Psychology, Phenomenology, & the Indeterminate Content of Unreflective Racial Bias.’ In preparation to appear in Emily Lee (ed.), *Race and Phenomenology*, Rowman & Littlefield.

Madva, Alex, and Michael Brownstein. (2017). ‘Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind.’ *Noûs*, doi:10.1111/nous.12182.

McConnell, Allen R., and Leibold, Jill M. (2001). ‘Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes.’ *Journal of experimental Social psychology*, *37*(5), 435-442.

Mele, Alfred. (2006). *Free Will and Luck*. New York: Oxford University Press.

Mills, Charles W. (1997). *The Racial Contract*. Ithaca: Cornell University Press.

Nier, Jason A. (2005). ‘How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach.’ *Group Processes and Intergroup Relations* 8, 39-52.

Norton, Michael I., Sommers, Samuel R., Apfelbaum, Evan P., Pura, Natassia, and Ariely, Dan. (2006). ‘Color blindness and interracial interaction playing the political correctness game.’ *Psychological Science*, *17*(11), 949-953.

Oswald, Frederick L., Gregory Mitchell, Hart Blanton, James Jaccard, and Philip E. Tetlock. (2013). ‘Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies.’ *Journal of Personality and Social Psychology* 105 (2): 171–92. doi:10.1037/a0032734.

Pickard, Hanna. (2011). ‘Responsibility without blame: empathy and the effective treatment of personality disorder.’ *Philosophy, psychiatry, & psychology: PPP,18*(3), 209.

Prescott-Couch, Alexander. (2005). ‘What Is a Mood?’ *The Yale Philosophy Review*, 2, 42-58.

Richeson, Jennifer A., & Shelton, J. Nicole. (2007). ‘Negotiating Interracial Interactions Costs, Consequences, and Possibilities.’ *Current Directions in Psychological Science*, *16*(6), 316-320.

Ranganath, Kate A., Smith, Colin T., and Nosek, Brian A. (2008). ‘Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods.’ *Journal of Experimental Social Psychology*, *44*(2), 386-396.

Raz, Joseph. (2010). ‘Being in the world.’ *Ratio*, *23*(4), 433-452.

Rotton, James, Barry, Timothy, Frey, James, and Soler, Edgardo. (1978). ‘Air Pollution and Interpersonal Attraction.’ *Journal of Applied Social Psychology*, *8*(1), 57-71.

Salvatore, Jessica, and Shelton, J. Nicole. (2007). ‘Cognitive costs to exposure to racial prejudice.’ *Psychological Science*, 18, 810-815.

Saul, Jennifer. (2013). ‘Unconscious Influences and Women in Philosophy.’ In *Women in Philosophy: What Needs to Change?* 39-60, edited by F. Jenkins and K. Hutchison. Oxford: Oxford University Press.

Scaife, Robin, Stafford, Tom, Bunge, Andreas, and Holroyd, Jules. (Manuscript). The effects of moral interactions on implicit bias.

Sher, George. (2009). *Who Knew? Responsibility Without Awareness*. New York. NY: Oxford University Press.

Schlenker, Barry R., and Darby, Bruce W. (1981). 'The use of apologies in social predicaments.' *Social Psychology Quarterly* 4, 271-278.

Schwitzgebel, Eric. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, *91*(4), 531-553.

Siemer, Matthias. (2009). 'Mood Experience: Implications of a Dispositional Theory of Moods.' *Emotion Review* 1 (3), 256-263.

Sims, Carra S., Drasgow, Fritz, and Fitzgerald, Louise. (2005). 'The effects of sexual harassment on turnover in the military: time-dependent modeling.' *Journal of Applied Psychology* 90, 1141-1152.

Sinnott-Armstrong, Walter. (2013). 'Are Addicts Responsible?' *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*, 122-142.

Sizer, Laura. (2000). 'Towards a computational theory of mood.' *The British journal for the philosophy of science*, *51*(4), 743-770.

Smith, Angela M. (2004). 'Conflicting attitudes, moral agency, and conceptions of the self.' *Philosophical Topics*, *32*(1/2), 331-352.

Smith, Angela M. (2005). 'Responsibility for attitudes: Activity and passivity in mental life.' *Ethics*, *115*(2), 236-271.

Sprigle, Ray. (1949). *In the Land of Jim Crow*. New York: Simon and Schuster.

Strawson, Peter F. (1962). *Freedom and resentment and other essays*. London: Methuen.

Sue, Derald W., Capodilupo, Christina M., Torino, Gina C., Bucceri, Jennifer M., Holder, Aisha M.B., Nadal, Kevin L., and Esquilin, Marta. (2007). 'Racial microaggressions in everyday life: Implications for clinical practice.' *American Psychologist* 62, 271-286.

Valian, Virginia. (1998). *Why so slow? The advancement of women*. Cambridge, MA: M.I.T. Press.

Washington, Natalia T., and Kelly, Daniel. (2016). 'Who's Responsible for This? Moral Responsibility, Externalism, and Knowledge about Implicit Bias.' In Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume 2: Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

Yancy, George. (2008). 'Elevators, social spaces and racism A philosophical analysis.' *Philosophy & Social Criticism*, *34*(8), 843-876.

Young, Iris Marion. (1990). *Justice and the Politics of Difference*. Princeton: Princeton University Press.

Young, Iris Marion. (2011). *Responsibility for Justice*. New York: Oxford University Press.

Zheng, Robin. (2016). 'Attributablity, Accountability, and Implicit Bias.' In Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume 2: Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

[1] The ambiguity of such rude behavior is often one of its harms. Members of stigmatized groups can find ambiguous but potentially biased behavior more unsettling than outright discrimination (Salvatore and Shelton, 2007; Sue et al. 2007). Such ambiguity also contributes to the difficulty of measuring implicit biases and their effects on discriminatory behavior (cf. Oswald et al., 2013; Greenwald, Banaji, and Nosek 2015), although I argue elsewhere that concerns about the replicability and real-world import of implicit bias research are overblown, conflating important-but-localized outstanding empirical questions with hyperbolic doubts about the entire field (Madva and Brownstein 2017; Brownstein, Madva, and Gawronski, ms). See also note #26.

[2] On structural-institutional interventions, see Valian (1998), Kalev et al. (2006), Anderson (2010), Levinson and Smith (2012), and Madva (2016b, forthcoming).

[3] See Madva (2012, 2016a,b, forthcoming, ms), although I intend to say more in future work about the forward-looking role that responsibility and blame can play in motivating activism to initiate and implement structural-institutional reform. Cf. Calhoun (1989), Young (1990, 2011), Zheng (2016), and Scaife, Stafford, Bunge, and Holroyd (manuscript).

[4] See also Levy (2014a,b; 2015; 2017), Saul (2013), and the references in note #**Error! Bookmark not defined.**17.

[5] This paper is written in a compatibilist spirit. See, e.g., Fischer and Ravizza (1998) for a compatibilist interpretation of control, and Levy (2014b) for an empirically-informed account of how awareness is necessary for control.

[6] See Holroyd (2012) for trenchant discussion of the differences between these questions, and a persuasive argument that individuals may sometimes be responsible for simply acquiring and harboring implicit biases.

[7] The same pattern was found in studies that measured, rather than manipulated, participants' views about whether their gut reaction reflected their genuine attitude ('my own beliefs,' 'my real attitude'), and whether their 'gut reactions reflect accurate beliefs about homosexuals' (Cooley et al. 2014, 2015). Angela Smith (2004, 2005) argues that individuals *should* accept that their unreflective prejudices are genuinely their 'own' attitudes, because taking ownership in this way will prompt individuals to take responsibility to do something about them. At first glance, these findings suggest that 'owning' these attitudes leads participants to reflectively endorse them, rather than to resist them. See Zheng (2016, 80-2) regarding further studies with similar upshots. Identifying the downstream effects of various ways of conceptualizing our biases is ultimately an empirical question, about which I say more in (ms). See also Scaife et al. (ms).

[8] Perhaps none of our attitudes are directly introspectible (King and Carruthers 2012), and all self-knowledge of attitudes is indirect or inferential. If so, this research suggests that our capacities to infer the contents of our implicit and explicit attitudes are surprisingly comparable. See also note #14.

[9] See Holroyd (2012, 292-4) for discussion of follow-up studies.

[10] As reported in Farinola and Freedman (October 21, 2001).

[11] As reported by Emanuel Levy.

[12] Crowther (November 12, 1947).

[13] For one survey of earlier psychological research and theories, and analysis of their normative and legal implications, see Lawrence (1987).

[14] In a similar way, Levy's (2017) Exonerating arguments (e.g., that implicit biases are highly uncontrollable, unintegrated with other attitudes, and unresponsive to reason) often apply equally well to explicit attitudes. In fact, although Levy's essay is purportedly focused on implicit attitudes, much of the research he discusses is actually about self-reported attitudes (e.g., he discusses 'celebrity contagion' research, which finds that participants *explicitly report* that they'd pay less money for clothing previously worn by celebrities if the clothing has been washed). Thus his arguments might suggest that human beings are, in general, simply too irrational to be rightly held responsible for anything, but they don't establish a categorical (non-graded), normative difference between implicit and explicit attitudes.

[15] Cf. Gawronski et al. (2012) and Cooley et al. (2014, 2015).

[16] See Glasgow (2016) and Faucher (2016, 125-31) for compelling accounts of additional factors that help to explain our intuitions regarding moral responsibility for implicit discrimination.

[17] Similarly, Kelly and Roedder (2008, 532) suggest that implicit biases might be morally wrong and even 'condemnable' without being blameworthy; Fricker (2010) distinguishes between 'moral-epistemic disappointment' and blame; Anderson (2010) distinguishes between merely 'racially stigmatizing' and genuinely racist behavior; and Pickard (2011, 209, 216) argues, in the context of clinical treatment, that addicts should be held accountable for controlling their impulses but not saddled with the 'sting' of 'affective blame' when they fall short. Zheng cites several earlier theorists who take roughly this line, including Young (2011), who (citing theorists like Lawrence, 1987) argued as early as 1990 that individuals were forward-looking responsible but not backward-looking blameworthy for their unreflective prejudices.

[18] While Levy (2011, 2014a,b, 2017) takes the hardline stance that individuals are categorically not responsible for local control over their implicit biases, his writings on moral responsibility, awareness, control, and reasons-responsiveness make frequent reference to degrees. For example, he writes that, 'the degree of accessibility of information seems to correlate (roughly) with the degree of moral responsibility of the agent for failing to utilize it' (2014b, 32). Insofar as the evidence suggests that implicit biases are in-between accessible to individuals, why not conclude that individuals are in-between responsible for accessing them? Levy instead asserts that only 'easy and effortless' accessibility suffices for direct responsibility (2014b, 33), but this assertion is difficult to square with the

prior quote about degrees of accessibility. Setting the threshold here (or anywhere) seems arbitrary given that we are clearly ‘dealing with a continuum’ (2014b, 83).

A *PPQ* referee points out that there is at least one very important context in which we make binary responsibility judgments, namely, in guilty-or-not legal decisions, before engaging in the more complex questions of sentencing and punishment. As I say in the following paragraph and §4, my view is that individuals typically *are* responsible and blameworthy for their implicit discriminatory acts (if forced to choose, my verdict is ‘guilty’), such that it is morally incumbent upon us to make good on them and do better, but the degraded accessibility, controllability, reasons-responsiveness, etc., of our implicit biases are mitigating factors.

[19] For empirical investigation into academic and non-academic intuitions and views about moods, and, in particular, how to distinguish moods from emotions, see Beedie et al. (2005).

[20] Some who argue that moods have no necessary connection to qualitative experience (e.g., Griffiths, 1989, 28; Prescott-Couch, 2005, 56) overlook the possibility that moods might be felt without being noticed. See Siemer (2009) for an account of moods that privileges their experiential and affective components.

[21] Cf. Coates and Swenson (2012).

[22] Cf. Smith (2004, 344-5).

[23] A more frequently discussed mitigating factor, more extreme than the moods case, is depression (Coates and Swenson, 2012). As Korsgaard (1997, 41) suggests, ‘people’s terror, idleness, shyness, or depression… [are] forces that block their susceptibility to the influence of reason.’ The so-called mood disorders are psychologically and normatively very different from transient moods (Pickard 2011). Thanks to Karen Harkins for insightful discussion of this issue.

[24] See, e.g., Schlenker and Darby (1981) and Glasgow (2016). It could be that part of what explains the divergence of intuitions surrounding responsibility for implicit discrimination is that Exonerators are more focused on the (typically) smaller-scale, local consequences of implicit discrimination while non-Exonerators are focused on the large-scale, aggregate consequences.

[25] Thanks to Susanna Siegel for insightful questions and advice about distinguishing between effects of moods.

[26] Studies such as this one reveal how implicit attitudes and moods are more than analogically related. Chartrand et al. (2006) explore numerous relationships and interactions between moods and implicit attitudes, for example, finding that priming techniques that activate valenced implicit attitudes induce moods with the same valence. Holland et al. (2012) found, like Forgas, that participants in happy moods acted on the basis of their implicit attitudes, while those in sad moods acted on the basis of their reflective beliefs. Such findings also highlight some of the many contextual mediators and moderators that determine when and how implicit attitudes influence behavior, and help to explain why meta-analyses that ignore such variables are likely to be misleading (Brownstein, Madva, and Gawronski, ms). See note #1.

[27] Cf. Adams (1985, esp. 25-7) on cognitive sins, Sher (2009, 21) on culpable ignorance, and Raz (2010) on domains of secure competence.

[28] It must be acknowledged that many women live with experiences of vulnerability that partly inform what’s going on in this encounter. It’s possible, for example, that this individual’s past experiences with harassment make it the case that she’d be uncomfortable alone with any man in this situation. Yancy explores the gendered, intersectional complexity of this scenario, and the woman’s potential perspectives on the encounter, in far greater depth. However, if we stipulate, e.g., that they are riding the elevator in an otherwise-crowded, publicly accessible building in the middle of the day in an extremely low-crime area, etc., it is reasonable to expect that many white women would not feel equally uncomfortable around a similarly dressed white man. There are also countless cases of white *men* who react with instinctive fear or discomfort toward black men, and black women, despite being in obviously safe conditions. Again, this is effectively *what 70% of participants admitted* in Devine et al. (1991).

[29] This essay benefited enormously from feedback and conversation with many people, including Taylor Carman, Katie Gasdaglis, Tamar Szabó Gendler, Lydia Goehr, Christia Mercer, Patricia Kitcher, and Virginia Valian; from the audience at the 2013 Pacific Division meeting of the American Philosophical Association, especially from insightful commentary by Manuel Vargas; from extensive discussion with Michael Brownstein, Natalia Washington, and Daniel Kelly during a symposium on implicit bias and moral responsibility at the 2013 meeting of the Southern Society for Philosophy and Psychology, including excellent questions from Susanna Siegel and Elizabeth Schechter; and from students in my Spring 2013 seminar at Berkeley, including Shannon Doberneck, Karen Harkins, Lianna Mecano, Dylan Murray, Jeremy Pober, and Jen Kanyuk. This essay was revised with institutional support from the Mellon Postdoctoral Fellowship at the University of California, Berkeley; Vassar College; and California State Polytechnic University, Pomona.