Alex Madva Pacific APA, March 2013

Implicit Bias, Moods, and Moral Responsibility

Abstract

Implicit social biases can influence our behavior in harmful and unjust ways, but they seem to operate outside of consciousness and control. Are individuals morally responsible for their implicitly biased behavior, or are they merely non-culpably implicated in a broader systematic injustice? One reason some philosophers have denied that individuals are responsible for their biases is that they are not sufficiently aware of them. They fail the necessary "awareness condition" for moral responsibility. However, recent empirical evidence suggests that individuals are often aware of their biases in some senses but not others. To argue that this state of partial awareness satisfies the awareness condition, I offer an argument by analogy to a close relative of implicit bias: moods. The degree of awareness individuals have of their moods meets the awareness condition, and the type of awareness individuals have of their implicitly biased behavior is importantly similar.

I. Introduction

In a 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 nevertheless exhibited racial bias on an indirect measure. Subsequently, their explicit reports best predicted the friendliness of what they said to a 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 black interlocutors than white interlocutors. Strikingly, the white participants generally formed positive impressions of the interactions, while their black interlocutors believed 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."[1]

There is good reason to think that these unfriendly microbehaviors make significant contributions to large-scale injustices. Widespread and frequent repetitions of such microbehaviors constitute "microinequities," which can stack up over time to reinforce macro-level disparities between social groups.[2] For example, Lilia Cortina and colleagues have found that women, and especially women of color, tend to report experiencing more interpersonal *incivility* in the workplace than do men, whether the workplace is a city government, a law enforcement agency, or the US Military.[3] In many cases, the incivility does not consist in overt harassment, or involve explicit reference to gender or race, or result from any obvious intent to do harm; rather it consists in generic forms of rudeness, such as speaking condescendingly to or interrupting a colleague. Discourteous behavior of this sort is deeply ambiguous: since just about *everybody* interrupts and gets interrupted *sometimes*, it is extremely difficult to identify any particular instance of interruption as expressive of bias, as opposed to, say, of enthusiasm for the topic of conversation.[4] But because women, and especially women of color, are evidently treated in such uncivil ways much more often, it should come as no surprise that they are, as Cortina also found, much more likely to report the intention to quit. Cortina's findings contribute to broader patterns of evidence that suggest that experiencing a work environment as hostile leads one to quit.[5]

---

[1] As reported by Carpenter (2008).
[2] Valian's (1998) account of the gradual accumulation of advantage for men and disadvantage for women captures the structure of this phenomenon well.
[3] Cortina et al. (2008, 2011)
[4] The ambiguity of such rude behavior is often precisely one of its harms. Evidence suggests that members of stigmatized social groups can find ambiguous but potentially biased behavior highly unsettling and cognitively taxing. See Salvatore and Shelton (2007) and Sue et al. (2007).
[5] Feeling mistreated at work is about as good a reason to leave as any. Cortina and colleagues also find that merely *observing* uncivil behavior does harm and leads to turnover intentions. Their data is correlative rather than causal,

Findings like these suggest that the subtle expressions of implicit biases can have significant negative consequences. Reflection on how to correct bias leads to the pressing philosophical question: are individuals morally responsible and blameworthy for their biased microbehaviors, or are they merely non-culpably implicated in a broader systematic injustice? Philosophers have begun to weigh in on this question, often taking the stance that individuals are not responsible for their implicitly biased behavior because they are not sufficiently aware of or in control of it.[6] These views are misguided. Here I offer a first step towards a larger argument that individuals are responsible for their implicitly biased behavior. After considering the apparent reasons to deny personal responsibility for implicit bias (§II), I review empirical evidence that suggests that individuals are often more aware of their biases than previously recognized (§III). Then I begin to explain how this kind of awareness might make individuals responsible for their biased behavior (§IV).

II. Précis of the Case for Blamelessness

---

so it is possible, e.g., that a given individual's thoughts of quitting preceded observations and experiences of incivility, but see, e.g., Sims et al. (2005) for evidence suggesting causal effects of more blatant forms of harassment on actual turnover behavior.

[6] See Kelly and Roedder (2008), Levy (2012), and Saul (forthcoming). A noteworthy exception is Holroyd (forthcoming), the only paper I know of in the "analytic" philosophical tradition that defends individual responsibility for implicit bias outright. I am in broad and detailed agreement with Holroyd, although I advance a different argument for similar conclusions in what follows. I add the qualifier about *analytic* philosophy because there seems to be a greater willingness to attribute responsibility for tacit prejudice among theorists working outside mainstream Anglo-American philosophy (e.g., Sullivan 2006), although even in these domains, assertions of responsibility typically remain somewhat ambiguous, as I explain in Madva (2012, Ch.3, §II).

Of course, to consider the responsibility individuals might bear for their implicitly biased behavior is not to assume that the harms associated with implicit bias are primarily "individual" rather than "social." The harms and inequities suffered by individuals on the basis of race and gender depend to a great extent on social-institutional forces, and institutional change is necessary for redressing those harms. But institutions are composed both of a set of rules and laws as well as a set of individuals, and, if we want to bring about lasting change, we have to understand the roles that each plays in contributing to these large-scale harms. Here I focus on the role of individuals and, in particular, the responsibility they bear for their own implicit discriminatory behaviors.

There are a number of reasons one might think individuals are blameless for possessing and expressing implicit social biases. Saul (forthcoming) summarizes several:

> A person should not be blamed for an implicit bias that they are completely unaware of, which results solely from the fact that they live in a sexist culture. Even once they become aware that they are likely to have implicit biases, they do not instantly become able to control their biases, and so they should not be blamed for them. (They may, however, be blamed if they fail to act properly on the knowledge that they are likely to be biased— e.g. by investigating and implementing remedies to deal with their biases.)

Saul here invokes several considerations that seem to mitigate responsibility for bias, the most powerful of which are (1) that individuals do not seem to be aware of them, and (2) when made aware, that they cannot control their biased behavior in the right way.[7] I take up (2), the control condition for moral responsibility, in Madva (2012, Ch.3, §VI-VII), and I focus here on (1), the awareness condition for moral responsibility. For example, the participants in Dovidio's study might not have been aware that implicit 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. Awareness seems like a non-negotiable necessary condition for responsibility, and these subjects seemed to lack it.

III. Consciousness of Implicit Biases

Is it typically the case that individuals are, as Saul suggests, "completely unaware" of their implicit biases? Implicit biases were originally thought to be unconscious because participants did not or could not report them, as in studies like Dovidio's, where participants' explicit and

[7] I discuss the case for blamelessness in greater depth in Madva (2012, Ch.3, §II). There may also be a strategic (i.e., non-metaphysical) basis for denying responsibility for bias: it might be counterproductive for those motivated to effect meaningful social change to foist personal responsibility onto people for their biases, which could lead them to become defensive and resentful. See Saul (forthcoming) and Holroyd (forthcoming) for further discussion.

implicit attitudes stood in stark contrast.[8] But recent evidence suggests otherwise. Researchers have since found a number of ways to draw explicit and implicit attitudes together. If participants are first told to "focus on their feelings," their self-reports tend to move closer to their performance on indirect measures.[9] A similar effect occurs when participants are told that the implicit measure is "the closest thing to a lie detector that social psychologists can use to determine your true beliefs about race."[10] In fact, simply telling participants whether their "gut feelings" do or do not reflect their "genuine" views may influence their self-reports.[11] If told that the gut reaction people have to photos of homosexual couples "usually reflects people's genuine attitude towards homosexuality," participants (who tend to have negative gut reactions to such photos) are more likely to report that gay people should not be allowed to marry or join the military. By contrast, if told that gut reactions do *not* reflect genuine attitudes, participants are more likely to report that gay people *should* be allowed to marry. Many empirical questions remain unanswered, but it is clear that we cannot simply cast implicit biases into what popular authors such as Malcolm Gladwell (2005) call "the locked door of the unconscious." The working hypothesis should be that the affective elements of implicit biases contribute, or are available, to conscious experience, although in many instances without being the object of explicit attention. As researchers explain, "hard evidence that people have attitudes and beliefs that they don't know about, or can't know about when they try, is difficult to find."[12]

There are also familiar examples from film and literature of our awareness of implicit bias. In the film *Gentleman's Agreement*, a reporter pretends to be Jewish for six weeks in order to study anti-Semitism. In effect, he encounters repeated expressions of implicit bias. When he

[8] See Greenwald and Banaji (1995).
[9] Gawronski and LeBel (2008).
[10] Nier (2005, 43).
[11] Payne (April 2012).
[12] Hall and Payne (2010, 222).

explains to his romantic interest 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 re-shot because her original look of dismay was too overt; it had to be more subtle. 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."[13] *Gentleman's Agreement* came out in 1947, winning Academy Awards for Best Picture, Best Director, and Best Supporting Actress. While research in contemporary social psychology is expanding our understanding of implicit bias, the phenomenon has been part of our collective awareness for quite some time, and with it, I argue, a degree of responsibility.

IV. Awareness, Moods, and Reactive Attitudes

The findings discussed above can be understood in light of distinctions familiar from the philosophy of mind between the content of one's phenomenology (i.e., that which is experienced) and the content of one's focal attention. In Block's terms, implicit attitudes seem to be *phenomenally conscious*, if not (always) *access conscious*.[14] They may often be *felt* without being *noticed*, just as a person can be in a grumpy or lighthearted mood without noticing as much.[15]

Individuals seem to be aware of their biases in some senses but not others. But is this state of partial awareness sufficient to satisfy (1), the awareness condition for moral

[13] As reported by Emanuel Levy (http://www.emanuellevy.com/review/gentlemans-agreement-1947-6/).
[14] Block (1995). My argument could be reformulated in terms of theories that deny the distinction between phenomenal and access consciousness, such as a higher-order theory, by describing the states which I here call phenomenally conscious as *potentially* conscious. Agents *can* become focally conscious of them.
[15] Social psychologists have also begun to recognize the relevance of these distinctions to research on implicit attitudes. See Gawronski et al. (2006) and Hall and Payne (2010).

responsibility? To answer this question, I offer an argument by analogy to a close relative of implicit bias: moods. I argue that the degree of awareness individuals have of their moods meets the awareness condition. The type of awareness individuals have of their implicitly biased behavior is importantly similar.

Affective phenomenal experience plays a central role in our ordinary practices of assigning responsibility and blame. The reactive attitudes that Strawson (1974) cited as integral to our understanding of responsibility were not the cold cognitive evaluations judges and jurors are expected to make in determining the scope of a defendant's criminal responsibility. Rather, they were automatic and affective. We care about whether people bear us "good will" or "ill will," whether they appreciate or resent us, smile or frown. Expressions of good or ill will are as much a matter of the initial automatic reactions we observe in others as they are a matter of the reflective judgments they make about those reactions. In turn, we take these automatic, affect-expressive behaviors of others to license certain affect-laden responses from us, such as when I feel resentful toward you for being short with me.

These affective reactions are part of phenomenal awareness. Agents 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. Consider how we commonsensically understand the effects of *moods* on behavior, and how this understanding in turn figures in our practices of attributing responsibility and blame. I might not notice that I'm in a grumpy mood, but I may be in one just the same, feeling it all the while. I might have no knowledge of the causal source of my mood and my mood might have all sorts of unknown effects on what I think and do. Still, it would be hasty to conclude that the mood itself was in any deep respect unconscious. Perhaps my mood passes in and out of focal awareness, or

perhaps it just hovers in the periphery. These are empirical questions, albeit ones notoriously difficult to tackle.

But the fact that I fail to *notice* my grumpy mood would not simply exonerate my rude behavior.[16] We routinely hold others and ourselves responsible for the things said and done because of bad moods. It's true that being in a bad 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 (perhaps it is less intentional or less "personal"), somehow mitigating the severity of the offence. 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.

Compare the case of mood-related incivility to Nagel's (1979) example of an extremely safe driver who gets into an accident because her brakes suddenly stop working.[17] Here, it seems, the driver should bear no responsibility or blame for the resultant harms. Nevertheless, we can easily imagine the driver feeling awful and apologizing profusely, and we can imagine victims of the incident harboring ill will toward the driver. But when the driver expresses regret for the accident, would the (appropriate) reply be, "Apology accepted!" or would it instead be, "You can't beat yourself up about this. You did nothing wrong"? I think, clearly, something more like the latter.

---

[16] Some philosophers identify mood-related misbehavior as blameless. For example, Levy (2011, 245) writes, "George's shortness with his colleagues might be excused because of the stress he has been under recently," because George is not properly aware of the reasons for his acting that way. A more frequently discussed case, which I take to be more extreme but structurally similar to the moods case, is whether *depression* mitigates responsibility. For example, Korsgaard (1997, 41) suggests that, "people's terror, idleness, shyness, or depression… [are] forces that block their susceptibility to the influence of reason." Broadly speaking, I agree that these emotional influences play a mitigating role, but these philosophers err in concluding that they fully exculpate bad behavior. A conception of responsibility as *coming in degrees* can straightforwardly accommodate these cases (Madva 2012, Ch.3). For more on "degrees of responsibility," see Mele (2006, 129-132) and Coates and Swenson (2012).

[17] One might also recall the recent spate of Toyota cars that would not stop accelerating.

Does "being in a bad mood" or "being stressed out" have the same blame-defeating force? When an individual snaps at a friend, and subsequently apologizes, she might say, "I'm sorry for being obnoxious. 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 the friend respond in this case? Would the friend say, "Come now, you have *nothing* to apologize for. You didn't do anything wrong"? More likely, the friend 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 in this case as it is for the unlucky driver. This is 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 does not transport an individual out of the realm of responsibility and blame, but shifts her location within that space.

Of course, it makes a difference what sort of behavior is supposed to be illuminated by reference to the bad mood. If a bad mood leads an individual to under-evaluate a job applicant, or to punch someone in the face, then citing it will do considerably less exculpatory work. Other things being equal, the more serious the consequences of the behavior, the less mitigating we'll take a bad mood to be.[18] 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 her rude impulses in the presence of an armed mugger than in the presence of a close friend, or in the presence of her bosses than in the presence of subordinates.

[18] See Schlenker and Darby (1981) for evidence that the severity of the consequences is directly related to people's tendencies of giving "nonperfunctory apologies… expressing remorse, and offering to help the victim" (271).

Suppose these external factors do influence how easy it is to control the influence of moods on our behavior. The upshot is not that agents ought to be excused for mood-related misbehavior when it is difficult to control or the stakes are low, 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 try to capture all the nuances of the 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 unwittingly influenced by this unnoticed psychological state does not transport her out of the realm of responsibility for her behavior.

Now consider a case of behavior that might express not (merely) a bad mood but an implicit bias. If Stephanie wrinkles her nose and shows other signs of automatic disgust (rather than sympathy) at her friend Dennis' choice of attire, she can reasonably be held responsible and blameworthy for that automatic reaction, even if she might not endorse it upon reflection. Dennis, the target of the automatic disgust, would not, I take it, be blamed for resenting the disgusted reaction and holding it against her. It would likely make *some* difference if Stephanie were to apologize for the disgusted reaction and explain that it did not represent her reflective commitments, but it would not simply undo the harm or exonerate the reaction entirely. The affect-laden reaction is, like it or not, an expression of how she feels, and precisely the sort of behavior most apt to elicit feelings of bitterness or disappointment from Dennis. It might even be true that Dennis can benefit psychologically by thinking about all the external forces (such as growing up in a prejudiced society) that could have led Stephanie to feel automatic disgust. But even if it helps Dennis to think about such mitigating factors, the presence of these factors would

not make it the case that Stephanie had nothing to apologize for, nothing for which to be forgiven. She might not be quite as blameworthy for the reaction as she would be if she reflectively endorsed it, but she is *more* blameworthy than she would be for a mere behavioral reflex, like blinking in response to a bright light. We should not conclude of her, or each other more generally, that we are all bad prejudiced people, but it is fair to conclude that she could be, in an important sense, better than she is.

Regardless where the empirical chips fall, taking seriously the possibility that individuals are (often, in certain senses, relatively) aware of their implicit biases invites us to ask a more general question insufficiently explored in philosophical discussions of responsibility. What *kind* of awareness do we have in mind when we assert that awareness is necessary for responsibility? If awareness comes in degrees, just how aware of a potential influence on our behavior must we be in order to be responsible for resisting it?

Bibliography

Block, N. 1995: On a confusion about the function of consciousness. *Behavioral and Brain Sciences* 18, 227-247.

Carpenter, S. April/May 2008: Buried Prejudice. *Scientific American Mind*, 35.

Coates, D.J. and Swenson, P. 2012: Reasons-responsiveness and degrees of responsibility. *Philosophical Studies*, doi:10.1007/s11098-012-9969-5.

Cortina, L.M. 2008: Unseen injustice: Incivility as modern discrimination in organizations. *Academy of Management Review* 33, 55-75.

Cortina, L.M., Kabat Farr, D., Leskinen, E., Huerta, M. and Magley, V.J. 2011: Selective incivility as modern discrimination in organizations: Evidence and impact. *Journal of Management*.

Dovidio, J.F., Kawakami, K., and Gaertner, S.L. 2002: Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology* 82, 62-68.

Gawronski, B., and LeBel, E.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.

Gawronski, B., Hofmann, W., and Wilbur, C.J. 2006: Are "implicit" attitudes unconscious? *Consciousness and Cognition* 15, 485-499.

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

Greenwald, A.G., and Banaji, M.R. 1995: Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102, 4-27.

Hall, D., and Payne, B. K. 2010: Unconscious attitudes, unconscious influence, and challenges to self-control. In *Self-control in Society, Mind, and Brain*, edited by Y. Trope, K. Ochsner, and R. Hassin, 221-242. New York, NY: Oxford University Press.

Holroyd, J. Forthcoming: Taking Responsibility for Bias. In Special Edition of *Journal of Social Philosophy*, edited by M. Crouch & L. Schwartzman.

Kelly, D., and Roedder, E. 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, C.M. 1997: The Normativity of Instrumental Reason. In *Ethics and Practical Reason*, edited by G. Cullity and B. Gaut, 27-68. Oxford: Clarendon Press.

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

Levy, N. 2011: Expressing Who We Are: Moral Responsibility and Awareness of our Reasons for Action. *Analytic Philosophy* 52 (4), 243-261.

Levy, N. 2012: Consciousness, Implicit Attitudes, and Moral Responsibility. *Noûs*, doi: 10.1111/j.1468-0068.2011.00853.x.

Madva, A. 2012: *The Hidden Mechanisms of Prejudice: Implicit Bias & Interpersonal Fluency*. Doctoral dissertation, Columbia University, NY.

Mele, A. 2006: *Free Will and Luck*. New York: Oxford University Press.

Nagel, T. 1970: *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.

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

Payne, B.K., Cooley, E., Lei, F. 2012 April: Who owns implicit attitudes? Testing a meta-

cognitive perspective. Presentation for *The Implicit Bias & Philosophy Workshop*. University of Sheffield, UK.

Salvatore, J., and Shelton, J.N. 2007: Cognitive costs to exposure to racial prejudice. *Psychological Science*, 18, 810-815.

Saul, J. Forthcoming: Unconscious Influences and Women in Philosophy. In *Women in Philosophy: What Needs to Change?*, edited by F. Jenkins and K. Hutchison.

Schlenker, B.R., and Darby, B.W. 1981: The use of apologies in social predicaments. *Social Psychology Quarterly* 4, 271-278.

Sims, C., Drasgow, F., and Fitzgerald, L. 2005: The effects of sexual harassment on turnover in the military: time-dependent modeling. *Journal of Applied Psychology* 90, 1141-1152.

Strawson, P.F. 1974: *Freedom and resentment and other essays*. London: Methuen.

Sue, D.W., Capodilupo, C.M., Torino, G.C., Bucceri, J.M., Holder, A.M.B., Nadal, K.L., and Esquilin, M. 2007: Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist* 62, 271-286.

Sullivan, S. 2006: *Revealing Whiteness: The Unconscious Habits of Racial Privilege*. Bloomington, IN: Indiana University Press.

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