Rev.Phil.Psych. (2016) 7:765–786
DOI 10.1007/s13164-015-0287-7

Attributionism and Moral Responsibility for Implicit Bias

Michael Brownstein[1]

Published online: 8 October 2015

Abstract Implicit intergroup biases have been shown to impact social behavior in many unsettling ways, from disparities in decisions to "shoot" black and white men in a computer simulation to unequal gender-based evaluations of résumés and CVs. It is a difficult question whether, and in what way, agents are responsible for behaviors affected by implicit biases. I argue that in paradigmatic cases agents are responsible for these behaviors in the sense that the behavior is "attributable" to them. That is, behaviors affected by implicit biases reflect upon who one is as a moral agent.

1 Introduction

There is an important and multifaceted connection between psychological research on implicit bias and philosophical research on moral responsibility. One facet of this connection is that implicit bias is simply very morally weighty. It is well-established in the empirical literature that implicit biases contribute to many kinds of discriminatory behavior.[1] It has also been shown that apparently minor acts—such as interrupting women more than men or giving slightly better scores on papers to white students than black students—can add up to very significant patterns of discrimination.[2] Implicit bias *matters*, and anyone concerned with fairness and justice should want to know who's responsible for it.[3]

[1] For review see Greenwald et al. (2009), Jost et al. (2009), and Nosek et al. (2007a). For a critique of measures of implicit bias, see Oswald et al. (2013); see Greenwald et al. (2015) for reply. For a more detailed introduction to implicit bias, see Brownstein (2015).

[2] See Valian (2004) on the "accumulation of advantage." See also Greenwald et al. (2015).

[3] One possibility is that no one in particular—no individual—is responsible for implicit bias. By this I mean that implicit bias is an effect of legacies of historical inequality, patterns of residential and occupational segregation, discriminatory laws and political policies, and so on. There are multiple ways to render this "institutional" approach to understanding implicit bias. For examples, see Anderson (2010) and Haslanger (2015). I am sympathetic to these approaches, although I think there are significant reasons to examine implicit bias in terms of individuals too, at least alongside social-institutional considerations. For a response to the institutional critique of research on implicit bias, see Madva (ms a). One aim of this paper is to show why the question of individual moral responsibility for implicit bias matters.

✉ Michael Brownstein
msbrownstein@gmail.com

[1] John Jay College/CUNY, 180 Carlton Avenue, Brooklyn, NY 11205, USA

Another facet of the connection between implicit bias and moral responsibility has to do with the role implicit attitudes more generally play in day-to-day life. "Implicit bias" is a term of art referring to implicit attitudes directed toward individuals in virtue of their social group membership. Implicit attitudes themselves are evaluative states—"likings" or "dislikings" in the empirical literature—that can be directed toward anything, including consumer products, self-esteem, food, alcohol, political values, and so on. Implicit attitudes, then, are hugely pervasive in daily life, arguably affecting many, if not all, of our opinions, judgments, and decisions.[4] Surely we should want to know whether, and in what sense, we are responsible for these.

A third facet of the connection between implicit bias and moral responsibility has to do with the peculiar psychological structure of implicit attitudes. In short, what we know about implicit attitudes suggests they do not easily fit into traditional philosophical approaches to theorizing about moral responsibility. For example, in the empirical literature, "implicit" typically means outside of conscious awareness or control. [5] Philosophers who think that moral responsibility hinges on agentive control over, or consciousness of, one's attitudes may take this usage to suggest that people are not responsible for their implicit biases. But the data are complex, showing that we do have some degree of control over our implicit biases and that we are in fact aware of them in some sense (see §3). Philosophers who think instead about moral responsibility in terms of the reasons-responsiveness of an agent's action-guiding psychological states, or in terms of identification and alienation, may similarly be stymied by ambiguous data. In general, implicit biases seem to live at the margin between ordinary cognitive states like belief and mere psychological forces.[6] Implicit bias is therefore a good test case for theories of moral responsibility that aim to accommodate the messy reality revealed by the contemporary sciences of the mind.

There is a small but growing philosophical literature focused on responsibility and implicit bias.[7] Few authors, however, have approached the question from the perspective of "attributionist" theories of moral responsibility.[8] In general, for an action to be attributable to an agent is for it to "reflect upon" the agent "herself." When this is the case, the agent becomes open to evaluations that target her moral character, evaluations like "kind," "selfish," and so on. A common way of speaking is that these "aretaic"

[4] Tamar Gendler's (2008a,b) influential account of implicit attitudes as "aliefs" argues that these mental states are responsible for the management of much of "moment-to-moment" behavior.

[5] See, for example, Hardin and Banaji (2013).

[6] As Levy (2014) and Mallon (forthcoming) put it, terms of moral assessment are connected to folk psychological concepts like "belief." So one reason the question of moral responsibility for implicit bias is important is because of the way in which implicit biases are, and are not, belief-like. Levy (2014) argues that implicit biases are neither beliefs nor mere associations; because of this, neither blame nor excuse for them is appropriate. The view I develop here can be understood as (mostly) accepting Levy's account of the psychological structure of implicit attitudes, but rejecting the claim that neither blame nor excuse is appropriate. For discussion of whether implicit biases are beliefs, see Brownstein (2015), Gendler (2008a,b), Levy (2012, 2014), Madva (2012, forthcoming), Mandelbaum (2015), and Schwitzgebel (2010).

[7] See Faucher (forthcoming), Glasgow (forthcoming); Holroyd (2012); Kelly and Roedder (2008); Levy (2012, 2014, forthcoming); Madva (2012, ms b); Saul (2013); Sie and Vorst Vader-Bours (forthcoming); Washington and Kelly (forthcoming); and Zheng (forthcoming).

[8] To my knowledge, those few are Zheng (forthcoming), Glasgow (forthcoming), and, briefly, Smith (2012). Faucher (forthcoming) discusses related issues. For key accounts of attributionism, see Arpaly (2004), Frankfurt (1971), Hieronymi (2008), Jaworska (2007), Scanlon (1998), Sher (2009), Shoemaker (2003), Smith (2005, 2008, 2012), Sripada (2010, 2015), and Watson (1975, 1996).

evaluations are appropriate in virtue of an action reflecting upon the agent's "real" or "deep" self. In general terms, the real or deep self is a functional concept representing an agent's stable and identity-grounding attitudes. Historically, there is reason to think that attributionism in its contemporary form traces back at least to Hume, who distinguished between something like an agent's deep and superficial psychological attitudes.[9] Contemporary attributionist theories use this distinction to make sense of cases in which it seems appropriate to hold a person morally responsible for actions that are non-conscious (e.g., "failure to notice" cases), non-voluntarily (e.g., actions stemming from strong emotional reactions), or otherwise divergent from an agent's will. For example, it seems intuitive to say that I would not be morally responsible for inadvertently stepping on a stranger's toe on a crowded subway. Doing so might be bad, but nothing about *me* is open for evaluation as a result. However, if I step on someone's toes because I am aggressively pushing my way toward my favorite spot near the window, then there is something about me—something about who I am as a moral agent—that is expressed through my behavior, and *I* now appear to be open to evaluation for this act. This is to say that the action appears to be attributable to me, even if a number of putatively "exculpating" conditions obtain. I might not know that I have stepped on anyone's toes, and might not have intended to do so, and I might even have tried hard to avoid everyone's toes while I raced to my favorite spot. Regardless whether my action is non-conscious and non-volitional in this way, or whether I disavow "New York Style" subway riding, what I've done expresses something morally important about me. I'm not just klutzy, which is a kind of "shallow" or "grading" evaluation (Smart 1961). Rather, a bystander would be quite right to think "what a jerk!" as I push by.

The attributionism literature contains many case studies more vivid than this, and to a large extent I am going to presume without argument that agents can, in principle, be thought of as responsible for actions that reflect upon their deep selves, even if those actions are non-conscious, non-volitional, and "non-tracing" (i.e., the agent's responsibility does not trace back to some previous action or decision (Fischer and Ravizza 1998)). I will say, however, that one reason attributionism is appealing is that it is well-suited for making sense of moral responsibility in light of the changing conception of the human mind found in contemporary science. Cognitive and social psychology, behavioral economics, neuroscience, etc. are coalescing around a picture of the mind as "boundedly rational"—driven more than we used to think by affect, non-conscious processes, associative learning, and so on—and a good theory of moral responsibility should be well-suited to these findings.[10] Moreover, a good theory of moral responsibility should be relatively consistent with—or make sense of—common folk attitudes toward responsibility, since the concept of responsibility is itself a deeply social one. And there is indeed evidence that folk attributions of moral responsibility are

[9] Sripada (2010) quotes from Hume's *Treatise*: "Actions are by their very nature temporary and perishing; and where they proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honour, if good, nor infamy, if evil, the action itself may be blameable; it may be contrary to all the rules of morality and religion: But the person is not responsible for it: and as it proceeded from nothing in him, that is durable or constant, and leaves nothing of that nature behind it, 'tis impossible he can, upon its account, become the object of punishment or vengeance" (Treatise, bk. 11, Pt. 111, sec. 2).

[10] I entirely sidestep the question of whether this emerging picture of the mind threatens free will.

influenced by judgments about whether others' actions reflect something like their deep selves (e.g., Sripada 2010; Newman et al. 2014; Strohminger and Nichols 2014). Thus while I do not expect this essay to undo anyone's antecedently-held strong skepticism toward attributionism, I do hope to make the case for attributionism stronger by showing how it helps to make sense of moral responsibility in light of the emerging scientific picture of the mind. My aim is not to build an original case for attributionism, in other words, but rather to add to its appeal by putting it into practice for one particularly challenging and important case—the case of implicit bias.

That said, there are significant unresolved questions about attributionism, in particular what constitutes the deep self and what it means for an action to reflect upon the deep self.[11] If successful, what follows will help to resolve both of these questions, as well as clarify whether people are responsible for their behaviors that express implicit bias. First I make a few brief stage-setting remarks (§2). Then I introduce the empirical literature on implicit bias in a bit more depth, emphasizing key features of implicit attitudes that are relevant to moral responsibility (§3). After this, I present a schematic conception of the deep self based on an agent's "cares" (§4), and then discuss what it means for an action to reflect upon one's deep self (§5).[12] §1-§5 comprise the bulk of the paper, with my conclusion unfolding somewhat quickly. The conclusion is that behaviors affected by implicit bias can, and in paradigmatic cases do, reflect upon our cares, such that those behaviors are attributable to us (§6). This means that behaviors affected by implicit bias redound on one's standing as a moral agent. I respond to potential objections to this claim (§7) and then make some concluding remarks about moral responsibility in a broader sense (§8).

2 Terminology, etc

Five brief points will help to clarify what follows. First, while I have spoken about responsibility for implicit bias in a general sense, my focus will be on moral responsibility for *behaviors* affected by implicit biases, or what I will call behavioral expression of implicit bias (BEIB). That said, much will hang on the psychological structure of implicit attitudes. But my concern will not be with our responsibility for having these attitudes; rather, I will be concerned with what the structure of these attitudes tells us about responsibility for the behaviors in which they are implicated.[13]

Second, the term "attitudes" is used differently in philosophy and psychology. In psychology, attitudes are understood as likings or dislikings; or, more formally, as associations between a concept and an evaluation (Nosek and Banaji 2009). Philosophical usage tends to be more expansive, treating beliefs, desires, intentions,

[11] I am indebted to Chandra Sripada (2015) for this way of dividing up the central questions facing attributionism.

[12] As will be clear, this is a different sense of caring than the more familiar one developed by Frankfurt (1988).

[13] I focus on responsibility for behavior rather than for attitudes simply because I find the former to be more tractable. In future work I hope to consider the question of responsibility for implicit bias itself. For discussion of responsibility for attitudes, see Smith (2005, 2008, 2012). Also see Zheng (forthcoming) for discussion of the conditions under which an implicit bias is implicated in behavior, as it relates to questions about moral responsibility.

imaginations, and more as attitudes. Unless otherwise indicated, hereafter I'll discuss attitudes in the psychological sense.

Third, while what I discuss bears upon the appropriateness of judging others and ourselves for our implicit biases, nothing I say should be taken to imply that implicit bias is tantamount to racism, sexism, etc., in any of the generic uses of these terms. Implicit bias is a form of prejudice, but I make no claim about whether it amounts to full-blown racism, etc.

Fourth, unless otherwise noted, I will use the terms "action" and "behavior" interchangeably.

Finally, a note of caution about the term the "deep self." In the experimental literature, the deep self is usually defined as "the person you really are, deep down inside" (Newman et al. 2014). This seems to imply that the deep self is singular, harmonious, and definitional for agents.[14] It implies, in other words, that each of us has one core deep self, which is internally consistent, and which fundamentally defines who we are "at bottom". This is *not* what I mean when I speak of the deep self. As I use the term, as I said above, the deep self is a functional concept representing an agent's stable and identity-grounding attitudes. The deep self is that to which actions that are attributable to us are attributed. When an action reflects upon a person's deep self in this sense, the person is thereby open, in principle, to aretaic evaluations (from others or even from and toward oneself). This functional conception of the deep self has two immediate implications. The first is that one's deep self may be comprised of multiple sources (e.g., habits, desires, beliefs, etc.), which may be in conflict with one another, and may or may not have the "final say" about who one is. I say more about this in §7.1. The second is that while attributability reflects one sense of what it is to be morally responsible for some action—namely that one is open to aretaic evaluation—it leaves open questions about other important senses in which an agent may be morally responsible for an action. I briefly discuss the relationship between attributability and what Shoemaker (2011) calls "answerability" and "accountability" in §8.

3 Implicitness as Arationality

What makes the question of responsibility for BEIB distinct from questions about responsibility for other kinds of behavior is not, perhaps surprisingly, that implicit attitudes are unconscious or automatic. Despite that this is by far the commonest way philosophers, and some psychologists (as I noted above), characterize implicit attitudes,[15] the empirical literature is quite mixed with respect to whether people are aware of their implicit attitudes and whether they can control them. I have reviewed this

[14] Perhaps deep down one can be fundamentally conflicted, as I briefly discuss in §7.1, but this possibility is not usually reflected in the experimental literature on the deep self. Thanks to Alex Madva and Susanna Siegel for pushing me to clarify this. An important question for future research is whether and how the account I develop here integrates with experimental approaches to folk conceptions of the deep self.

[15] In their influential 2008 paper, for example, Kelly and Roedder write, "the IAT requires subjects to make snap judgments that must be made quickly, and thus without moderating influence of introspection and deliberation and often without conscious intention. Biases revealed by an IAT are often thought to implicate relatively automatic processes" (525). Jennifer Saul (2012, 244) describes implicit biases as "unconscious tendencies to automatically associate concepts with one another." And elsewhere I have called implicit biases "relatively unconscious and relatively automatic features of prejudiced judgment and social behavior" (Brownstein 2015).

literature elsewhere (Brownstein forthcoming). At present, it suggests that people are often aware of the content of their implicit attitudes, largely in the form of "gut feelings," but are often unaware of the effects their implicit attitudes have on their behavior. Crucially, this is also the case with explicit attitudes. We generally know what our explicit preferences are, but often don't know how they end up affecting our behavior. The likeness between implicit and explicit attitudes with respect to control is similar. We can typically exercise "long range" control (Holroyd 2012) over our implicit attitudes but can't control them "directly" through sheer force of will. The same holds with explicit attitudes. We can shape them through habituation and practice, but (usually) can't simply will ourselves into explicitly liking or disliking something.

A better characterization focuses on the arationality of implicit mental states (Brownstein forthcoming). Implicit attitudes are distinct from explicit attitudes because they are largely insensitive to what we explicitly take to be true or good.[16] It is important to note that this does not prohibit implicit attitudes from being consistent with what we explicitly take to be true or good. In cases of topics of relatively low social sensitivity, like brand preferences as compared to racial preferences, implicit and explicit attitudes are typically correlated (Nosek et al. 2007b). But when implicit and explicit attitudes converge, they do so via different routes. The psychological model that best captures the difference between these "routes" is Bertram Gawronski and Galen Bodenhausen's "Associative-Propositional" model of evaluation (APE; Gawronski and Bodenhausen 2006, 2011). APE treats implicit and explicit attitudes as behavioral manifestations of two distinct kinds of mental process.[17] According to APE, information is stored in the mind in the form of associations. For example, the statement, "black people are a disadvantaged group" represents the association between "black people" and "disadvantaged group" (Gawronski and Bodenhausen 2011). When we encounter relevant cues, the associations stored in our memories become activated. Hearing the name "Malcolm X," for example, might activate the thought that black people are a disadvantaged group.[18] APE refers to this process of the activation of associations as *associative processing*. Sometimes, however, we are concerned to validate the information supplied by associative processing. That is, sometimes we are concerned with whether a given association is true or false. APE refers to this process of validation as *propositional processing*. The result of propositional processing might be the thought that "it is true (false) that black people are a disadvantaged group."[19] Thus the fundamental difference between

[16] An immediate objection stems from the fact that a good number of our explicit attitudes—in the philosophical sense of attitudes—fail to be sensitive to what we think to be true or good. The large literature on belief perseverance is one testament to this (see Anderson and Lindsay (1998) for review). My view—discussed in more depth elsewhere (Brownstein, M. Manuscript, The Implicit Mind (Unpublished))—is that explicit attitudes like beliefs have the possibility of being irrational because part of their function is to respect certain basic rules of rationality (e.g., to be sensitive to logical operators like negation, to be "inferentially promiscuous," and so). As I suggest below, implicit attitudes don't have this function. They don't play the game of rationality. Thus they are arational, not irrational. Thanks to Susanna Siegel for pushing me to clarify this point.

[17] The rest of this paragraph, as well as the paragraph following, are adapted from Brownstein (forthcoming).

[18] Of course, a person will have many associations with the name Malcolm X, just as they will with virtually any cue. APE offers a complex account of which associations will be activated in a given context. This account is largely in keeping with connectionist models of cognition.

[19] A note on potential terminological confusion: APE focuses on what it calls propositional *processes*, not propositional states (i.e., not mental states with propositional structure, the kind with which philosophers of mind are typically concerned). For more, see Brownstein (2015).

associative and propositional processes, according to APE, is that propositional processes alone reflect an agent's subjective assessment of truth.

APE is put to work to distinguish implicit and explicit attitudes in the following way. When a person reads through a pile of résumés (for example), she may notice (consciously or unconsciously) the names of the job candidates. These names will trigger associations with particular social groups (e.g., Jamal may trigger associations with black men; Emily may trigger associations with white women, etc.). In addition, people often associate positive and negative stereotypes with particular social groups. For example, many white Americans associate negative stereotypes such as "lazy" and positive stereotypes such as "athletic" with black men.[20] Upon registering the name "Jamal," these stereotypes may become activated. Because this is an associative process, the name Jamal may activate the concept lazy independently of whether the person believes it to be true or false that people with the name Jamal tend to be lazy. Such activated associations may enter into the résumé reader's conscious awareness as a vague negative gut feeling, although this emergence into consciousness is not a defining feature. What *is* crucial, according to APE, are the ways in which an activated association gives rise to behavior. One possibility is that associative processing alone "guides" the résumé reader's response. This is the situation manufactured by the IAT and other indirect measures. A second possibility is that the association is transformed into a proposition (e.g., "black people are lazy") which the agent then endorses or rejects. This is the situation manufactured by questionnaires and other direct measures of attitudes.

A number of philosophers have endorsed the characterization of implicitness in terms of arationality. For example, Tamar Szabó Gendler's (2008a,b) account of implicit attitudes as "aliefs" counts among the core characterizations of these states that they are arational.[21] Alex Madva (forthcoming) relatedly develops a view of implicitness in terms of a mental state's insensitivity to logical operations like negation (i.e., treating *P* and not-*P* as the same). Neil Levy (2014) argues for a related view, albeit in somewhat different terms. Gendler, Madva, and Levy all put these points into the service of showing why implicit attitudes are not beliefs.[22]

All of this might incline the theorist of moral responsibility to think that people shouldn't be held responsible for their BEIB, at least in paradigmatic cases. If the relevant behaviors stem from arational states, then one might think the agent can't be responsible for the behavior, in any meaningful sense of "responsible." An account of attributability based on one's cares suggests otherwise.

[20] But note that stereotypes such as "athletic" can be positive in some contexts but negative in others. "Athletic" is often associated with "unintelligent," for example. On the relationship between implicit stereotypes and evaluation, see Amodio and Devine (2006), Holroyd and Sweetman (forthcoming), and Madva and Brownstein (ms).

[21] Notably, Bodenhausen and Gawronski (2014, 957) write that the "distinction between associative and propositional evaluations is analogous to the distinction between 'alief' and belief in recent philosophy of epistemology."

[22] Although note that Levy (2014) does not endorse an associative picture of implicit attitudes. See Mandelbaum (2015) for a contrasting view.

4 Caring and the Deep Self

Consider the provocative first sentences of a recent article on *Slate* ("Heel," 30 May 2013):

> I'm a stay-at-home dad to twin 4-year-old girls who are already smarter than me, and my wife is a brilliant doctor who kicks ass and saves lives every day. I grew up with big sisters and a mom whose authority was unbreachable. I celebrate every inroad that women make into business, technology, science, politics, comedy, you name it, and I get angry about "slut-shaming" or "stereotype threat" or whatever is the affront du jour. And yet, in the caveman recesses of my imagination, I objectify women in ways that make Hooters look like a breakout session at a NOW conference.

What is the right assessment of the author's deplorable pattern of awareness? Is it reflective of the author himself or is it rather something he does, but only in virtue of his living in a social world suffused with sexualized images of women?[23] In virtue of what can we answer this question?

In order to answer this kind of question, attributionists have focused on varying criteria, including hierarchies of agents' desires, agents' values, and agents' evaluative judgments.[24] I will focus instead on an agent's cares.[25] The most well-known philosophical account of caring is Harry Frankfurt's (1988). But in recent years a different kind of account of caring—one not focused on an agent's strongest desires or volitional necessities—has surfaced. As in Frankfurt's work, the upshot of this alternative view is that cares are inherently internal to agents. Cares are the source of the deep self, in other words. But this alternative view of caring stresses three distinct features of cares: (1) a distinction between "ontological" and "psychological" senses of caring; (2) a tight link between caring and emotion; and (3) a particular dispositional profile of cares.[26]

In the psychological sense, what one cares about are the things, people, places, ideas, and so on that one perceives oneself as valuing. Cares in the psychological sense track the agent's own perspective. In the ontological sense, cares are, by definition, just those attitudes (in the philosophical sense) that belong to an agent, in contrast to the rest of the "sea of happenings" in her psychic life (Jaworska 2007, 531; Sripada 2015). One can easily be wrong about one's cares in the ontological sense, and one can discover what one cares about too (sometimes to one's own surprise). The care-based view of the deep self is concerned with cares in the ontological sense. Hereafter, when I discuss cares, I do so in this sense.

[23] Indulge me in a brief bit of textual interpretation. The fact that the author says he gets angry about slut-shaming, stereotype threat, or "whatever is the affront du jour" strikes me as dismissive of gender prejudice, as if slut-shaming and stereotype threat are passing fads. This seems to be a nice example of unintended bias being expressed in writing. This is striking, since it's in an essay *about unintended biases*.

[24] See, respectively, Frankfurt (1971), Watson (1975), and Scanlon (1998).

[25] In the same vein as what I said in §1 with respect to attributionism in general, my central aim is not to promote care-based approaches as such. So I will not offer arguments for the superiority of the care-based approach to these other attributionist theories. I do hope, however, to strengthen the case for a care-based account by showing how it can make sense of responsibility for BEIB.

[26] Different theorists place emphasis on these features in different ways. I draw largely upon Jaworska (1999, 2007), Shoemaker (2003, 2013), and Sripada (2010, 2015), although the synthesis I present in this section is my own.

If not by reference to what an agent takes herself to care about (or what she takes herself to identify with), then by reference to what do we pick out an agent's cares, in the ontological sense? Some stress the deep link between caring and feeling. When we care about something, we feel "with" it. We are emotionally tethered to it (Shoemaker 2003, 94). To care about your first-born, your St. Bernard, the fate of Ft. Greene Park, or the history of urban beekeeping is to be psychologically open to the fortunes of these people, animals, places, narratives, and so on. Another way to put this is that caring is a way for something to *matter* to you (Shoemaker 2003, 95; Sripada 2015). Mattering in this sense is dispositional. For example, I can be angry about the destruction of the Amazon Rainforest—in the sense that its fate matters to me—without experiencing the physiological correlates of anger at any particular moment. But to care about the Amazon involves the disposition to feel certain things at certain moments, like anger when you are reminded of the Amazon's ongoing destruction.

What matters to us is inherently motivational. In particular, things that matter to us motivate a broad suite of feelings, judgments, and behaviors, which manifest over time and across situations.[27] Emotions are tightly tied to cares precisely because of their dispositional profile, which is cross-situational, durable, and multiform (i.e., emotions are expressed through many channels, such as feelings, judgments, behavior, etc.). For this reason, emotions (in the dispositional sense) can underwrite identity and psychological continuity (Jaworska 2007, 549). One reason they can do so is because emotional dispositions are constituted by webs of referential connections, in Michael Bratman's (2000) sense. Jaworska explains this idea using the example of grief, which involves painful thoughts, a tendency to imagine counterfactuals, disturbed sleep, and so on, all of which point referentially to the person or thing for which one grieves. "In this way," Jaworska writes (2007, 553), "emotions are constituted by conceptual connections, a kind of conceptual convergence, linking disparate elements of a person's psychology occurring at different points in the history of her mental life."

While it is necessary to comprehend the importance of something in order to care about it, it is not necessary to comprehend anything consciously, nor is it necessary that one's cares are consistent with one's evaluative judgments.[28] While cares and explicit judgments often correlate, it is clear, I think, that caring about something does not require judging the thing to be true or good, nor does judging something true or good require caring about it (Shoemaker 2003, 96; Jaworska 2007, 562, fn 94). I can believe in compatibilism without particularly caring about it, and I can care about what I wear to work without believing that what I wear to work matters very much.[29]

This synthesis of the care-based view of the deep self is necessarily brief and schematic. If plausible, however, it should lead directly to a second pressing question: what does it mean for an action to reflect upon one's cares, and thus upon the deep self?

[27] Sripada's (2015) closely related view is that "cares involve a complex syndrome of motivational, commitmental, evaluative, and affective dispositions."

[28] For a dissenting view on the role of consciousness in caring, see Levy (2011).

[29] A slightly stronger claim than that cares and explicit judgments are typically correlated is that caring about something disposes an agent to judge the thing to be good (Sripada 2015). Also, there is another potential point of disagreement between theorists about cares having to do with just how fundamental cares are. One possibility is that cares, alongside evaluative judgments, are inherently internal (Jaworska 2007, 559, fn 88). Another possibility is that cares are the source of all states that are internal in the relevant sense (Shoemaker 2003).

5 Reflection

As Levy (2011) puts it in a challenge to deep self theories of responsibility, it is important to be able to spell out the conditions under which a care *causes* an action, and "in the right sort of way."[30] The right sort of causality should be "direct and nonaccidental," and must also be a form of mental causality, as others have stressed (Scanlon 1998; Smith 2005, 2008; Sripada 2015). These conditions are to rule out deviant causal chains and the like, but also to distinguish actions that reflect upon agents from just actions as such. For an action to say something about me requires more than it simply being an action that I perform (Shoemaker 2003; Levy 2011). Answering my ringing cell phone is an action, but it doesn't *ipso facto* express anything about me. In the terms I have suggested, this is because answering the phone is not something I particularly care about. My cares are not the effective cause of my action, in this case. (I will return to this example below.)

A causal connection between a care and an action is a necessary condition for the latter to reflect the former, but it doesn't appear to be sufficient. To illustrate this, consider Jack, who volunteers at a soup kitchen every weekend.[31] Jack genuinely cares about helping people in need, and his caring about helping people in need causes him to volunteer. At the soup kitchen, Jack meets Jill, whom he likes and wants to impress. The following week, Jack goes to the soup kitchen, not because he is moved to help people in need, but rather to impress Jill. The question is: on which of Jack's cares does his action reflect? Does it reflect his caring about people in need or his caring about impressing Jill? The answer seems to be that his volunteering this particular time reflects his caring about impressing Jill. But both of his cares helped to cause him to volunteer this particular time. His caring about those in need caused him to volunteer earlier, which caused him to meet Jill, which caused him to volunteer this particular time. And his caring about impressing Jill also caused him to volunteer this particular time. The broad problem is that agents have cares that are causally connected to their actions yet aren't reflected in those actions.

How can it be discerned when an action is both caused by one's cares and reflects upon those cares?[32] Considering Jack's actions in a broader context in order to identify patterns of caring and concern can help. Does he continue to go to the soup kitchen on days when Jill isn't going to be there? Has he acted on charity-directed cares in other contexts? Answering these kinds of questions helps

[30] Most researchers writing on attributionism accept that a causal connection between an agent's identity-grounding states and her attitudes/actions is a necessary condition for reflection (e.g., Levy 2011; Scanlon 1998; Sripada 2015). Note also two points of clarification on Levy's (2011) terminology. First, he identifies a set of propositional attitudes that play the same role in his discussion as what I have been calling cares. That is, he simply refers to an agent's identity-grounding states as "attitudes." Second, Levy distinguishes between an action "expressing," "reflecting," and "matching" at attitude (or, as I would put it, an action or attitude expressing, reflecting, and matching an agent's cares). The crucial one of these relations is expression. It is analogous to what I will call "reflection."

[31] I am grateful to Sripada (p.c.) for this illustration.

[32] To answer this question, Sripada (2015) proposes the "motivational support account of expression." Roughly, an action is said to reflect upon one's deep self, on this view, if and only if the motive expressed in the action is one of the agent's cares. A motive is said to be expressed in an action if and only if the motive exerts influence of sufficient strength on the agent's "action-directing psychological mechanisms."

to illuminate *which* of an agent's causally effective cares are reflected in a given action. It can also help to illuminate *whether* an agent's cares are effective. Consider again the case of answering the cell phone. Imagine a couple in the midst of a fight. John says to work-obsessed Larry, "I feel like you don't pay attention to me when I'm talking." Larry starts to reply but is interrupted by his ringing phone, which he starts to answer while saying to John, "hang on, I need to take this." "See!," says John, exasperated. In this case, answering the phone does seem to reflect upon Larry's cares. His cares are causally effective in this case, in a way in which they aren't in the normal answering-the-phone case. We infer this causal effectiveness from the pattern toward which John's exasperation points.

Patterns of actions that are most relevant to making inferences about an agent's cares are multitrack. I use this term in the way that theorists of virtue use it, to describe patterns of action that are durable and arise in a variety of contexts, and are also connected to a suite of thoughts, feelings, and so on (e.g., Hursthouse 1999). Both Jack's and Larry's actions appear to reflect their cares because we infer counterfactually that their Jill-directed and work-directed cares would manifest in past and future thoughts and actions across a variety of situations. Jack might also go to see horror movies if Jill did; Larry might desire to work on the weekends; and so on. Patterns in this sense make future actions predictable and past actions intelligible. They do so because their manifestations are diverse yet semantically related. That is, similar cares show up in one's beliefs, imaginations, hopes, patterns of perception and attention, and so on. Another way to put all of this, of course, is that multitrack patterns indicate dispositions, and inferences to dispositions help to justify attributions of actions to what a person cares about.[33]

6 Implicit Bias and the Deep Self

Are cares reflected in paradigmatic BEIB? Which cares? And what reason do we have to think that BEIB really reflect those cares, rather than stand as some kind of unreflecting byproduct of them?

[33] Levy (2011, 252–253) makes a similar point in the case of omissions: "patterns of lapses are good evidence about agents' attitudes for reasons to do with the nature of probability. From the fact that there is, say, a 50 % chance per hour of my recalling that it is my friend's birthday if it is true that I care for him or her and if internal and external conditions are suitable for my recalling the information (if I am not tired, stressed, or distracted; my environment is such that I am likely to encounter cues that prompt me to think of my friend and of his or her birthday, and so on), and the fact that I failed to recall her birthday over, say, a 6-h stretch, we can conclude that one of the following is the case: either I failed during that stretch to care for him or her *or* my environment was not conductive to my thinking of him or her *or* I was tired, stressed, distracted, or what have you, *or* I was unlucky. But when the stretch of time is much longer, the probability that I failed to encounter relevant cues is much lower; if it is reasonable to think that during that stretch there were extended periods of time in which I was in a fit state to recall the information, then I would have had to have been *much* unluckier to have failed to recall the information if I genuinely cared for my friend (Levy 2011). The longer the period of time, and the more conducive the internal and external environment, the lower the probability that my failure to recall is a product of my bad luck rather than of my failure to care sufficiently. This is part of the reason why ordinary people care whether an action is out of character for an agent: character, as manifested in patterns of response over time, is good evidence of the agent's evaluative commitments in a way that a single lapse cannot be."

In §4 I stressed three features of cares: (1) a distinction between ontological and psychological senses of caring; (2) a tight link between caring and emotion; and (3) a particular dispositional profile of cares. And in §5 I suggested that actions reflect cares when they are caused by those cares in the right—nonaccidental and mental-causal—way. This kind of causality can be inferred from agents' multitrack patterns of thought and action.

Consider two paradigmatic cases of BEIB. The first is "shooter bias." In a computer simulation that shows subjects a series of pictures of black and white men holding guns or harmless objects like cell phones, in which the goal is to shoot all and only those men shown holding guns, most white subjects are more likely to shoot unarmed black men than unarmed white men and to fail to shoot armed white men compared to armed black men (Correll et al. 2002; Glaser and Knowles 2008). While the data have been mixed, a meta-analysis of shooter bias studies ominously finds that police officers tend to fare no better than average participants in terms of unbiased performance (Mekawi and Bresin 2015).

Participants' shooting decisions appear to reflect a care about the purported violent tendencies of black men. It is crucial to remember that this is meant in the ontological, not psychological, sense. The agent need not recognize the care as her own. But the purported violent tendencies of black men *do* matter to agents in this context, as is manifest in study participants' emotional and dispositional profiles. Emotionally, the shooter bias test elicits fear. Not just generic or undifferentiated fear, moreover, but fear that is specifically linked to the perception of black faces. This is suggested in various ways. One is that shooter bias appears to be linked to perceiving black targets as threatening. Threat detection is known to enhance neurophysiological response in what's called the P200 component of "event-related potentials" (ERPs), a measure of fluctuations in electrical activity in the brain. Ito and Urland (2005) show that P200 response is more pronounced when participants are presented with black faces compared with white faces, and Correll et al. (2006) show that the degree of differentiation in P200 response between presentation of black and white faces predicts participants' magnitude of shooter bias. The linkage between shooter bias and fear is also made clear in consideration of the interventions that do and do not affect task performance. For example, participants who adopt the conditional plan, "whenever I see a Black face on the screen, I will think 'accurate!'" do no better than controls at being unbiased. However, participants who adopt the plan, "whenever I see a Black face on the screen, I will think 'safe!'" demonstrate significantly less shooting bias (Stewart and Payne 2008). While the emotion that shooter bias elicits is specific, the care itself has a broad dispositional profile. For example, shooter bias is correlated with "implicit motivation to control prejudice," which is defined as the interaction of one's implicit attitudes toward prejudice and one's implicit associations between oneself and prejudice. People who have weak associations between "prejudice" and "bad" and strong associations between "prejudice" and "self" tend to display the most biased simulated shooting behavior (Glaser and Knowles 2008). This suggests a link between states of fear, motivation, and value, which coalesce around what we can call a care.

Shooting behavior on the computer simulation also appears to reflect the agent's cares in the right way. The fact that tests of black-violence implicit associations predict biased responses on the shooter bias test (Glaser and Knowles 2008) suggests that the agent's behavior is indeed caused by those attitudes that reflect what she cares about.

These behavioral predictions are buttressed by studies in which manipulations of the black-violent implicit association lead to changes in shooting behavior. Correll et al. (2007) show that participants who read a newspaper story about a black criminal before taking the shooter bias test demonstrate more anti-black bias than participants who read a newspaper story about a white criminal. Moreover, the multitrack patterns of behavior that tests of black-violence associations appear to predict rule out the possibility that participants' behavior is caused by their cares but that their behavior doesn't reflect their cares (as in the Jack and Jill case in §5). Tests such as the IAT have moderate test-retest reliability (Nosek et al. 2007b), which demonstrates that the relevant associations are relatively durable and not simply reflections of the testing context. Also, caring about the purported violent tendencies of black men doesn't manifest in shooting behavior alone, but gives rise to a pattern of related results, such as ambiguous word and face detection (Eberhardt et al. 2004), social exclusion (Rudman and Ashmore 2007), and the allocation of attention (Donders et al. 2008). These patterns may be hard to notice in daily life, since social norms put pressure on people to behave in unprejudiced ways, and most people are fairly good at self-regulating their implicit attitudes. This is precisely the reason controlled experiments are needed. They have the unique ability to create the conditions under which multitrack patterns of behavior emerge.

A second paradigmatic example of implicit bias yields similar results. A long list of studies demonstrates gender bias in reviews of CVs, résumés, and job application materials (e.g., Bertrand and Mullainathan 2003; Moss-Racusin et al. 2012). Many of these studies suggest that participants in them care about the purported intellectual superiority—and related traits such as agentiveness and competence—of men compared to women. Again, the discordance of this care with agents' own explicit beliefs or judgments does not speak against its status as a care in the ontological sense. In virtue of the specific emotions and specific behaviors the relevant stimuli elicit, it seems right to say that the comparative intellectual status of men vs. women *does* matter to participants in these studies. The link between the agent's cares and specific emotions is more complex in this case (compared with the shooter bias case), since it may appear that gender-intelligence associations are more "coldly cognitive" than black-violence associations (Anderson 2010; Valian 2005). But emotion actually plays a crucial role in these cares too, as is evident in research on "halo" and "compensation" effects (e.g., Fiske et al. 2002) and "benevolent sexism" (e.g., Dardenne et al. 2007). Broadly, this research shows that people commonly compensate for negative stereotypes about particular groups with positive feelings toward them. Compensation effects have been found not only in studies of explicit attitudes but in studies of implicit intergroup attitudes as well (Carlsson and Björklund 2010). These findings suggest that even the most coldly cognitive stereotypes engender positive and negative affect.[34] This form of affect may be "low-level" (see §7.3), but along with a broader dispositional profile it seems to demonstrate that gender-intelligence associations reflect upon the agent's cares. And, indeed, this broader dispositional profile is evident. The gender-career IAT is moderated by participants' attitudes toward authority (Rudman and Kilianski 2000) and friendliness (Rudman and Glick 2001), for example.

Biased behavior appears to reflect upon the agent's cares in this case too. Biased evaluations of CVs and résumés are predicted by IATs (Bertrand et al. 2005), and

[34] See footnote #20.

manipulations of the relevant cares, such as attractiveness of candidates, lead to changes in behavior (Quereshi and Kay 1986). The predicted behaviors are also multitrack. A person who is likely to give lower scores to CVs with a woman's name at the top compared to a man's name is also more likely to offer women lower starting salaries and less career mentoring (Moss-Racusin et al. 2012).

There is much still to learn, of course, about shooter bias, CV evaluations, and other BEIB. What's currently underexplored is how implicit attitudes change over time, both generally across the lifespan and specifically as a result of interventions.[35] Longitudinal studies could help to clarify how durable implicit biases are; the durability of these states is a key component of their issuing in the kinds of dispositions that reflect cares. Moreover, longitudinal studies examining multiple kinds of behavior as dependent variables, and that examine these behaviors across varied contexts, would also help to clarify exactly which token implicit biases, or types of implicit biases, really do reflect upon what agents care about.

In addition, more research is needed on folk judgments of responsibility for BEIB. In the one published study (of which I am aware) addressing this, Cameron et al. (2010) compared folk attributions of responsibility for BEIB under three conditions: (1) when implicit bias is defined as operating outside consciousness; (2) when implicit bias is defined as being difficult to control (i.e., automatic); and (3) when no definition of implicit bias is given. They found that participants were significantly more likely to pardon biased behavior when it is described as unconscious than when it is described as automatic (or when it is described as neither unconscious nor automatic). This seems crucial on a voluntarist view of moral responsibility. But on the view I have been urging, we need to know additional facts. Consider the details of Cameron and colleagues' study. After reading the vignette about "John," who acts in a biased way, participants were asked to agree or disagree (on a 5 point Likert scale) with four statements, which together constituted the study's "Moral Responsibility Scale." The statements were: (1) "John . . . is morally responsible for his treating African Americans unfairly;" (2) "John should be punished for treating African Americans unfairly;" (3) "John should not be blamed for treating African Americans unfairly" (reverse coded); and (4), "John should not be held accountable for treating African Americans unfairly" (reverse coded). While intuitively related, these items pick out distinct judgments (namely, judgments about moral responsibility as such, punishment, blame, and accountability). Participants overall responses to this scale don't necessarily indicate whether they distinguish between John *being* responsible, in the attributability sense, and *holding* John responsibility, in the sense of demanding that John explain himself or demanding that John be punished (see §8). It's possible that Cameron and colleagues' participants' judgments were driven entirely by their assessments of whether John ought to be held to account for his actions, particularly since none of the 4 items on the Moral Responsibility scale single out aretaic appraisals.[36]

[35] For some data, see Dunham et al. (2013) and Devine et al. (2012).

[36] See Redford & Ratliff (ms) for data suggesting that participant judgments about what agents have an obligation to foresee (rather than what they do in fact foresee) mediates their responsibility judgments for BEIB. Note also that Cameron and colleagues report that the 4 items on the scale had an "acceptable" internal consistency (Cronbach's $\alpha=.65$). This means that participants' answers to the 4 items on the scale were somewhat, but not strongly, correlated with each other, and that the scale may in fact reflect multiple distinct concepts.

7 Objections

For all of this, though, readers may have a number of reasonable objections to my argument.

7.1 Objection #1: The Deep Self Represents the Agent's Fundamental Evaluative Stance

In §3 I argued that implicitness is best characterized as a form of arationality. Implicit attitudes are mental states that are functionally insensitive to what agents explicitly take to be true or good. How can an attitude that is insensitive to what we explicitly take to be true or good reflect upon our cares? Levy, for example, argues that attributionists are forced to excuse agents for BEIB because implicit biases—while morally significant—do not reflect an agent's evaluative stance. He writes that attributionists are "committed to excusing agents' responsibility for actions caused by a single morally significant attitude, or cluster of such attitudes insufficient by themselves to constitute the agent's evaluative stance, when the relevant attitudes are unrepresentative of the agent. This is clearest in cases where the attitude is inconsistent with the evaluative stance" (2011, 256). Levy then goes on to describe implicit biases in these terms, stressing their arationality (or "judgment-insensitivity"), concluding that attributionists "ought to excuse [agents] of responsibility for actions caused by these attitudes."

Levy assumes a homogeneous conception of an agent's evaluative stance. This is what I take him to mean when he says that an attitude or cluster of attitudes "constitutes" the agent's "global" evaluative stance. If correct, and agents fundamentally have one more or less unified evaluative stance, which is not internally conflicted, then Levy is right that attributionists must excuse agents' BEIB in paradigmatic cases. But it is more plausible to think of the deep self as heterogeneous and (potentially) internally conflicted. Sripada (2015) calls this the "mosaic" conception of the deep self. This conception stresses the difference between cares and reflective states like belief (though of course they are often connected). The difference is precisely in the degree of arationality these states tolerate. "To believe X, believe that Y is incompatible with X, and believe Y is irrational," Sripada (2015) writes, but "to care for X, believe that Y is incompatible with X, and care for Y is not irrational." In other words, cares need not be internally harmonious, in contrast with (an ideally rational set of) beliefs or reflective judgments. While a reflective conception of the deep self would thus suffer for being internally conflicted, a care-based conception doesn't. Both one's BEIB and one's reflective egalitarian judgments can be thought of as reflecting one's deep self.

7.2 Objection #2: BEIB are Non-actions or are Wanton Actions

A related worry is that arational attitudes like implicit biases can't reflect cares because they cause mere behavior, not action as such. BEIB are not explicitly intended, and in the case of biased microbehavior might not even count as intentional actions. A related worry is that BEIB are wanton actions. Is it clear that BEIB manifest the kind of concern for the worth of one's actions that distinguishes nonwanton from wanton actions?[37]

These objections are not convincing, at least with respect to the kinds of cases I have discussed. The central cases of BEIB I have discussed are cases of ordinary intentional action. They run contrary to agents' intentions, but that does not suggest that they are non-actions. Moreover, in the central cases I've discussed—decisions about whether to shoot or not shoot a person, evaluations of CVs, etc.—agents demonstrate a manifest concern for the quality of their decisions and actions.

7.3 Objection #3: BEIB Involve the Wrong Kind of Emotion

A more difficult challenge for my view is the idea that BEIB are intentional, nonwanton actions that don't reflect the ordinary profile of a care. In particular, it is reasonable to think that BEIB don't reflect the tight link between caring and emotion that I stressed in §4. In the ordinary kind of case discussed by attributionists, if I care for my dear old St. Bernard, my feelings will rise and fall with his well-being. Above I discussed the fact that implicit attitudes are affective states, but is this enough to show that they really reflect something the agent cares about?

On Jaworska's understanding of emotion, this would not be enough. In arguing that emotions are referentially connected to attitudes and actions in a way that enables them to underwrite identity, Jaworska (2007, 555) draws a distinction between so-called primary and secondary emotions.[38] Only secondary emotions, on her view, are apt for underwriting identity. This is because secondary emotions alone involve self-awareness, deliberation, an understanding of one's situation, and systematic thought about the significance and consequences of one's feelings. Jaworska offers gratitude, envy, jealousy, and grief as examples of secondary emotions. A paradigmatic example of a primary emotion, on the other hand, is fear, of the sort that precipitates the driver of a car slamming on the brake when the car in front of her stops short, or precipitates a mouse crouching when it detects the shadow of a hawk pass overhead. (This is a telling example, given my argument above that shooter bias reflects a care which manifests as fear in the face of black men.) Other ostensible primary emotions are disgust, rage, and surprise. All of these, on Jaworska's view, are more or less stimulus-bound reflex-like responses that require no understanding of one's situation.

I am inclined to resist this approach for two reasons. First, it draws the theory of the deep self perilously close to voluntarism about moral responsibility, to which attributionism is meant to be an alternative.[39] If what it takes for an emotion to demonstrate the right kind of connection to an agent's cares is conscious deliberation and an understanding of one's situation, then it becomes unclear what role the emotion itself—rather than the agent's deliberation and understanding—is playing in underwriting identity. Second, Jaworska's division between primary and secondary emotions is too stark. Some emotions might

[37] Shoemaker (2003, 97) discusses why only nonwanton actions reflect cares. See also Lippert-Rasmussen (2003) for discussion of "whim" cases in which agents don't seem to have any significant attitudes toward their actions.

[38] See also Shoemaker (2003, 93–94) and the distinction between pleasure and "central affective states" in Haybron (2013).

[39] For illuminating discussion of how attributionism can seem to collapse into voluntarism about moral responsibility, see Holly Smith (2011).

match these descriptions, but many fall between the extremes. Fear, for example, is often neither a reflex-like stimulus-bound response to a cue nor a fully cognitively mediated attitude. Imagine that you are afraid of tomorrow's chemistry test. This feeling involves an understanding of your situation, a form of projecting consequences into the future, and valuing things like academic success. But the fear might easily escape your self-awareness, much like a mood might affect you without your noticing it, and the fear might run contrary to your overall understanding of the situation, which is that you're well-prepared, it's a minor test, and getting an A isn't everything.[40]

The affective component of BEIB fall into this middle zone. Fear induced by the sight of black faces is the product of encoding social stereotypes and evaluations, is sensitive to some of an agent's motivations and beliefs, and is mediated by the agent's perception of context. But it doesn't issue in coordinated plans, policies, and intentions, as Jaworska suggests secondary emotions do (2007, 557). Halo and compensation effects are similar. They're cognitively complex but aren't integrated into deliberation and planning. In both cases, the affective component of implicit attitudes is too low-level to figure into deliberation or conscious judgment, but this does not mean the attitudes are mere reflexes as a result.

Jaworska is right that not all affective reactions are linked to an agent's cares. Those that are display the kind of "conceptual convergence" I discussed in §4. This convergence does not depend on an emotion entering into an agent's self-awareness, deliberation, or policies.

7.4 Objection #4: BEIB Don't Reflect the Right Kind of Mental-Causal Links to Cares

A final objection worth considering focuses on how resistant one's cares can be to revision in light of what one values or believes. The specific objection is that BEIB can't reflect cares because BEIB aren't susceptible to mental causal pressure. In other words, the fact that I can't persuade myself to be unbiased suggests that these states aren't reflective of my cares.

The response to this objection is that while BEIB do resist revision by "reason alone," they are in fact susceptible to revision by a number of "indirect" self-regulation strategies.[41] For example, the expression of implicit attitudes in behavior changes with changes in socialization experiences and perceived group membership (Gawronski and Sritharan 2010). There is also evidence supporting the idea that the associations underlying implicit attitudes can themselves be changed. Evidence for this is found in studies of approach-training (Kawakami et al. 2007; Kawakami et al. 2008; Phills et al. 2011), evaluative conditioning (Olson and Fazio 2006), and increasing individuals' exposure to images, film clips, or even mental imagery depicting members of stigmatized groups acting in stereotype-discordant ways (Blair et al. 2002; Dasgupta and Greenwald 2001; Wittenbrink et al. 2001). For example, exposure to exemplars of counter-stereotypes appears to alter an agent's "statistical map" of social stereotypes (Dasgupta and Greenwald 2001 and Gawronski et al. 2008). While none of these techniques are classic forms of rational persuasion, they are tools for exerting mental causal pressure on one's own, or another's, mind. The fact that

[40] For views of emotions that fall in this "middle zone" between reflex-like and reason-like responses, see D'Arms and Jacobson (2000) and Prinz (2004). I am indebted to Madva (ms b) for the idea that one can be in a mood without noticing it.

[41] Although see Mandelbaum (2015) for argument that implicit attitudes are sometimes even susceptible to revision by reason alone. See also Holroyd (2012) for discussion of what makes a self-regulation strategy "indirect."

these tools appear to change behavior and attitudes suggests that BEIB do indeed reflect the right kind of mental-causal links to one's cares.

8 Conclusion

I have argued that agents are morally responsible for actions that reflect upon what they care about, in the sense that these actions open them to being evaluated as moral agents. I have proposed a view of what it means to care about something and what it means for an attitude or action to reflect upon that care. And I have argued that in paradigmatic cases, BEIB reflect upon agents' cares.

A pressing question for future work has to do with how this approach relates to other central features of the concept of moral responsibility. Does attributability justify blaming people for BEIB? Does it justify intervening as a bystander when one observes discrimination unfolding? Answering these questions is not only important for practical ethics. It is also important for responding to what I take to be a very reasonable incredulity one might still have in reaction to my argument. For all that I've said, one might think that it just can't be right to treat people as responsible for acting in ways they don't intend to act, or for acting in ways that they don't know they're acting. Relatedly, doesn't my argument entail a far too expansive conception of responsibility, one that would entail holding people responsible for phobic reactions, actions resulting from brainwashing, addictive behavior, and so on?

I think this objection may be borne of the thought that responsibility itself is a singular concept. But responsibility admits of kinds.[42] On one plausible view, there are at least three: attributability, answerability, and accountability (Shoemaker 2011). I take myself as having given here the argument for attributability for BEIB.[43] Answerability requires that an agent "be able (in principle) to cite what she took to be justifying reasons for her action or attitude" (Shoemaker 2011, 628, fn 62). It seems to me that agents are not answerable for BEIB in this sense, although of course I cannot explore this question in depth here. As Levy (2011, 256) puts it, "it makes no sense at all to ask me to justify my belief that *p* when in fact I believe that not-*p*; similarly, it makes no sense to ask me to justify my implicit racism when I have spent my life attempting to eradicate it."[44] Finally, accountability pertains to how we *hold* one another responsible, including intervening, seeking an apology or retribution, punishing,

[42] Or perhaps responsibility has multiple "faces," as Watson (1996) and Shoemaker (2011) suggest.

[43] Whether phobic reactions, actions resulting from brainwashing, or addictive behaviors are attributable to agents depends on whether the empirical facts meet the conditions I discussed in §4 and §5.

[44] One might demand that I attend to or explain (in an etiological sense) why I seem to believe both *p* and not-*p*. But demanding that I *justify* these conflicting beliefs does not make sense. This is, on my view, a problem for Smith's (2005, 2008) theory that moral responsibility just is answerability. This is relevant because Smith argues that agents are answerable for (what I would call) BEIB. She writes: "I think it is often the case . . . that we simply take or see certain things as counting in favor of certain attitudes without being fully aware of these reasons or the role they play in justifying our attitudes. And I think these normative 'takings' or 'seemings' can sometimes operate alongside more consciously formulated judgments to the effect that such considerations do not serve to justify our attitudes. So, for example, a person may hold consciously egalitarian views and yet still find herself taking the fact of a person's race as a reason not to trust her or not to hire her. In these cases, I think an answerability demand directed toward her racist reactions still makes perfect sense—a person's explicitly avowed beliefs do not settle the question of what she regards as a justifying consideration" (2012, 581, fn 10). I agree with Smith that "takings" and "seemings" may fall within the domain of moral responsibility, but I do not think this is because we are answerable for them, in the sense of being reasonably demanded to justify them. I address Smith's view in Brownstein, M. Manuscript, The Implicit Mind (Unpublished).

and more. Questions about accountability are questions about social obligations and rights; they are about what we owe one another in virtue of our social roles and relationships.[45] I suspect that when people think that we just can't be responsible for BEIB, what they may have in mind is that *holding* people responsible for BEIB seems inappropriate. This is a separate—albeit related—question from whether implicit biases reflect upon who we are as moral agents.[46] Often the expression of implicit bias in our behavior *does* reflect on us as moral agents, and this puts all of us on notice to figure out what we are accountable for doing about it.

Acknowledgments For invaluable suggestions and discussion, I am grateful to Jules Holroyd, Daniel Kelly, W. John Koolage, Joshua Knobe, Alex Madva, Mytro Mylopoulos, Jeremy Pober, Eric Schwitzgebel, David Shoemaker, Susanna Siegel, Holly Smith, Chandra Sripada, Natalia Washington, Robin Zheng, an anonymous reviewer for *The Review of Philosophy and Psychology*, and audiences at the fall 2012 Wittgenstein Workshop, the 2013 meeting of the Southern Society for Philosophy and Psychology, and the 2013 Pacific Division meeting of the American Philosophical Association. Thanks also to the Leverhulme Trust for funding the "Implicit Bias and Philosophy" workshops at the University of Sheffield from 2011–2013, out of which this work grew.

References

Amodio, D., and P. Devine. 2006. Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology* 91(4): 652.

Anderson, E. 2010. *The Imperative of Integration*. Princeton: Princeton University Press.

Anderson, C.A., and J.J. Lindsay. 1998. The development, perseverance, and change of naïve theories. *Social Cognition* 16: 8–30.

Arpaly, N. 2004. *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Banaji, M., R. Bhaskar, and M. Brownstein. 2015. When bias is implicit, what should be the model for repairing harm? *Current Opinion in Psychology* 6: 183–188.

Bertrand, M. and Mullainathan, S. 2003. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market. No. 9873, NBER Working Papers from National Bureau of Economic Research, Inc.

Bertrand, M., Chugh, D., and Mullainathan, S. 2005. Implicit discrimination. *American Economic Review*, 94–98.

Blair, I., C. Judd, M. Sadler, and C. Jenkins. 2002. The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology* 83(1): 5.

Bodenhausen, G., and B. Gawronski. 2014. Attitude Change. In *The Oxford Handbook of Cognitive Psychology*, ed. D. Reisberg. New York: Oxford University Press.

Bratman, M. 2000. Reflection, planning, and temporally extended agency. *The Philosophical Review* 109(1): 38.

Brownstein, M. 2015. Implicit Bias. The Stanford Encyclopedia of Philosophy (Spring 2015 Edition). Ed. E. Zalta. <http://plato.stanford.edu/entries/implicit-bias/>. Accessed 1 April 2015.

Brownstein, M. Forthcoming. Implicit Bias and Race. In *The routledge companion to the philosophy of race*, eds. P. Taylor, L. Anderson, and L. Alcoff. New York: Routledge.

Cameron, C.D., B.K. Payne, and J. Knobe. 2010. Do theories of implicit race bias change moral judgments? *Social Justice Research* 23(4): 272–289.

Carlsson, R., and F. Björklund. 2010. Implicit stereotype content: Mixed stereotypes can be measured with the implicit association test. *Social Psychology* 41(4): 213–222.

[45] On the social nature of accountability, see, for instance, Watson (1996, 262): "Holding people responsible is not just a matter of the relation of an individual to her behavior; it also involves a social setting in which we demand (require) certain conduct from one another and respond adversely to one another's failures to comply with these demands."

[46] For discussion of holding-responsible for BEIB, see Banaji et al. (2015).

Correll, J., B. Park, C. Judd, and B. Wittenbrink. 2002. The police officer's dilemma: Using race to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology* 83: 1314–1329.

Correll, J., G. Urland, and T. Ito. 2006. Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology* 42: 120–128.

Correll, J., B. Park, C.M. Judd, B. Wittenbrink, M.S. Sadler, and T. Keesee. 2007. Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology* 92(6): 1006.

Dardenne, B., M. Dumont, and T. Bollier. 2007. Insidious dangers of benevolent sexism: Consequences for women's performance. *Journal of Personality and Social Psychology* 93(5): 764–779.

D'arms, J., and D. Jacobson. 2000. Sentiment and value. *Ethics* 110(4): 722–748.

Dasgupta, N., and A.G. Greenwald. 2001. On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology* 81: 800–814.

Devine, P., P. Forscher, A. Austin, and W. Cox. 2012. Long-term reduction in implicit race bias; A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology* 48(6): 1267–1278.

Donders, N.C., J. Correll, and B. Wittenbrink. 2008. Danger stereotypes predict racially biased attentional allocation. *Journal of Experimental Social Psychology* 44(5): 1328–1333.

Dunham, Y., E.E. Chen, and M.R. Banaji. 2013. Two signatures of implicit intergroup attitudes developmental invariance and early enculturation. *Psychological Science* 24(6): 860–868.

Eberhardt, J., P. Goff, V. Purdie, and P. Davies. 2004. Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology* 87(6): 876.

Faucher, L. Forthcoming. Revisionism and Moral Responsibility for Attitudes. In *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*, eds. Brownstein, M. and Saul, J. Oxford University Press.

Fisher, J.M., and M. Ravizza. 1998. *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Fiske, S.T., A.J. Cuddy, P. Glick, and J. Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* 82(6): 878.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.

Frankfurt, H. 1988. *The importance of what we care about*. Cambridge: Cambridge University Press.

Gawronski, B., and Bodenhausen, G.V. 2006. Associative and Propositional Processes in Evaluation: Conceptual, Empirical, and Metatheoretical Issues: Reply to Albarracín, Hart, and McCulloch (2006), Kruglanski and Dechesne (2006), and Petty and Briñol, (2006). *Psychological Bulletin* 132:5, 745–750.

Gawronski, B., and G. Bodenhausen. 2011. The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology* 44: 59–127.

Gawronski, B., and R. Sritharan. 2010. Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In *Handbook of implicit social cognition: Measurement, theory, and applications*, ed. B. Gawronski and B.K. Payne. New York: Guilford Press.

Gawronski, B., R. Deutsch, S. Mbirkou, B. Seibt, and F. Strack. 2008. When "Just Say No" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology* 44: 370–377.

Gendler, T.S. 2008a. Alief and belief. *The Journal of Philosophy* 105(10): 634–663.

Gendler, T.S. 2008b. Alief in action (and reaction). *Mind and Language* 23(5): 552–585.

Glaser, J., and E. Knowles. 2008. Implicit motivation to control prejudice. *Journal of Experimental Social Psychology* 44: 164–172.

Glasgow, J. Forthcoming. Alienation and Responsibility. In *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics,* eds. Brownstein, M. and Saul, J. Oxford University Press.

Greenwald, A.G., T.A. Poehlman, E.L. Uhlmann, and M. Banaji. 2009. Understanding and using the implicit association test: III meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97(1): 17–41.

Greenwald, A., M. Banaji, and B. Nosek. 2015. Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology* 108(4): 553–561.

Hardin, C. and Banaji, M. 2013. The Nature of Implicit Prejudice: Implications for personal and public policy. *The Behavioral Foundations of Public Policy*, 13–30.

Haslanger, S. 2015. Social structure, narrative, and explanation. *Canadian Journal of Philosophy.* doi:10.1080/00455091.2015.1019176.

Haybron, D. 2013. *Happiness: A very short introduction*. Oxford: Oxford University Press.

Hieronymi, P. 2008. Responsibility for believing. *Synthese* 161: 357–373.

Hinds, A. 30 May 2013. "Heel." <http://www.slate.com/articles/double_x/doublex/2013/05/stay_at_home_dad_sexual_fantasies_why_i_d_like_to_stop.html>

Holroyd, J. 2012. Responsibility for implicit bias. *Journal of Social Philosophy* 43(3): 274–306.

Holroyd, J. and J. Sweetman. Forthcoming. The Heterogeneity of Implicit Biases. In *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, eds. M. Brownstein and J. Saul. Oxford: Oxford University Press.

Hursthouse, R. 1999. *On virtue ethics*. Oxford: Oxford University Press.

Ito, T., and G. Urland. 2005. The influence of processing objectives on the perception of faces: An ERP study of race and gender perception. *Cognitive, Affective and Behavioral Neuroscience* 5: 21–36.

Jaworska, A. 1999. Respecting the margins of agency: Alzheimer's patients and the capacity to value. *Philosophy and Public Affairs* 28(2): 105–138.

Jaworska, A. 2007. Caring and internality. *Philosophy and Phenomenological Research* 74(3): 529–568.

Jost, J.T., L.A. Rudman, I.V. Blair, D.R. Carney, N. Dasgupta, J. Glaser, and C.D. Hardin. 2009. The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior* 29: 39–69.

Kawakami, K., J.F. Dovidio, and S. van Kamp. 2007. The impact of counterstereotypic training and related correction processes on the application of stereotypes. *Group Processes and Intergroup Relations* 10(2): 139–156.

Kawakami, K., J. Steele, C. Cifa, C. Phills, and J. Dovidio. 2008. Approaching math increases math=me, math=pleasant. *Journal of Experimental Social Psychology* 44: 818–825.

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

Levy, N. 2011. Expressing who we are: Moral responsibility and awareness of our reasons for action. *Analytical Philosophy* 52(4): 243–261.

Levy, N. 2012. Consciousness, implicit attitudes, and moral responsibility. *Noûs* 48: 21–40.

Levy, N. 2014. Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs*. doi:10.1111/nous.12074.

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

Lippert-Rasmussen, K. 2003. Identification and responsibility. *Ethical Theory and Moral Practice* 6: 349–376.

Madva, A. 2012. The hidden mechanisms of prejudice: Implicit bias and interpersonal fluency. PhD dissertation. Columbia University.

Madva, A. Manuscript a, Biased Against De-Biasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle Against Prejudice.

Madva, A. Manuscript b, Implicit Bias, Moods, and Moral Responsibility.

Madva, A. and M. Brownstein, Manuscript, The Blurry Boundary Between Stereotyping and Evaluation in Implicit Cognition.

Madva, A. Forthcoming. Why Implicit Attitudes are (Probably) Not Beliefs. *Synthese.*

Mandelbaum, E. 2015. Attitude, association, and inference: On the propositional structure of implicit bias. *Noûs*. doi:10.1111/nous.12089.

Mekawi, Y., and K. Bresin. 2015. Is the evidence from racial bias shooting task studies a smoking gun? Results From a Meta-Analysis. *Journal of Experimental Social Psychology*. doi:10.1016/j.jesp.2015.08.002.

Moss-Racusin, C., J. Dovidio, V. Brescoll, M. Graham, and J. Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of the Sciences*. doi:10.1073/pnas.1211286109.

Newman, G.De., J. Freitas, and J. Knobe. 2014. Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*. doi:10.1111/cogs.12134.

Nosek, B. A. and Banaji, M. R. 2009. Implicit attitude. *Oxford Companion to Consciousness*, 84–85.

Nosek, B.A., A.G. Greenwald, and M.R. Banaji. 2007a. The implicit association test at age 7: A methodological and conceptual review. In *Automatic Processes in Social Thinking and Behavior*, ed. J.A. Bargh. Philadelphia: Psychology Press.

Nosek, B.A., F.L. Smyth, J.J. Hansen, T. Devos, N.M. Lindner, K.A. Ranganath, C.T. Smith, K. Olson, D. Chugh, A.G. Greenwald, and M.R. Banaji. 2007b. Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology* 18(1): 36–88.

Olson, M., and R. Fazio. 2006. Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin* 32: 421–433.

Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P.E. 2013. Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology.* doi: 10.1037/a0032734.

Phills, C., K. Kawakami, E. Tabi, D. Nadolny, and M. Inzlicht. 2011. Mind the Gap: Increasing the associations between the self and blacks with approach behaviors. *Journal of Personality and Social Psychology* 100: 197–210.

Prinz, J. 2004. *Gut reactions: A perceptual theory of emotion*. New York: Oxford University Press.

Quereshi, M.Y., and J.P. Kay. 1986. Physical attractiveness, age, and sex as determinants of reactions to resumes. *Social Behavior and Personality: An International Journal* 14(1): 103–112.

Rudman, L.A., and R.D. Ashmore. 2007. Discrimination and the implicit association test. *Group Processes and Intergroup Relations* 10(3): 359–372.

Rudman, L.A., and P. Glick. 2001. Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues* 57(4): 743–762.

Rudman, L.A., and S.E. Kilianski. 2000. Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin* 26(11): 1315–1328.

Saul, J. 2012. Skepticism and implicit bias. *Disputatio Lecture* 5(37): 243–263.

Saul, J. 2013. Unconscious influences and women in philosophy. In *Women in philosophy: What needs to change?* ed. F. Jenkins and K. Hutchison. Oxford: Oxford University Press.

Scanlon, T.M. 1998. *What we owe each other*. Cambridge: Harvard University Press.

Schwitzgebel, E. 2010. Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly* 91: 531–553.

Sher, G. 2009. *Who Knew? Responsibility without Awareness*. Oxford: Oxford University Press.

Shoemaker, D. 2003. Caring, identification, and agency. *Ethics* 118: 88–118.

Shoemaker, D. 2011. Attributability, answerability, and accountability: Towards a wider theory of moral responsibility. *Ethics* 121: 602–632.

Sie, M. and Vorst Vader-Bours, N. Forthcoming. Personal Responsibility vis-à-vis Prejudice Resulting from Implicit Bias. In *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics,* eds. Brownstein, M. and Saul, J. Oxford University Press.

Smart, J. J. C. 1961. Free-will, praise and blame. *Mind*, 291–306.

Smith, A. 2005. Responsibility for attitudes: Activity and passivity in mental life. *Ethics* 115(2): 236–271.

Smith, A. 2008. Control, responsibility, and moral assessment. *Philosophical Studies* 138: 367–392.

Smith, H. 2011. Non-tracing cases of culpable ignorance. *Criminal Law and Philosophy* 5: 115–146.

Smith, A. 2012. Attributability, answerability, and accountability: In defense of a unified account. *Ethics* 122(3): 575–589.

Sripada, C.S. 2010. The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies* 151(2): 159–176.

Sripada, C. 2015. Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*. doi:10.1007/s11098-015-0527-9.

Stewart, B., and B. Payne. 2008. Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin* 34: 1332–1345.

Strohminger, N., and S. Nichols. 2014. The essential moral self. *Cognition* 131(1): 159–171.

Valian, V. 2004. Beyond gender schemas: Improving the advancement of women in academia. *NWSA Journal* 16(1): 207–220.

Valian, V. 2005. Beyond gender schemas: Improving the advancement of women in academia. *Hypatia* 20: 198–213.

Washington, N. and Kelly, D. Forthcoming. Who's responsible for this? Implicit bias and the knowledge condition. In *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*, eds. Brownstein, M. and Saul, J. Oxford University Press.

Watson, G. 1975. Free agency. *Journal of Philosophy* 72: 205–220.

Watson, G. 1996. Two faces of responsibility. *Philosophical Topics* 24(2): 227–248.

Wittenbrink, B., C.M. Judd, and B. Park. 2001. Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology* 81(5): 815.

Zheng, R. Forthcoming. Attributability, Accountability and Implicit Attitudes. In *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*, eds. Brownstein, M. and Saul, J. Oxford University Press.