Michael Brownstein

Forthcoming in Brownstein, M. and Saul, J. (Eds). *Implicit Bias and Philosophy Volume II: Moral Responsibility, Structural Injustice, and Ethics*. Oxford University Press.

Implicit Bias, Context, and Character

Word Count: 7,610 (including footnotes and references)

§1. Introduction

Early research on implicit attitudes and implicit biases emphasized the "directness" of the link between apparent triggers of those attitudes and behavior. For example, Bargh and colleagues argued that there is a direct link between the perception of cues relevant to one's implicit attitudes and behavior; they write (1996, 231): "social behavior is often triggered automatically on the mere presence of relevant situational features." The implication of this view is that merely being in the presence of a member of a socially stigmatized group may be sufficient to activate implicit biases and to cause one to act in biased ways.[1] In a related vein, in her seminal paper, Devine (1989) argued that scores on measures of implicit bias reflect mere knowledge of cultural stereotypes. She showed that both egalitarians and non-egalitarians associate Blacks with negative stereotypes, and suggested that there is a direct relationship between merely having cultural knowledge of stereotypes and behaving in biased ways.[2]

More recently, though, context has been shown to be a significant and pervasive moderator of the activation and expression in behavior of implicit attitudes and implicit biases. Some philosophers have begun to consider the ramifications of this development in the empirical literature. Understandably, most who have considered the effects of context on implicit attitudes and biases have focused on the metaphysical ramifications of this fact, in particular whether implicit attitudes represent a durable or singular kind, or rather whether the implicit attitude construct stands for a number of related (or perhaps unrelated) cognitive and affective processes.[3] In this paper, I focus on a separate philosophical issue stemming from findings about the contextual moderation of implicit biases. There are, I will argue, ethical ramifications stemming from these findings. Any comprehensive "ethics of implicit bias," I will argue, must focus on outlining how agents can cultivate the right sort of relationships with the situations and contexts that affect their behavior.

---

[1] See also Dijksterhuis and Bargh (2001).

[2] See Nosek and Hansen (2008) for discussion.

[3] See, for instance, Machery (this volume); Holroyd and Sweetman (this volume); Madva and Brownstein (in prep.); and discussion between Mazarin Banaji and Tamar Gendler on "The Mind Report" (http://bloggingheads.tv/videos/15811).

This notion, of cultivating the right sort of "ambient" relationships, has been under-described by most ethical thinking about implicit bias, which usually focuses on the relationship between attitudes or mental states *within* agents.

First, I will outline recent data on the effects of context on the activation and expression in behavior of implicit biases (§2). I will then briefly describe the predominant ways of thinking about the ethics of implicit bias (§3). Both philosophers and psychologists typically emphasize the "harmonizing" of one's internal and external states. Such harmonizing is no doubt central in the fight against implicit bias. However, the crucial ambient relationship between agents and their situations and contexts is not easily incorporated into this ideal. My aim is thus to provide a tentative outline of an "ethics of ambient harmony." I will first distinguish my view from related two claims. In re-focusing their attention outward, toward their ambient relationships, agents who want to act in egalitarian ways need not think that their only options are to "change the world" or to simply avoid ethically compromising situations, as some have argued in the related "situationist" literature (§4). Rather, agents can work to cultivate the right sorts of relationships between themselves and their contexts; I provide three examples (§5).

§2. Context and Implicit Bias

In a recent paper, Gawronski and Cesario (2013) argue that implicit attitudes are subject to "renewal effects," which is a term used in animal learning literature to describe the recurrence of an original behavioral response after the learning of a new response. As Gawronski and Cesario explain, renewal effects usually occur in contexts other than the one in which the new response was learned. For example, a rat with a conditioned fear response to a sound may have learned to associate the sound with an electric shock in context A (e.g. its cage). Imagine then that the fear response is extinguished or counter-conditioned in context B (e.g. a different cage). An "ABA renewal" effect occurs if, upon being placed back in A (its original cage), the fear response returns. Gawronski and Cesario argue that implicit attitudes are subject to renewal effects like these. For example, a person might learn biased associations while hanging out with their friends (context A), effectively learn counter-stereotypical associations while taking part in a psychology experiment (context B), then exhibit behaviors consistent with biased associations when back with their friends. Gawronski and Cesario discuss several studies (in particular, Rydell and Gawronski (2009) and Gawronski et al. (2010)) that demonstrate renewal effects like these in implicit attitudes in controlled laboratory settings. The basic method of these studies involves an impression formation task in which participants are first presented with valenced information about a target individual who is pictured against a background of a particular color. This represents context A. Then, the same target individual is presented with oppositely valenced information against a background of a different color, representing context B. Participants' evaluations of the target are then assessed using an affective priming task in which the target is presented against the color of context A. In this ABA pattern, participants' evaluations reflect what they learned in context A. Gawronski and Cesario report similar renewal effects in AAB (where the original information and new information

are presented against the same background color A, and evaluations are measured against a novel background B) and ABC (where original information, new information, and evaluation all take place against different backgrounds) patterns.

These studies suggest that minor features of agents' context—like the background color against which an impression of a person is formed—can influence the activation of implicit attitudes, even after those attitudes have been "unlearned." These are striking results, but they are consistent with a broad array of findings about the influence of context and situation on the activation and expression in behavior of implicit biases and attitudes. Context is, of course, a broad notion. Imagine taking an IAT. The background color against which the images of the target subjects are presented is part of the context. Perhaps before having taken the IAT, the experimenter asks the subject to imagine herself as a manager at a large company deciding whom to hire. Having imagined oneself in this powerful social role will have effects on one's implicit evaluations (see below), and these effects too can be thought of as part of one's context. Similarly, perhaps one had a fight with one's friend before entering the lab, and began the task feeling an acute sense of disrespect, or was hungry and jittery from having drank too much caffeine, etc. . . .[4]

As I will use the term, context can refer to any stimulus that moderates the way an agent evaluates or responds behaviorally to a separate conditioned stimulus. Anything that acts, in other words, as what theorists of animal learning call an "occasion setter" can count as an element of context.[5] A standard example of an occasion setter is an animal's cage; a rat may demonstrate a conditioned fear response to a sound when in its cage, but not when in a novel environment. Similarly, a person may feel or express biased attitudes toward members of a social group only when in a particular physical setting, when playing a certain social role, when feeling a certain way, etc. To briefly review the extant data, I will note three kinds of contextual elements—perceptual, conceptual, and motivational—that have been shown to influence the activation and expression in behavior of implicit biases. These three kinds of contextual elements overlap significantly, of course. I carve them up this way only as a device for presentation of the empirical literature.

Perceptual elements of context influence implicit biases primarily in light of their visual, aural, or other sensory properties. For example, Schaller and colleagues (2003) found that the relative darkness or lightness of the room in which participants sit shifts scores on implicit racial evaluations across several indirect measures, including the IAT.[6] The sort of renewal effects Gawronski and Cesario discuss (above) are also examples of perceptual elements of context influencing the activation of implicit attitudes.

Conceptual elements of context influence implicit biases primarily in light of individuating group membership and social roles. Barden and colleagues (2004), for example, varied the category

[4] Perhaps it would be better to refer to all of these various elements as "situational factors" rather than elements of context. See below for discussion. I make no principled distinction between context and situational factors.
[5] I am indebted to Gawronski and Cesario (2013) for the idea that context acts as an occasion setter. On occasion setting and animal learning, see Schmajuk and Holland (1998).
[6] These results obtained only for subjects with chronic beliefs in a dangerous world.

membership of targets by presenting the same individual in a prison context dressed as a prisoner and dressed as a lawyer; implicit evaluations of the person dressed as prisoner were considerably more negative. Similarly, Mitchell and colleagues (2003) showed that implicit evaluations of the same individual—Michael Jordan—depended on whether the individual was categorized by race or occupation. Conceptual elements of context also include one's own social role; Guinote and colleagues (2010), for example, led participants in a high-power condition to believe that their opinions would impact the decisions of their school's "Executive Committee," and these participants showed more racial bias on both an IAT and an Affect Misattribution Procedure (AMP) than those in a low-power condition, who were led to believe that their opinions would not affect the committee's decisions.[7]

Finally, motivational elements comprising context influence implicit biases, and they do so primarily in light of fluctuations in mood and emotion. Dasgupta and colleagues (2009), for instance, found that salient emotions selectively influence the activation of implicit biases. Participants who were induced to feel disgust had more negative evaluations of homosexuals on an IAT, although their implicit evaluations of Arabs remained unchanged. However, participants who were induced to feel anger had more negative evaluations of Arabs, while their evaluations of homosexuals remained unchanged.[8]

These data should not be surprising. Anecdotally, it is not hard to imagine a person who treats her colleagues fairly regardless of race, but (unwittingly) grades her students unfairly on the basis of race. Perhaps being in the superordinate position of professor activates this person's prejudices while being in an equal-status position with her colleagues does not. Perhaps perceptual and motivational factors play a role too. There could be posters on the wall of her classroom for films that propagate stereotypes, and these posters might render her stereotypical associations more psychologically available. As might simply being in a bad mood, or being under-slept, etc.

§3. The Ethics of Internal Harmony

One reason implicit biases are ethically pernicious is that in many cases they tend to persist and to influence the behavior of individuals who don't endorse the validity of those very biases

[7] Richeson and Ambady (2001) obtained similar results with respect to implicit gender bias; they found that assigning male participants to a subordinate role in a dyadic interaction with a woman led the men to have more negative implicit evaluations of women on an IAT, while male participants assigned to an equal-status or superordinate role showed favorable evaluations of the women.

[8] See Gawronski and Sritharan (2010) for summary and discussion of these data. For more on context and implicit bias, one could look to the literature on the relationship between context and habit (on the plausible assumption that the behavioral upshots of implicit bias are habit-like). Neal and colleagues (2011) showed, for example, that people who habitually eat popcorn at the movies will be minimally influenced by hunger or by how much they like what they are eating, but only when the right physical context-cues obtain. People eat less out of habit if they are not in the right place—e.g. a meeting room rather than a cinema—or if they cannot eat in their habitual way—e.g. if forced to eat with their non-dominant hand.

(Gawronski and Bodenhausen, 2006; Nosek and Hansen, 2008).[9] They are aversive with respect to agents' reflective or moral commitments, in other words, giving rise to "aversive racism" (Dovidio et al., 2000, 2004). Philosophers writing on implicit bias have focused on this fact, and for good reason.[10] The idea that implicit biases can persist and influence the behavior of individuals who disavow them raises important ethical questions (not to mention metaphysical and epistemological questions). For example, Gendler (2008b) describes the ethical conflict arising from the pervasiveness of implicit bias in terms of agents being put into a state of "internal disharmony." This state is the result of discord between one's "aliefs"—which for all relevant purposes here we can think of as one's implicit attitudes—and one's beliefs. The relevant ideal to which one can aspire in order to combat internal disharmony is, of course, internal harmony. This is an ideal with a long history, stretching all the way back to Plato, who claimed that a just person is one who "puts himself in order, harmonizes . . . himself . . . [and] becomes entirely one, moderate and harmonious . . ." (*Republic*, 443de in Plato, 380BCE/1992; quoted in Gendler, 2008b, 572).

This ideal of internal harmony—or something quite like it—is also found in the psychological literature. Payne and Cameron (2010, 445), for example, write, "the message of implicit social cognition is that the thoughts people introspect and report about do not tell the whole story of why they believe the things they believe and why they do the things they do." It is important that people learn this message, furthermore, because implicit cognition, like life in Hobbes' state of nature, can be "nasty, brutish, and short-sighted" (2010, 445). When we learn the facts about implicit social cognition, in particular how it "can cause our ethicality to corrode," we can "engage [in] better moral self-regulation in pursuit of our ideals" (2010, 456). Payne and Cameron's chapter in the *Handbook of Implicit Social Cognition* even begins with this epitaph from Rousseau: "virtue is a state of war, and to live in it we have always to combat with ourselves."

It is worth noting that these related ways of thinking about the ethics of implicit bias are *agent-centered* and conceptually *bilateral*. They are agent-centered in the sense that they recommend self-regulatory effort aimed at controlling or changing mental and emotional states internal to agents. Internal harmony describes an occurrent or dispositional status of one's own cognitive and emotional states. Gendler considers two strategies for regulating one's "belief-discordant aliefs" in the hopes of becoming more internally harmonious (2008b, 554): the "cultivation of alternative habits through deliberate rehearsal" and/or "refocusing of attention through directed imagination." Both of these strategies focus directly on the agent herself, and in two senses. First, both the cultivation of habits and the refocusing of attention are things one does to oneself; the object of attention and regulatory effort are one's own habits or one's own patterns of attention. Second, both the cultivation of habits and the refocusing of attention are things one does in order to create harmony between one's implicit and explicit attitudes, both of which "belong to" the agent.

The ethics of internal harmony is bilateral in the sense that it conceives of there being two relevant "forces" calling for self-regulation: one's implicit attitudes and one's explicit attitudes. This

[9] Of course, the greater *moral* problem is that implicit bias perpetuates injustices.
[10] See, for instance, Gendler (2008a,b and 2011); Madva (2012); Huebner (2009); and Kelly and Roedder (2008).

is particularly salient in Payne and Cameron's use of the metaphor of battle. Virtue is a state of war, presumably the state in which one's explicit attitudes win the battle (or keep trying to win it) for the control of one's judgment and behavior against one's nasty and brutish implicit attitudes.

Nothing I say in this paper should be taken as a complaint against the cultivation of habits, the refocusing of one's attention, or the constant effort to be vigilant against internal barriers to acting as one hopes to act. These efforts are all of the utmost value. Nor should anything I say be construed as an attack on the ideal of internal harmony *per se*; there is no doubt that treating others unfairly on the basis of race (or gender, sexual preference, age, or membership in other socially stigmatized groups) puts one in a distinct ethical dilemma insofar as one is explicitly committed to egalitarian values.[11] However, *prima facie*, data attesting to the pervasive moderating effects of context on implicit biases suggest that agents ought to focus their attention outward, onto the relevant details of their contexts, at least in addition to focusing their self-regulatory attention on themselves. For if context acts as an occasion setter for implicit biases, why wouldn't one focus on *it* (i.e. the context)? An exclusively agent-centered ethics risks ignoring the moderating effects of context, thus raising the possibility that one could be ideally internally harmonious yet nevertheless behave in biased ways in certain situations. The bilateralism of the ethics of internal harmony offers another way to articulate this concern. If the constitutive combatants are one's implicit and explicit attitudes, what is the role for crucial moderating elements like context? Which side of the battle are the causes of one's mood on? Are social roles (like those discussed above) merely peripheral players in the battle for virtue?

The internal harmonist might reply to worries like these by suggesting that I'm attacking a straw man. No one recommends that the way to combat implicit biases is to simply focus on oneself. The ideal of internal harmony is broad, one might argue; it simply articulates the long-term goal, and is not itself a practical recommendation for where to fix one's attention. Rather, in a related vein, the internal harmonist might argue that the ideal provides something like the normative ground for the self-regulation of implicit bias. So, even if one can successfully regulate one's implicit biases by looking outward and attending to one's context (in a way not yet specified), the reason one would do so is in order to bring one's behavior and judgment in line with one's explicit ideals. In other words, the ideal of internal harmony simply provides the abstract normative grounds for any practical effort—inwardly or outwardly focused—to regulate one's implicit biases.

If the ideal of internal harmony is abstract in this sense—that is, that it is not a practical recommendation about *how* to regulate one's biases but rather a formulation of *what* one is doing when trying to regulate one's biases—then there is no conflict in principle between the context-dependence of implicit biases and the ideal of internal harmony. However, I have several concerns about this interpretation of the ideal. For one, I do not think the proponents of the ideal of internal harmony mean it to be simply an abstract account of what one is doing when one is fighting against

[11] Elsewhere Alex Madva and I have argued, however, that there are some cases in which it is ethically valuable for one's implicit attitudes to guide one's behavior, in spite of, or in the absence of concordance with one's explicit attitudes. See Brownstein and Madva (2012).

bias or a theoretical grounding for whatever self-regulation strategy one chooses. This is why Gendler, Payne and Cameron, and others offer specific strategies for promoting internal harmony. Second, and more substantively, an abstract account of an ideal should have a fairly tight motivational connection to the particular activities that support the ideal. For example, people often say that they want to be "be happy" above all else. Empirical research suggests, however, that happiness might be self-deflecting; like falling asleep or relaxing, trying to get it can make it harder to attain.[12] At some level of generality, of course, finding happiness indirectly, by spending time with one's family or pursuing hobbies or whatever, instead of pursuing it directly, is still a way of fulfilling the ideal of finding happiness. But this doesn't do much to vindicate the ideal of being happy above all else. The relevant practical strategies don't follow in any obvious way from the ideal, and even seem counterintuitive from its perspective. Similarly, if there are important strategies for combatting implicit bias that require one to cultivate particular relationships with elements of one's context, then these strategies would stand in a similar relationship to the ideal of internal harmony. They would not be inconsistent with the ideal, but wouldn't follow from it in any obvious way, and might even seem counterintuitive from its perspective.

Thus far I hope to have shown that some kind of outwardly-focused ethical ideal is demanded by evidence demonstrating the context-dependence of implicit biases. In the next section, I will briefly discuss two outwardly-focused ethical ideal which are valuable, but also both limited in important ways.

§4. The World-First and Seek/Avoid Strategies

Some have argued that the fight against prejudice and bias cannot be won unless the world itself changes. Generally, the idea behind this argument is that prejudice and bias are sustained by unjust political, economic, and social institutions, and that so long as these institutions persist unchanged, no amount of individual self-regulation can undo the pernicious effects of racism, sexism, etc. Call this the "world-first" strategy. Put a bit crassly, it states that the only way to change individuals is to change the world. Huebner (2009, 88) seems to defend the world-first strategy, for example, arguing that "the only way in which we will be able to adequately modify our psychology is by modifying the world in which we live."[13] Huebner's view is that in a propaganda-filled world like ours, where even the most harmonious agent continues to be bombarded with stereotypes, etc., any ethical strategy focused on making change at the individual level amounts to too little, too late.

[12] See Schooler et al. (2003) and Green et al. (2003). John Stuart Mill also articulated this idea long ago, writing that those are happy who "have their minds fixed on some object other than their own happiness; on the happiness of others, on the improvement of mankind, even on some art or pursuit, followed not as a means, but as itself an ideal end" (1873/1944, p. 100).

[13] Huebner does, however, discuss some self-regulatory strategies for modifying one's psychology without first changing the world. See also Haslanger (this volume). For a defense of "psychological" approaches to combating implicit bias, see Machery et al. (2010).

The world-first strategy is one way of focusing one's attention outward in an effort to combat bias. As a call to action, the world-first strategy is undoubtedly valuable. We *should* try to change the unjust political, economic, and social institutions that sustain prejudice and bias. But as I understand it the world-first strategy is not a call for simultaneous political activism and self-regulation. Rather, it is the notion that self-regulation at the individual level is either hopeless or subsidiary to change in the world at large. But of course this is not a valuable way to combat prejudice. One ineliminable feature of social change is change in those individuals themselves who populate unjust institutions.[14]

A second kind of outwardly-focused ethics is articulated in the "situationist" literature. Critics of virtue ethics like Harman (1999) and Doris (2002) might take the data about the context-dependence of implicit biases as reason to claim that implicit biases are yet another example of how unexpected and seemingly trivial features of the situations that we are in can strongly affect our behavior and attitudes, regardless how harmonious or peaceful a character we (think we) have. There is a sense in which Gawronski and Cesario's background colors in IAT trials are analogous to the oft-cited Isen and Levin (1972) dime in a payphone; similarly, the conceptual and motivational elements of context that affect implicit biases are analogous in some respects to other central situationist examples, like Darley and Batson (1973) and Milgram (1974/2009).[15]

Situationists tend to be wary of agent-centered ethical ideals. Instead, some situationists argue that the best avenue to ethical action is to focus our attention on the situations we are in and the effects those situations have on us. This recommendation is similar to the one I am making. However, situationists have articulated this ideal in terms of what Sarkissian (2010) calls a "seek/avoid" strategy. This strategy is to seek out situations that are likely to promote wanted attitudes and behavior and avoid situations that are likely to compromise wanted attitudes and behavior. Harman (2003, 91) articulates the seek/avoid strategy clearly: "If you are trying not to give into temptation to drink alcohol, to smoke, or to eat caloric food, the best advice is not to try to develop 'will-power' or 'self-control'. Instead, it is best to head [*sic*] the situationist slogan, 'People! Places! Things!' Don't go to places where people drink! Do not carry cigarettes or a lighter and avoid people who smoke! Stay out of the kitchen!"

Unfortunately, the seek/avoid strategy is often hamstrung in real life situations, as Sarkissian makes clear. He offers four reasons (2010, 5). First, one has to know which situations to avoid ahead of time, and many situations are neither good nor bad *simpliciter* such that one can know to seek or avoid them ahead of time. Second, some problematic situations are practically unavoidable. Third, there are times when one's ethical commitments require one to enter compromising situations. And fourth, the problematic variables inherent in ordinary situations are so finely

[14] The novelist Tom Robbins put it nicely, if not a bit snarkily, when he wrote in *Still Life with Woodpecker*, "Political activism is seductive because it seems to offer the possibility that one can improve society, make things better, without going through the personal ordeal of rearranging one's perceptions and transforming one's self."

[15] Nothing I say in this paper should be construed as support or defense of the specific situationist critique of virtue ethics. I do not know whether whatever traits people have substantiate the moral psychology required by virtue ethicists.

individuated that it is hard to know how agents could ever discriminate between them. Consider these arguments cashed out in terms of deciding whether to go to a party while trying to quit smoking: first, you do not know whether people will be smoking at the party; second, some parties are more-or-less obligatory (e.g. office parties); third, you may have overriding reasons to go to the party, such as talking to a friend who is going through a difficult divorce; and fourth, any number of other situational influences present at the party might complicate the way you have individuated it (perhaps the presence at the party of a colleague recovering from lung cancer means that this is a situation that would in fact help you to quit smoking).

Each of these concerns applies to the use of the seek/avoid strategy to combat implicit bias. I myself had the following experience. Having spent the day hearing talks *at a workshop on implicit bias*, I was excited to unwind over a good meal with the rest of the conference attendees. Unbeknownst to the conference organizers, the meal at the chosen restaurant included a belly-dancing performance. This particular performance struck me (and others, I think) as uncomfortably suffused with familiar and problematic gender associations. Just imagine the optics: in a restaurant full of buttoned-up academics, a scantily dressed woman circles the tables, symbolically prostrating herself in front of people who pay her some—but not much—attention while they eat. This would seem to be a good situation to avoid if one is trying to combat implicit associations between the concept of "sexual object" and women. However, first, no one could have known ahead of time that this situation was one to avoid. Second, there was some sense of professional obligation to attend the conference dinner, so even if one wanted to avoid the situation, doing so would have come at some cost. Third, while recognizing the problematic nature of the situation, some of the workshop participants took themselves to have an obligation to show solidarity with the belly dancer, and decided to dance with her instead of trying to avoid her. And, fourth, the situation (for me at least) became difficult to individuate, since the exposure to gender depicted this way (presumably) reinforced stereotypical associations in me, yet exposure to my colleagues' unexpected show of solidarity arguably had the opposite effect, helping me to think of new ways to interact with stereotyped individuals. The seek/avoid strategy is no doubt useful at times, but its usefulness is limited.

The context-dependence of implicit bias renders agent-centered self-regulatory ideals incomplete, at least to the extent that they motivate a singular focus on the relationship between states internal to agents. Some outwardly-centered ethical ideals, such as the world-first and seek/avoid strategies, are available and valuable, but each has significant limitations. What's left?

§5. Self-Regulation via Context

In what follows I will discuss three self-regulatory strategies that take as a starting point the relationship between context and implicit attitudes. The first focuses on using seemingly minor elements of context to produce renewal effects of desirable behavior; the second on using specified cues across situations as ways to "off-load" the regulation of one's behavior to the environment; and

the third on conceptualizing our own behavior as an element of *others'* context. What holds these strategies together is that they promote egalitarian behavior in virtue of cultivating the right sort of ambient relationships between agents and their environments. I propose these examples as case studies in the ethics of ambient harmony, a fuller articulation of which I plan to purse in future work.

§5.1 Desirable renewal effects

In the same paper discussed above, Gawronski and Cesario (2013) suggest that context plays a key role in the regulation of implicit attitudes. While the literature they discuss emphasizes the return of undesirable, stereotype-consistent attitudes in ABA, AAB, and ABC patterns, they also discuss patterns of context-change in which participants' ultimate evaluations of targets reflect the counter-conditioning information they learned in the second block of training. These are the ABB and AAA patterns. The practical upshot of this, Gawronski and Cesario suggest, is that one ought to learn counter-stereotyping information in the same context in which one aims to be unbiased (ABB and AAA renewal). And what counts as "the same" context can be specified. Gawronski and colleagues (in prep.) show that renewal effects are more responsive to the perceptual similarity of contexts than they are to conceptual identity or equivalence. So it is better, for example, to learn counter-stereotyping information in contexts that look like one's familiar environs than it is to learn them in contexts that one recognizes to be conceptually equivalent. It may matter less that a de-biasing intervention aimed at classroom interactions is administered in another "classroom," for example, than that it is administered in another room that is painted the same color as one's usual classroom. Finally, Gawronski and Cesario suggest that if it is not possible to learn counter-stereotyping interventions in the same or similar context in which one aims to be unbiased, one ought to learn counter-stereotyping interventions across a variety of contexts. This is because both ABA and ABC renewal is weaker when counter-attitudinal information is presented across a variety of contexts, rather than just one. The reason for this is that fewer contextual cues are incorporated into the agent's representation of the counter-attitudinal information when the "B" context is varied. This signals to the agent that the counter-attitudinal information generalizes to novel contexts.

All of these are new findings, but they suggest that the self-regulation of implicit bias is neither only a matter of getting one's internal states into the right relationship, nor is it only a matter of changing the situations we are in. Rather, it is a matter of getting oneself into the right relationship *with* one's context.

§5.2 Off-loading the regulation of behavior to the environment

Implementation intentions are "if-then" plans that appear to be a remarkably effective strategy for achieving a wide range of goals. An implementation intention specifies a goal-directed response

which an individual plans to perform when she encounters an anticipated cue. In the usual experimental scenario, participants who hold a goal, "I want to X!" (e.g. "I want to eat healthy!") are asked to supplement their goal with a plan of the form, "And if I encounter opportunity Y, then I will perform goal-directed response Z!" (e.g. "And if I get take-out tonight, then I will order something with lots of vegetables!") Forming a plan to implement one's goal in this specific conditional format significantly improves self-regulation in a wide variety of domains, including (to name just of few of what is a long, long list) dieting, exercising, recycling, restraining impulses, maintaining focus (e.g. in athletics), avoiding binge drinking, and performing well on memory, arithmetic and Stroop tasks (Gollwitzer and Sheeran, 2006). If-then planning has also been shown to be effective in the regulation of biased implicit attitudes.

For example, if-then planning has been shown to be effective in reducing bias on IAT scores (Webb et al., 2010), a weapons identification task (Stewart and Payne, 2008), and a shooter-bias task (Mendoza et al, 2010). Mendoza and colleagues instructed participants in the implementation intention condition to adopt the plan, "and if I see a gun, then I will shoot!" A simple plan like this is thought to work by increasing the accessibility of cues relevant to a particular goal and automatizing the intended behavioral response. Here the relevant cue—"gun"—is made more accessible—and the intended response—to shoot when one sees a gun—is made more automatic. Gollwitzer et al. (2008) explain this cue-response link in associative terms. They write, "an implementation intention produces automaticity immediately through the willful act of creating an association between the critical situation and the goal-directed response" (326). It is this link between *the critical situation* and one's behavioral response that is crucial for broader ethical questions. It suggests that if-then planning represents a "context-first" approach to self-regulation. In order to attain one's egalitarian goals, one focuses one's attention outward, away from oneself, toward the context cues that act as instigators for goal-consistent behaviors. Gollwitzer (1993, 173) puts it this way: "by forming implementation intentions people pass the control of their behavior on to the environment."

§5.3 Be part of the situation

Finally, one can form a productive ethical strategy for combatting implicit biases by conceptualizing one's own behavior as an element of *others'* context. Sarkissian advances this strategy in response to the situationist critique of virtue ethics. He writes (2010, 12, emphasis in original):

> We hardly notice it, but oftentimes a kind smile from a friend, a playful wink from a stranger, or a meaningful handshake from a supportive colleague can completely change our attitudes. Such minor acts can have great effects. If we mind them, we can foster a form of *ethical bootstrapping* — that is, we can prompt or lift one another toward our joint moral ends. If situationism is true, then whether any individual will be able to meet her ethical aims on

> any particular occasion will hinge on the actions and manners of others in her presence, which in turn will hinge on her own. In being mindful of the interconnectedness of our behavior, we not only affect how others react to us, but also thereby affect the kinds of reactions we face with in turn. The bootstrapping is mutual.

Seemingly minor things that we do in the presence of others, in other words, help to form the context that shapes how others behave, which in turn affects us. Sarkissian takes advice from Confucian ethics in determining which "minor things" we can control which in turn may have ethical bootstrapping effects. These include mannerisms, tone of voice, and posture, each of which is a source of "*de*," or moral power or moral charisma (Sarkissian, 2010, 9). Empirical literature also supports Sarkissian's point; he cites literature showing that smiling, winking, and handshaking increases trust and cooperation between strangers (Scharlemann et al., 2001 and Manzini et al, 2009).

Shaping one's interpersonal context by attending to mannerisms, tone of voice, posture, etc. is particularly valuable in the case of implicit bias because these seemingly minor behaviors are precisely the medium through which implicit bias are often expressed. Such so-called "micro-expressions" of prejudice involve, for example, making more eye contact with white colleagues than black colleagues during a meeting or referring to male scholars by their last names and female scholars by their first names.[16] Attending to these micro-expressions of prejudice is certainly in keeping with the ethics of internal harmony. But the proximal goal motivating one's attention is external to the agent as well. The proximal goal is the creation of a certain interpersonal atmosphere, so to speak. The regulation of one's own internal states comes about as happy by-product of the aim to create this mutually supportive atmosphere.

§6. Conclusion

Implicit biases are not "directly" activated or expressed in behavior by critical cues or mere cultural knowledge of stereotypes. Rather, implicit biases are highly context-dependent. This fact has ethical ramifications. In particular, it points to the need to supplement or amend the ethics of internal harmony with an outward-focused ethics of ambient harmony. The options comprising an outward-focused ethics are not only changing the world or avoiding or seeking out particular situations. Rather, ethical regulation of one's implicit biases necessarily includes putting oneself into the right relationship with one's context. Doing so appropriately incorporates the metaphysical complexities of implicit social cognition into our ethical responses to it.

[16] On bias and microbehavior, see Dovidio et al. (2002); Valian (1998, 2005); Cortina et al. (2008, 2011); and Brennan (forthcoming).

Works Cited

Barden, J., Maddux, W., Petty, R., and Brewer, M. 2004. Contextual Moderation of Racial Bias: The Impact of Social Roles on Controlled and Automatically Activated Attitudes. *Journal of Personality and Social Psychology* 87:1, 5-22.

Bargh, J. A., Chen, M., and Burrows, L. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology* 71, 230-244.

Brennan, S. Forthcoming. Rethinking the Moral Significance of Micro-Inequities: The Case of Women in Philosophy. In Jenkins, F. and Hutchinson, K. (Eds.) *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press.

Brownstein, M. and Madva, A. 2012. Ethical Automaticity. *Philosophy of the Social Sciences*, 42:1, 67-97.

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*.

Darley, J., and Batson, C. 1973. From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology* 27, 100–108.

Dasgupta, N., DeSteno, D., Williams, L.A, and Hunsinger, M. 2009. Fanning the Flames of Prejudice: The Influence of Specific Incidental Emotions on Implicit Prejudice. *Emotion* 9:4, 585-591.

Devine, P. G. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* 56, 5–18.

Dijksterhuis, A. and Bargh, J. A. 2001. The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology* 33, 1-40.

Doris, J. 2002. Lack of character: Personality and moral behavior. Cambridge: Cambridge University Press.

Dovidio, J. F., and Gaertner, S. L. 2000. Aversive racism and selection decisions. Psychological Science 11, 319–323

Dovidio, J. F., and Gaertner, S. L. 2004. Aversive racism. In Advances in experimental social psychology 36, 1-51. M. P. Zanna (Ed.). San Diego, CA: Academic Press

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 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 Cesario, J. 2013. Of Mice and Men: What Animal Research Can Tell Us about Context Effects on Automatic Response in Humans. *Personality and Social Psychology Review* 17:2, 187-215.

Gawronski, B., Rydell, R.J., Vervliet, B., and de Houwer, J. 2010. Generalization Versus Contextualization in Automatic Evaluation. *Journal of Experimental Psychology* 139:4, 683-701.

Gawronski, B., Rydell, R. J., Ye, Y., and De Houwer, J. In preparation. Contextualized representation.

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

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.

Gollwitzer, P. 1993. Goal achievement: The role of intentions. *European Review of Social Psychology* 4, 141-185.

Gollwitzer, P., and Sheeran, P. 2006. Implementation intentions and goal achievement: A meta-analysis of effects and processes. In *Advances in experimental social psychology*, M.P. Zanna (Ed.), 69-119). US: Academic Press.

Gollwitzer, P., Parks-Stamm, E., Jaudas, A., and Sheeran, P. 2008. Flexible Tenacity in Goal Pursuit. In Shah, J. and Gardner, W. (Eds.). 2007. *Handbook of motivation science*. New York: Guilford Press.

Green, J., Sedikides, C., Saltzberg, J., Wood, J., and Forzano, L. 2003. Happy mood decreases self-focused attention. *British Journal of Social Psychology* 28, 147–157.

Guinote, A., Guillermo, B.W., and Martellotta, C. 2010. Social power increases implicit prejudice. *Journal of Experimental Social Psychology* 46, 299-307.

Harman, G. 1999. Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society* 99, 315–331.

Harman, G. 2003. No character or personality. *Business Ethics Quarterly*, 13:1, 87–94.

Haslanger, S. This volume. Schemas and the Materiality of Social Practices.

Holroyd, J. and Sweetman, J. This volume. The Heterogeneity of Implicit Bias.

Huebner, B. 2009. Trouble with Stereotypes for Spinozan Minds. Philosophy of the Social Sciences 39, 63-92.

Isen, A. and Levin, P. 1972. Effect of Feeling Good on Helping: Cookies and Kindness. *Journal of Personalioty and Social Psychology* 21:3, 384-388/.

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.

Machery, E. This volume. DeFreuding Implicit Attitudes.

Machery, E., Faucher, L., and Kelly, D. 2010. On the Alleged Inadequacies of Psychological Explanations of Racism. *The Monist* 93:2, 228-254.

Madva, A. 2012. The Hidden Mechanisms of Prejudice: Implicit Bias and Interpersonal Fluency. PhD dissertation.

Madva, A. and Brownstein, M. In preparation. Implicit Stereotyping and Evaluation are Not Independent Processes.

Manzini, P., Sadrieh, A., and Vriend, N. 2009. On smiles, winks and handshakes as coordination devices. *The Economic Journal*, 119:537, 826–854.

Mendoza, S.A., Gollwitzer, P.M., and Amodio, D.M. 2010. Reducing the Expression of Implicit Stereotypes: Reflexive Control Through Implementation Intentions. *Personality and Social Psychology Bulletin* 36:4, 512-523.

Milgram, S. 1974/2009. *Obedience to Authority*. New York: Harper and Row.

Mill, J.S. 1873/1944. *Autobiography*. New York: Columbia University Press.

Mitchell, J. P., Nosek, B. A., and Banaji, M. R. 2003. Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General* 132, 455-469.

Neal, D.T, Wood, W., Wu, M. and Kurlander, D. 2011. The Pull of the Past: When Do Habits Persist Despite Conflict With Motives? *Personality and Social Psychology Bulletin*, 1-10. doi: 10.1177/0146167211419863

Nosek, B.A., and Hansen, J.J. 2008. The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion* 22:4, 553-594.

Payne, B.K and Cameron, C.D. 2010. Divided Minds, Divided Morals: How Implicit Social Cognition Underpins and Undermines Our Sense of Justice. In B. Gawronski and B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, NY: Guilford Press.

Richeson, J.A. and Ambady, N. 2001. Who's in charge? Effects of situational roles on automatic gender bias. *Sex Roles* 44, 493-512.

Robbins, T. 1980. *Still Life with Woodpecker*. New York: Bantam Books.

Rydell, R. J. and Gawronski, B. 2009. I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion* 23, 1118-1152.

Sarkissian, H. 2010. Minor Tweaks, Major Payoffs: The Problems and Promise of Situationism in Moral Philosophy. *Philosophers' Imprint* 10:9, 1-15.

Schaller, M., Park, J.J., and Mueller, A. 2003. Fear of the dark: Interactive effects of beliefs about danger and ambient darkness on ethnic stereotypes. *Personality and Social Psychology Bulletin* 29, 637-649.

Scharlemann, J., Eckel, C., Kacelnik, A., and Wilson, R. 2001. The value of a smile: game theory with a human face. *Journal of Economic Psychology* 22, 617–640.

Schmajuk, N. A. and Holland, P. C. 1998. *Occasion setting: Associative learning and cognition in animals*. Washington, DC: American Psychological Association.

Schooler, J., Ariely, D., and Loewenstein, G. 2003. The pursuit and assessment of happiness may be self-defeating. In Brocas, I. and Carrillo, J. (Eds.). T*he psychology of economic decisions: Vol. 1. Rationality and well-being*. New York: Oxford University Press.

Stewart, B. and Payne, K. 2008. Bringing Automatic Stereotyping Under Control: Implementation Intentions as Efficient Means of Thought Control. *Personality and Social Psychology Bulletin* 34:10, 1332-1345.

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

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

Webb, T., Sheeran, P. and Pepper, A. 2010. Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. *British Journal of Social Psychology*, DOI:10.1348/014466610X532192