

3.2

Context and the Ethics of Implicit Bias

Michael Brownstein

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, John 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 attitudes and to cause one to act in biased ways.¹ In a related vein, Patricia Devine (1989) argued in a seminal paper 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 significantly affect the activation and expression in behavior of implicit attitudes. 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 have focused on the metaphysical ramifications of this fact; in particular, whether implicit attitudes represent a singular kind, or rather, whether the implicit attitude construct stands for a number of related (or perhaps

¹ See also Dijksterhuis and Bargh (2001).

² See Nosek and Hansen (2008) for discussion.

unrelated) cognitive and affective processes.³ In this chapter I focus on a separate philosophical issue stemming from findings about the effects of context on implicit attitudes. 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 attitudes and behavior. This notion, of cultivating the right sort of “ambient” relationships, has been underdescribed 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 (Section 2). I will then briefly describe the predominant ways of thinking about the ethics of implicit bias in the philosophical literature: the “ethics of internal harmony” (Section 3.1); the “world-first” strategy (Section 3.2); and the “seek/avoid” strategy (Section 3.3). Each of these conceptualizations reflects a crucial facet of the ethics of implicit bias, but each on its own is limited. I go on to describe a “contextualist” approach which incorporates elements of each of these by focusing on the relationship between agents’ internal states and their ambient environment and situations (Section 4). The “nodes” of this relationship between agents and their environments are various forms of contextual cues, I argue. In focusing on these cues, agents can start on the road toward creating the kind of world that promotes ethical thought and action within themselves and others.

2 Context and Implicit Bias

In a recent paper, Bertram Gawronski and Joseph 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 counterconditioned 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.

³ See, for instance, Machery (Volume I); Holroyd and Sweetman (Volume I); Madva and Brownstein (ms.); and discussion between Mazarin Banaji and Tamar Gendler on “The Mind Report.” <<http://bloggingheads.tv/videos/15811>>.

For example, a person might learn biased associations while hanging out with their friends (context A), effectively learn counterstereotypical 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 in controlled lab settings renewal effects like these in implicit attitudes. 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 (i.e. evaluations consistent with the valence of the information that was presented first) in the patterns 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).

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 attitudes. Context is, of course, a broad notion. Imagine taking an Implicit Association Test (IAT; Greenwald et al., 1998). The background color against which the images of the target subjects are presented is part of the context. Or perhaps before administering 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 the discussion later in this section), 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 drunk too much caffeine, and so on...⁴

⁴ Perhaps it would be better to refer to all of these various elements as “situational factors” rather than elements of context. See Section 3.3 for discussion.

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, or when feeling a certain way. 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 attitudes. 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 attitudes primarily in light of their visual, aural, or other sensory properties. For example, Mark Schaller and colleagues (2003) found that the relative darkness or lightness of the room in which participants sit shifts scores of implicit racial evaluations across several indirect measures, including the IAT.⁶ The sort of renewal effects which 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 attitudes primarily in light of perceived group membership and social roles. Jamie Barden and colleagues (2004), for example, varied the category 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, Jason Mitchell and colleagues (2003) showed that implicit evaluations of the same individual—Michael Jordan—depended on whether he was categorized by race or occupation. Conceptual elements of context also include one’s own social role. Ana 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; Payne et al., 2005) than those in a low-power condition,

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.

who were led to believe that their opinions would not affect the committee's decisions.⁷

Finally, motivational elements of context influence implicit attitudes primarily in light of fluctuations in mood and emotion. Nilanjana Dasgupta and colleagues (2009), for instance, found that salient emotions selectively influence the activation of implicit attitudes. 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 In a related vein, Jennifer Kubota and Tiffany Ito (2014) recently showed that the emotional expression of others' faces (e.g. smiling) moderates the activation of black-danger stereotypes.

These data should not be surprising. For example, 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 not getting enough sleep.

3 Three Ethics of Implicit Bias

What follows are three ways of thinking about the ethics of implicit bias. Each is connected to the others and is independently plausible. However, each on its own

⁷ In the case of implicit gender bias, Richeson and Ambady (2001) 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. This result is interesting to compare to Guinote and colleagues' findings. Both studies find effects of one's perceived status on one's implicit intergroup attitudes; but in the case of race, perceiving oneself in a superordinate status seems to amplify bias, while in the case of gender, perceiving oneself in a subordinate status seems to amplify bias.

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.

faces significant theoretical and practical challenges. The way forward includes elements of each and connects them by focusing on the concept of context, to which I return in Section 4.

3.1 *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 do not endorse the validity of those very biases (Gawronski and Bodenhausen, 2006; Nosek and Hansen, 2008). They can be aversive with respect to agents' reflective or moral commitments; in other words, giving rise to "aversive racism" (Dovidio and Gaertner, 2000, 2004). Philosophers writing on implicit bias have focused on this fact, and for good reason.⁹ 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, Tamar Szabó 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, 380 BCE/1992; quoted in Gendler, 2008b: 572).

This ideal of internal harmony—or something quite like it—is also found in the psychological literature. Keith Payne and Daryl 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 epigraph from

⁹ See, for instance, Gendler (2008a, 2008b); Madva (2012); Huebner (2009); and Kelly and Roedder (2008).

Rousseau: “Virtue is a state of war, and to live in it we have always to combat with ourselves.”

The “ethics of internal harmony” addresses at least one element of what is undeniably frightening about implicit bias: namely, that it can affect one’s own attitudes and behavior, even if one genuinely desires to be unbiased. However, the ethics of internal harmony faces challenges in representing an ethical response to the problem of implicit bias. First, the ideal of internal harmony is too permissive. A person with implicit biases, who is *also* explicitly biased, will (formally, at least) count as internally harmonious. Her ideals—as represented by her explicit attitudes—will line up with her implicit attitudes, as well as with those behaviors affected by her implicit attitudes. Minimally, this problem represents a *prima facie* challenge to the ethics of internal harmony.

More importantly, the ethics of internal harmony is predominantly *agent-centered*. It is agent-centered in the sense that it singularly recommends 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 “refocusing of attention through directed imagination.” While extremely valuable, both of these strategies focus exclusively on the agent, 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.

Others have articulated problems with an agent-centered ethics like this, problems having to do with the relative importance of social, institutional, and economic inequality (e.g. Huebner, 2009, Volume I; Anderson, 2010; Haslanger, 2015; Jacobson, this volume). I discuss this “institutional” critique of the self-regulation of implicit bias in Section 3.2. A related problem with an agent-centered ethics, not explicitly discussed in these critiques, is that it can have a framing effect on thought and action. For example, the overarching moral problem of implicit bias is not that it causes well-intentioned egalitarians to be internally disharmonious. The overarching moral problem is that implicit biases perpetuate injustice. The ethics of internal harmony is not formally inconsistent with this point, of course. One can aim to fight against injustice precisely by bringing one’s implicit attitudes and behavior in line with one’s reflective ends. But as an ethical

end unto itself, the ideal of internal harmony frames one's attention predominantly on oneself in such a way that it risks overshadowing the greater moral problem. For example, one might think one's ethical "work" is finished once one has reached a state of internal harmony. But in a world suffused with injustice, this would be an unsatisfying end-point for ethics. It does not address one's obligations to encourage one's peers to act in ethical ways, for instance. Nor does it commit anyone to changing the institutional structures that help to perpetuate injustice, such as non-anonymous review of student papers, job application materials, and so on.

3.2 *The world-first strategy*

Those who have stressed the importance of changing social, institutional, and economic inequalities in order to combat implicit bias (Huebner, 2009, Volume I; Anderson, 2010; Haslanger, 2015; Jacobson, this volume) have articulated what I call a "world-first" strategy. On this view, the task of an ethics of implicit bias is to change institutions, not individuals. Generally, the idea behind this argument is that prejudice and bias are sustained by material forms of inequality, and that so long as social institutions continue to promote these forms of inequality, no amount of self-regulation can successfully combat the pernicious effects of racism, sexism, and other prejudices on one's own behavior. Bryce Huebner (2009: 88), for example, argues that "the only way in which we will be able to adequately modify our psychology is by modifying the world in which we live."¹⁰ Huebner's view is that in a propaganda-filled world like ours, where even the most harmonious agent continues to be bombarded with stereotypes, any ethical strategy focused on making change at the individual level amounts to too little, too late.

The world-first strategy is appealing for several reasons. First, it connects ethical and moral concerns, by focusing one's attention "outward," off oneself and onto the world in which injustice is perpetrated. Second, I think proponents of the world-first strategy are correct in worrying that implicit biases may be relearned in social environments suffused with inequality.¹¹ Third, the world-first strategy serves as an important call for activism. Implicit bias is, of course, just one element of wider patterns of discrimination. Fighting to change housing laws, pay discrimination, racial profiling, and so on, can

¹⁰ Huebner does discuss some self-regulatory strategies for modifying one's psychology, but he is skeptical of their effectiveness without an attendant revolutionary politics. See also Dixon et al. (2012), who voice related worries about prejudice reduction and attitude change. For a defense of "psychological" approaches to combating implicit bias, see Machery et al. (2010).

¹¹ Although see Madva (ms.) for some important doubts about this "relearning" worry.

therefore serve the cause of combating implicit bias as well as other social-moral travesties.

There are weak and strong interpretations of the world-first strategy, however. On a weak interpretation, it suggests that one ought to *both* try to change one's own attitudes and try to change the world itself. On a strong interpretation, trying to change one's own attitudes through self-regulation is hopeless, because biases will inevitably be relearned in an unjust social world; or, trying to change one's own attitudes is problematically distracting from the "real" work of changing the social, institutional, and economic sources of bias and discrimination.

The weak interpretation of the world-first strategy is more appealing than the strong interpretations. One reason for this is that there are simply not yet enough data to confirm or deny the charge of hopelessness. Research on the self-regulation of implicit bias is in its relative infancy. There is reason at least to think that this research is promising.¹² Second, it is hard to see why changing one's own attitudes is not complementary to, rather than in conflict with, changing the world itself. And, third, pragmatically, no one can doubt that changing the world itself will take (a lot of) time. At least while this effort is ongoing, everyone individually simply has to cope with trying to be an ethical agent in an unjust world.

However, even the weak interpretation of the world-first strategy is, on its own, incomplete. The worry is that it does not speak to ethical concerns about personal responsibility for implicit bias. As several of the chapters of this volume suggest, it is a difficult and important question whether, and how, individuals ought to be held responsible for behavioral expressions of implicit bias. This question is no less pressing when directed toward oneself. What are *my* responsibilities, given the evidence that I likely hold and act upon implicit biases? On the strongest presentation of the worry, political activism can serve as a source of self-deception, leading one to think that one has discharged one's ethical obligations by "fighting the good (political) fight."¹³ It is hard to see how the world-first strategy speaks to these concerns. Perhaps it is complementary with them, but as an expression of what the ethics of implicit bias is all about, the world-first strategy is incomplete.

¹² For review, see Dasgupta (2013).

¹³ My point is not that any of the defenders of the world-first strategy are themselves self-deceived. The novelist Tom Robbins (2010) expressed my worry 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."

3.3 *The seek/avoid strategy*

A second kind of outwardly-focused ethics of implicit bias can be repurposed from the philosophical reception of the “situationist” literature in social psychology. Critics of virtue ethics like Gilbert Harman (1999) and John Doris (2002) have made clear how unexpected and seemingly trivial features of the situations that we are in can strongly affect our behavior and attitudes. It is not a stretch to think of context cues that affect implicit attitudes as examples of the kinds of situational influences on behavior with which these authors are concerned. 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 attitudes are analogous in many respects to other central situationist examples, such as Darley and Batson (1973) and Milgram (1974/2009).¹⁴

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. An obvious way to do so 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. Hagop Sarkissian (2010) calls this the “seek/avoid” strategy, which I here repurpose in the related context of combating implicit bias. 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 day-to-day life, 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 themselves require one to enter compromising situations. And fourth, the problematic variables inherent in ordinary situations are so finely individuated that it is hard to know how agents could ever

¹⁴ Nothing I say in this chapter 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 ethics.

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 may 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 conceptualized 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 had the following experience. Having spent the day hearing talks *at a workshop on implicit bias*, I was excited to unwind over dinner with my colleagues. 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, the seek/avoid strategy would be little help here. First, no one could have reasonably 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, for some of the workshop participants, this was not even a situation to avoid. Rather than ignore the dancer, they took this as an opportunity to express their ethical commitments by showing solidarity with a hard-working woman. And so they praised her great skill and—upon being invited—climbed onto the dinner table and danced with her! I took this to be the enactment of an ethical ideal that required entering an otherwise potentially compromising situation (i.e. an illustration of Sarkissian’s third point, above). Finally, these colleagues who showed solidarity with the dancer changed the meaning of the experience, for me at least. Rather than see it as a straightforwardly compromising situation, I now think of it as an illustration of how to be creatively ethical. This illustrates Sarkissian’s fourth point: the variables that determine whether a situation is likely to compromise or promote our ethical ends are extremely hard to individuate and identify.

4 A Contextualist Approach

My proposal is modest. It is a reframing of the way we think about the ethics of implicit bias, and it borrows from each of the proposals I discussed previously. When conceptualizing the fight against implicit bias, our proximal focus should not be on harmonizing our internal states alone, nor should it be on changing the world in a broad sense, nor should it be on seeking out the right kinds of situations and avoiding the wrong ones as such. What I propose instead is a contextualist approach that blends all three from the get-go. It focuses on precisely those nodes at which our attitudes are affected by features of the ambient environment, and the ambient environment is in turn shaped by our attitudes and behavior.¹⁵ I will give three examples of what I mean, and in each case I will try to clarify why I take the example to illustrate a contextualist approach to the ethics of implicit bias.

4.1 *Physical context cues and desirable renewal effects*

In the paper discussed previously, Gawronski and Cesario (2013) suggest that physical context cues can play an important 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 counterconditioning 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 counterstereotyping 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 empirically specified. Gawronski and colleagues (ms.) 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 counterstereotyping 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 debiasing 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.¹⁶ Finally, Gawronski and

¹⁵ This is, of course, not tantamount to a metaphysical claim about the boundaries between agents' minds and the outer world.

¹⁶ Of course, one cannot always make the relevant predictions, such as the wall-color of the classrooms. An anonymous reviewer raises the worry that this means that contextualism will fall prey to the same worries I raised about the seek/avoid strategy (Section 3.3). In some cases this is

Cesario suggest that if it is not possible to learn counterstereotyping interventions in contexts the same as or similar to those in which one aims to be unbiased, one ought to learn counterstereotyping interventions across a variety of contexts. This is because both ABA and ABC renewal are weaker when counterattitudinal information is presented across a variety of contexts, rather than just one. The reason for this is thought to be that fewer contextual cues are incorporated into the agent's representation of the counterattitudinal information when the "B" context is varied. A greater variety of contexts signals to the agent that the counterattitudinal information generalizes to novel contexts.

These are new findings, in need of further consideration, but they are promising. I take them to be contextualist in nature because they are focused on the fine-grained stimuli that act as occasion-setters for wanted behavior.¹⁷ Importantly, a strategy like this is not predicted by the foregoing ways of thinking about the ethics of implicit bias. The importance of attending to the color of the room in which one practices a debiasing procedure is not clearly predicted by the ethics of internal harmony, since the color of the room is a seemingly random feature of the environment from the perspective of one's reflective goals. Nor is this strategy clearly predicted by the world-first or seek/avoid strategies. It is by no means a form of revolutionary political activism, in which we reshape the social, institutional, and economic forces that perpetuate inequality. And by changing the color of the room, one is neither seeking out nor avoiding any particular kind of

true, since we simply cannot predict the future. In addition, it may not be feasible to fix the relevant context cues in the right way, just as it is not always practically feasible to avoid compromising situations. The difference between contexts, as I discuss them here, and situations, in the sense of the seek/avoid strategy, is one of scale. The seek/avoid strategy recommends seeking out good situations and avoiding bad situations wholesale. Do not go to the bar, for example, if you are trying not to drink. Manipulating physical context cues such as the color of the room in which one practices a debiasing procedure, by contrast, is a way of re-engineering small features of the situations in which we know we will be. To put that another way: while we cannot always predict what the situation will be like, we can predict which contextual elements of familiar situations will have wanted and unwanted effects on our behavior. This is precisely the kind of empirical prediction Gawronski and colleagues' research investigates. Future research might investigate other elements of context that might moderate the expression of implicit biases. Air temperature? Ambient noise? Moreover, how do these features of the context interact? Would a noticeably warm classroom elicit renewal effects despite similarities in the color of the walls? Would renewal effects be stronger if the room has the same color walls and is also a similar temperature to the rooms in which one ultimately interacts outside the lab?

¹⁷ An anonymous reviewer asks how fine-grained these contextual features can be before they become impossible to utilize outside the lab. Not very. We cannot often control the color of walls, air temperature, ambient noise, and so on. But the example I am discussing involves a debiasing procedure performed in the lab, which then has lasting effects outside the lab (hopefully). Certainly, these features of the context can be carefully controlled in the lab.

situation. Rather, one is seeking out small elements of situations and thereby changing the significance of the situation itself.

Promoting desirable renewal effects in this way is a relatively minor tool, all things considered, but it is one best conceived through a proximal focus on one's context as a node of exchange between agents and their ambient environment.

4.2 Behavioral context cues and if-then planning

Implementation intentions are "if-then" plans that appear to be remarkably effective for promoting a wide range of goals. An implementation intention specifies a goal-directed response that an individual plans to perform on encountering 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 a 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). Saaid Mendoza and colleagues, for example, instructed participants in an 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. Peter Gollwitzer and colleagues (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 illustrating why if-then planning is a tool of the kind of contextualist ethics I am recommending. The heavy-lifting in if-then planning is done by the specification of the situational cue to which one plans to respond (i.e. whatever

follows the premise in the “if” clause). Indeed, the effectiveness of this form of planning is strongly moderated by the accessibility of the specified cue, as well as the strength of association between the cue and the behavioral response (Webb and Sheeran, 2008). In order to attain one’s egalitarian goals, then, one can focus one’s attention outward, away from oneself, toward the context cues that act as instigators for goal-consistent behaviors. Gollwitzer (1993: 173) expresses it this way: “...by forming implementation intentions people pass the control of their behavior on to the environment.” This formulation may be too strong, if passing the control of one’s behavior to the environment is thought to entail a loss of agency. But I suspect that what Gollwitzer means is that by adopting an implementation intention one automatizes one’s goal-striving, and one does so by opening oneself to the effects of particular contexts on one’s attitudes and behavior. On this reading, if-then planning represents not just a valuable tool for self-regulation as such, but also represents an example of self-regulation via the strategic arrangement of one’s behavior-inducing ambient relations.

As in the previous example, if-then planning is not wholly divorced from the effort to be more internally harmonious, to change the world, or to seek out good situations and avoid bad ones. But what makes if-then planning powerful is not predicted by any of these approaches. It can help to make one more internally harmonious, but only by directing one’s attention away from one’s internal states. Perhaps if-then planning has the power to change the world, but only by changing the behavior of individuals, one by one. And if-then planning does not recommend seeking out or avoiding situations wholesale, but rather, attending to the crucial features of those situations that promote desirable behavior. In other words, it recommends attending to context.

4.3 *Social context cues and reciprocal bootstrapping*

A final example is the combating of 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.¹⁸

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 can have ethical bootstrapping effects. These include mannerisms, tone of voice, and posture, each of which is a source of “*de*,” or moral charisma (Sarkissian, 2010: 9). Empirical literature also supports Sarkissian’s point; he cites literature showing that smiling and handshaking increase trust and cooperation between strangers (Scharlemann et al., 2001; Manzini et al., 2009).

Shaping one’s interpersonal context by attending to mannerisms, tone of voice, posture, and so on, is particularly valuable in the case of implicit bias because these seemingly minor behaviors are precisely the medium through which implicit bias is often expressed. Such so-called “micro-expressions” of prejudice involve, for example, making more eye contact with white colleagues than with black colleagues during a meeting, or referring to male scholars by their last names and female scholars by their first names.¹⁹

Attending to these micro-expressions of prejudice is certainly in keeping with the ethics of internal harmony, as the end result is hopefully the harmonization of one’s internal states. But, again, the proximal goal motivating one’s attention in this case is external to the agent. The (hopeful) result as well is not just the harmonization of one’s own internal states, but the interpersonal harmonization of one’s attitudes and behavior with those of others. Reciprocal bootstrapping in this sense aims to create a harmonious interpersonal atmosphere, so to speak. The regulation of one’s own internal states can even be seen as a felicitous byproduct of the aim to create this mutually supportive atmosphere. One can conceptualize this changed atmosphere as a changed world, but not in the institutional sense meant by the world-first critique. Similarly, by adopting this form of contextualism, one certainly minds the effects of the situation on one’s behavior, but not simply by seeking out and avoiding particular situations. Rather, one’s behaviors help to constitute the situations themselves, as they bear on others, and ultimately on oneself too.

¹⁸ While I endorse giving others kind smiles and meaningful handshakes, I think one should probably avoid winking at strangers.

¹⁹ On bias and microbehavior, see Brennan (2013, this volume); Cortina (2008); Cortina et al. (2011); Dovidio et al. (2002); Olberding (2014); Valian (1998, 2005).

5 Conclusion

Implicit biases are not “directly” activated or expressed in behavior by mere cultural knowledge of stereotypes. Rather, they are highly context-dependent. This fact has ethical ramifications. In particular, it points to the need to amend our usual ways of conceptualizing the ethics of implicit bias with an outward-focused contextualist ethics, according to which agents and their environments are deeply connected. A contextualist ethics of implicit bias focuses on putting oneself into the right relationship with one’s context and thereby helping to create the kind of environment that promotes ethical thought and action. Doing so appropriately incorporates the metaphysical complexities of implicit social cognition into our ethical responses to it.

Acknowledgments

I am very grateful for feedback on this chapter from Asia Ferrin, Bryce Huebner, Daniel Kelly, Alex Madva, Kenny Marotta, Jennifer Saul, Natalia Washington, and two anonymous referees for Oxford University Press. Versions of this chapter were presented at the 2013 meeting of the Society for Philosophy and Psychology and the 2013 meeting of the Eastern Division Meeting of the American Philosophical Association. I benefited from many helpful comments and questions at these meetings.

References

- Anderson, E. (2010). *The Imperative of Integration*. Princeton, NJ: Princeton University Press.
- 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: 230–44.
- Brennan, S. (2013). “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: 180–96.
- Brennan, S. (this volume). “The moral status of micro-inequities: In favor of institutional solutions.”
- 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* 39(6): 1579–605.

- 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: 100–8.
- Dasgupta, N. (2013). "Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept." *Advances in Experimental Social Psychology* 47: 233–79.
- 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): 585–91.
- 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.
- Dixon, J., Levine, M., Reicher, S., and Durrheim, K. (2012). "Beyond prejudice: are negative evaluations the problem and is getting us to like one another more the solution?" *Behavioral and Brain Sciences* 35(6): 411–25.
- 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: 319–23.
- Dovidio, J. F. and Gaertner, S. L. (2004). "Aversive racism." In Zanna M. P. (ed.), *Advances in Experimental Social Psychology*, vol. 36. San Diego, CA: Academic Press: 1–51.
- Dovidio, J. F., Kawakami, K., and Gaertner, S. L. (2002). "Implicit and explicit prejudice and interracial interaction." *Journal of Personality and Social Psychology* 82: 62–8.
- 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): 745–50.
- 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. (ms.). "Contextualized representation."
- Gawronski, B. and Sritharan, R. (2010). "Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures." In Gawronski, B. and Payne, B. K. (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): 634–63.
- Gendler, T. S. (2008b). "Alief in action (and reaction)." *Mind and Language* 23(5): 552–85.

- Gollwitzer, P. (1993). "Goal achievement: The role of intentions." *European Review of Social Psychology* 4: 141–85.
- Gollwitzer, P., Parks-Stamm, E., Jaudas, A., and Sheeran, P. (2008). "Flexible tenacity in goal pursuit." In Shah, J. and Gardner, W. (eds.), *Handbook of Motivation Science*. New York, NY: Guilford Press: 325–41.
- Gollwitzer, P. and Sheeran, P. (2006). "Implementation intentions and goal achievement: A meta-analysis of effects and processes." In Zanna M. P. (ed.), *Advances in Experimental Social Psychology*. San Diego, CA: Academic Press: 69–119.
- Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). "Measuring individual differences in implicit cognition: The implicit association test." *Journal of Personality and Social Psychology* 74: 1464–80.
- 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: 315–31.
- Harman, G. (2003). "No character or personality." *Business Ethics Quarterly* 13(1): 87–94.
- Haslanger, S. (2015) "Social structure, narrative and explanation." *Canadian Journal of Philosophy*. DOI: 10.1080/00455091.2015.1019176.
- 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.
- Huebner, B. (this volume). "Implicit bias, reinforcement learning, and scaffolded moral cognition."
- Isen, A. and Levin, P. (1972). "Effect of feeling good on helping: Cookies and kindness." *Journal of Personality and Social Psychology* 21(3): 384–8.
- Jacobson, A. (this volume). "Reducing racial bias: Attitudinal and institutional change."
- Kelly, D. and Roedder, E. (2008). "Racial cognition and the ethics of implicit bias." *Philosophy Compass* 3(3): 522–40. doi:10.1111/j.1747-9991.2008.00138.x.
- Kubota, J. and Ito, T. (2014). "The role of expression and race in weapons identification." *Emotion* 14(6): 1115–24.
- Machery, E. (this volume). "De-Freuding implicit attitudes."
- Machery, E., Faucher, L., and Kelly, D. (2010). "On the alleged inadequacies of psychological explanations of racism." *The Monist* 93(2): 228–54.
- Madva, A. (2012). "The hidden mechanisms of prejudice: Implicit bias and interpersonal fluency. PhD dissertation.
- Madva, A. (ms.). "Biased against de-biasing: On the role of (institutionally sponsored) self-transformation in the struggle Against Prejudice."
- Madva, A. and Brownstein, M. (ms.). "The blurry boundary between stereotyping and evaluation in implicit cognition."
- Manzini, P., Sadrieh, A., and Vriend, N. (2009). "On smiles, winks and handshakes as coordination devices." *The Economic Journal* 119: 537, 826–54.
- 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): 512–23.
- Milgram, S. (1974/2009). *Obedience to Authority*. New York, NY: Harper and Row.

- Mitchell, J. P., Nosek, B. A., and Banaji, M. R. (2003). "Contextual variations in implicit evaluation." *Journal of Experimental Psychology: General* 132: 455–69.
- 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* 37: 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): 553–94.
- Olberding, A. (2014). "Subclinical bias, manners, and moral harm." *Hypatia* 29(2): 287–302.
- Payne, B. K. and Cameron, C. D. (2010). "Divided minds, divided morals: How implicit social cognition underpins and undermines our sense of justice. In Gawronski, B. and Payne, B. K. (eds.), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York, NY: Guilford Press: 1–18.
- Payne, B., Cheng, C. M., Govorun, O., and Stewart, B. (2005). "An inkblot for attitudes: Affect misattribution as implicit measurement." *Journal of Personality and Social Psychology* 89: 277–93.
- 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: 1118–52.
- 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: 637–49.
- 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: 617–40.
- Schmajuk, N. A. and Holland, P. C. (1998). *Occasion Setting: Associative Learning and Cognition in Animals*. Washington, DC: American Psychological Association.
- 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): 1332–45.
- Valian, V. (1998). *Why so Slow? The Advancement of Women*. Cambridge, MA: MIT Press.
- Valian, V. (2005). "Beyond gender schemas: Improving the advancement of women in academia." *Hypatia* 20: 198–213.
- Webb, T. and Sheeran, P. (2008). "Mechanisms of implementation intention effects: The role of goal intentions, self-efficacy, and accessibility of plan components." *British Journal of Social Psychology* 47: 373–9.
- 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* 51(1): 13–32. doi:10.1348/014666610X532192.