Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Biased against Debiasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle against Prejudice

ALEX MADVA
California State Polytechnic University, Pomona

Research suggests that interventions involving extensive training or counterconditioning can reduce implicit prejudice and stereotyping, and even susceptibility to stereotype threat. This research is widely cited as providing an "existence proof" that certain entrenched social attitudes are capable of change, but is summarily dismissed—by philosophers, psychologists, and activists alike—as lacking direct, practical import for the broader struggle against prejudice, discrimination, and inequality. Criticisms of these *debiasing* procedures fall into three categories: concerns about empirical efficacy, about practical feasibility, and about the failure to appreciate the underlying structural-institutional nature of discrimination. I reply to these criticisms of debiasing, and argue that a comprehensive strategy for combating prejudice and discrimination should include a central role for training our biases away.

1. Introduction

More than a decade of research suggests that implicit biases can be transformed (or at least considerably weakened) by interventions that involve extensive training or counterconditioning. In particular, Kerry Kawakami and colleagues have demonstrated the benefits of:

> *counterstereotype training*, which involves repeatedly affirming counterstereotypes, for example, by responding "yes" to an image of a black person paired with the word "friendly," or by repeatedly pairing images of women with words like "powerful" and "courageous;" and

Contact: Alex Madva <alexmadva@gmail.com>

http://dx.doi.org/10.3998/ergo.12405314.0004.006

approach training, which involves practicing approach-oriented behaviors toward stigmatized words and images, for example, by moving toward or nodding one's head in response to images of black faces or Arab-Muslim names.

In addition to reducing bias on a variety of indirect measures, including the Implicit Association Test (IAT), these training procedures:

- reorient unreflective social behaviors, for example, by leading white and Asian participants to instinctively sit closer to a black interlocutor;
- make stereotypes less likely to come to mind and color judgment, for example, by reducing the likelihood that participants recommend hiring a man over an equally qualified woman; and
- reduce susceptibility to stereotype threat, for example, by improving women's performance on math tests even though they have just been reminded of pervasive stereotypes about gender and mathematical aptitude, and would otherwise feel anxiety and underperform.

While this research is often cited as providing a sort of "existence proof" that certain entrenched social attitudes are capable of change, it is summarily dismissed by social scientists, philosophers, and activists as lacking direct, practical import for the broader struggle against prejudice and discrimination. For example, David Schneider's 568-page opus on social cognition, *The Psychology of Stereotyping*, devotes only a single paragraph to this research on "retraining," concluding that, "Obviously, in everyday life people are not likely to get such deliberate training" (2004, 423).

I find the widespread dismissal of these training procedures puzzling. Implicit biases influence whom we trust and whom we ignore, whom we promote and for whom we vote. They affect interactions between teachers and students, doctors and patients, police and civilians, and lawyers and jurors. The stakes are high, then, not just for individuals in prominent gatekeeper positions, but for us all.[1] Given the myriad ways in which our biases can harm others and ourselves across many spheres of life, I believe that each of us has an obligation to take steps to reduce our biases—if we can. I take this obligation to be relatively uncontroversial, at least among those who agree that widespread prejudice and discrimination are major social ills. Even those who deny that we are directly morally responsible for controlling our implicit biases, such as Levy (2017),

1. See, e.g., Blair, Dasgupta, and Glaser (2015) for a review of "meaningful life outcomes" affected by implicit attitudes. For more on the ethical and political implications of implicit bias, see the contributions to Brownstein and Saul (2016). Thanks to an *Ergo* referee for urging me to clarify the high stakes even for individuals outside gatekeeping positions.

grant that we may at least be obligated to take steps to reduce our biases so long as doing so is reasonably within our power. These procedures promise to put bias reduction within our power. So why are they so readily written off?

There are a handful of frequently cited reasons, which are taken either to override the defeasible obligation to reduce our biases, or to suggest that procedures like Kawakami's will not help us meet that obligation. These reasons fall roughly into three categories: concerns about empirical efficacy, about practical feasibility, and about the failure to appreciate the underlying structural-institutional nature of discrimination.

(EMPIRICAL INEFFICACY) While skeptics grant that these interventions may be somewhat effective in artificial lab-based settings, they press two further empirical concerns.[2] First, many suspect that individuals will quickly "relearn" their biases after the debiasing procedure. I call this the relearning worry (Section 4). Second, many suspect that the effects of debiasing will hold only in highly specific contexts, e.g., being visible inside the lab but not outside it. I call this the context-specificity worry (Section 5).

(PRACTICAL UNFEASIBILITY) Many allege (typically in passing) that, even if these debiasing procedures prove to be effective, they would still be too laborious and time-consuming to be practically feasible (Section 6).[3]

(INDIVIDUALISM) Others argue that the entire project of seeking out effective debiasing procedures is overly individualistic, a counterproductive distraction from what is at root an institutional problem that demands institutional solutions.[4] The idea is that we are wasting our time unless we are talking about directly changing the underlying material conditions or radically restructuring power relations (Section 7).

Here I reply to these criticisms of debiasing, although I give a more thorough response to the concerns about individualism elsewhere (Madva 2016b). I begin by surveying the relevant research (Section 2), because a clear appreciation of some key details of these studies will help to demonstrate that leading concerns about debiasing are less pressing than commonly portrayed, and may in some respects be altogether unfounded. After reviewing the research, I briefly discuss, and express puzzlement over, how theorists tend to construe the practical implications of these findings (Section 3). I then address each of the major concerns about debiasing (Sections 4–7).

Ultimately, I believe, a comprehensive strategy for combating prejudice and

2. See, e.g., Anderson (2012), Bargh (1999), Devine, Forscher, Austin, and Cox (2012), Gawronski and Cesario (2013), Huebner (2016), Mendoza, Gollwitzer, and Amodio (2010), Stewart and Payne (2008), Olson and Fazio (2006), and Wennekers (2013).

3. See the sources in the previous footnote.

4. See, e.g., Alcoff (2010), Anderson (2010, 2012), Banks and Ford (2008), Dixon, Levine, Reicher, and Durrheim (2012), Haslanger (2015), and Huebner (2016).

discrimination should include a central role for training our biases away. I will, however, sometimes retreat in what follows to the more modest claims that we should collectively take these training procedures far more seriously than we currently do, that we should be testing them in field studies, etc. Given that virtually no one—not even the researchers who have developed these procedures—is considering such field studies, lending some plausibility to these more modest claims would still constitute a major step forward for my position.

2. Research Survey

In this essay, I refer to an intervention as *debiasing* if one of its aims is to reduce or eliminate individuals' undesirable social prejudices or stereotypes.[5] Two clarificatory points about this usage of "debiasing" are in order. First, debiasing sometimes refers more broadly to any intervention that aims to reduce the impact of bias on outcomes (Beaulac and Kenyon 2014), including, e.g., anonymous review. Most likely, reviewing materials anonymously does little to *reduce* one's biases; rather, it leaves those biases in place but hopefully reduces their influence on judgment. While I am generally a proponent of anonymous review, it does not count as debiasing in the sense intended here. Second, debiasing often refers to efforts to combat general cognitive heuristics and biases (Morewedge et al. 2015), such as the confirmation bias, whereas my focus is combating undesirable attitudes specifically related to social group membership. What distinguishes *undesirable* social attitudes from desirable or innocuous ones? I am persuaded by Antony (2002) and Beeghly (2015) that these are empirical and context-sensitive questions (see also Section 5; Madva 2016a). An uncontroversial example of an undesirable stereotype would be a student who is very interested in math but has internalized stereotypes that members of her group are not good at math, such that these internalized stereotypes lead her to underperform on math tests. Some (very valuable) interventions aim to circumvent these internalized stereotypes by designing test-taking environments that make the stereotypes less likely to come to mind, or by teaching individuals cognitive techniques that mitigate the stereotypes' undesirable effects. My focus is on interventions that aim to change these stereotypes on a more profound psychological level, such that the need to circumvent or control them is reduced or eliminated altogether. If biases are like a disease, then I am interested here in potential cures and vaccines, rather than in ways to manage the problematic symptoms.

Among the many debiasing interventions that have been studied, I focus on

5. Thanks to an *Ergo* referee for pressing for more conceptual clarity about debiasing.

a specific class of procedures that aim, through active and targeted practice, to directly retrain the unreflective habits of thinking, feeling, and acting that underlie implicit biases. I'll argue that procedures of this type have been systematically overlooked and even misrepresented by scientists, activists, policymakers, and philosophers. I take these procedures to be less vulnerable to prominent concerns about debiasing than are other interventions (see especially Section 5), and I believe they may have a distinctive role to play in *complementing* existing efforts to combat prejudice, discrimination, and inequality (see especially Sections 3, 4, and 7; Madva 2016b).

In Kawakami, Dovidio, Moll, Hermsen, and Russin's (2000) seminal paper, participants repeatedly "negated" stereotypical associations and "affirmed" counterstereotypical associations. They saw images of racially typical black and white male faces paired with potentially stereotypical traits. In response to stereotypical pairings (a black face with the word "athletic"), they pressed a "NO" button. For counterstereotypical pairings (a white face with "athletic"), they pressed "YES." Participants worked through a total of 384 pairings, or "trials," which took under 45 minutes. Unlike other participants who repeatedly *affirmed* stereotypes or underwent no training, this group went from being biased to unbiased on a measure of implicit stereotyping, even if tested after 2, 4, 6, and 24 hours.[6] These participants were even *less* biased the next day, presumably because they were not fatigued from the training.

Further studies outside of Kawakami's lab have partially replicated but qualified these findings.[7] Two warrant specific mention because they will be relevant for addressing concerns about debiasing. First, Gawronski, Deutsch, Mbirkou, Seibt, and Strack (2008) observed that the original studies confounded counterstereotype affirmation with stereotype negation. They thus split participants into two groups, all of whom saw the same overall set of face-word pairings, but instructed some to simply affirm the counterstereotypes and others to negate the stereotypes. They found that affirming counterstereotypes reduced bias, but negating stereotypes *exacerbated* it. Gawronski and colleagues hypothesize that the primary debiasing factor was "enhanced attention" to one rather than another set of stimuli (2008: 375). However, Johnson, Kopp, and Petty (2016) argued that "more meaningful" forms of stereotype negation may be effective. Specifically, they found that responding "*No! That's wrong!*" to stereotypes reduced bias. They also found that this "meaningful negation" most effectively reduced im-

6. Lai et al. (2016) found that five minutes of counterstereotype training initially reduced implicit prejudice, but effects dissipated after several hours. However, Kawakami and colleagues (2000) already found that a mere five minutes of training (in contrast to 45) was too brief.

7. Cf. Calanchini Gonsalkorale, Sherman, and Klauer's (2013) replication of Kawakami et al. (2000) using generic positive and negative words, rather than stereotype-related terms.

plicit bias among individuals who were already strongly motivated to be unbiased. I highlight Gawronski and Johnson's studies because both exemplify how the *mere fact* of exposure to stereotypes (or counterstereotypes) is insufficient to affect implicit bias. The effects also depend on *what participants are doing*, i.e., how they attend, feel, think, and act in response to the stimuli in front of them. Such facts are, I'll argue, incredibly important for understanding how biases are learned, unlearned, and possibly relearned, and thus for addressing empirical, practical, and institutional concerns about debiasing.

In addition to reducing bias on indirect computer measures, these procedures also influence deliberative decision-making, unreflective social behavior, and test performance.

Deliberative decisions. In Kawakami, Dovidio, and van Kamp (2005; 2007), participants practiced gender counterstereotype training, pairing men's faces with words like "sensitive" and women's faces with words like "strong." In the 2007 study, participants next evaluated four job applications (résumés and cover letters) for a leadership position. All applicants were qualified, but two were men and two were women. With no training, only 35% of participants chose a woman. After counterstereotype training, however, 61% chose a woman.

There is, however, a "catch." The debiasing effect occurred only when the task of choosing the best candidate came *after* a prior "filler" task evaluating the candidates' traits. When participants had to choose the best candidate immediately after debiasing, only 37% chose a woman. Evidently, participants recognized that the researchers were trying to debias them, and tried to correct for this influence by deliberately responding in more stereotypical ways, at least at first. After the opportunity to explicitly resist the training, they then responded in counterstereotypical ways. Thus, individuals might express resentment *immediately after* a debiasing intervention, although the effects nevertheless emerge later. The researchers also recommend pursuing less "heavy-handed" strategies to avoid temporary backlash (Kawakami, Dovidio, & van Kamp 2007: 151).

Unreflective social behavior. One less heavy-handed intervention may be approach/avoidance training. In Kawakami, Phills, Steele, and Dovidio (2007), white and Asian participants repeatedly pulled a joystick toward themselves when they saw black faces and pushed it away when they saw white faces. In pulling the joystick in, for example, it is as if participants are bringing the perceived image closer, or approaching it. In some cases, participants were explicitly told that moving the joystick signified approaching the image. In other cases, the images were "masked" and shown so quickly participants didn't notice them, and instead believed that they were just responding to the words "approach" or "avoid." Both explicit and subliminal forms of training significantly reduced implicit bias on the IAT. Moreover, subliminal approach training influenced social

behavior, leading white and Asian participants to sit closer to a black interlocutor and face him head-on, rather than at an indirect angle.[8]

Test performance. These training procedures also help individuals cope with the stereotypes that might negatively affect *themselves*. Building on Kawakami, Steel, Cifa, Phills, and Dovidio (2008), Forbes and Schmader (2010) found that women who were subtly trained to associate the phrase "women are good at" with math-related words exhibited improved performance on tests of working memory and math questions from the GRE. The benefits of this gender-math counterstereotype training were visible even after a 24–30-hour delay.

Taken together, the evidence suggests that counterstereotype and approach training significantly affect individuals' cognitive, affective, and behavioral dispositions along a number of key dimensions. For the many individuals (like myself) who harbor biases despite being sincerely committed to fairness and egalitarianism, these procedures would seem to be a boon, a concrete way to bring our unreflective dispositions more into line with our considered commitments.

3. The General Reception of Counterstereotype and Approach Training

Kawakami's original 2000 study is widely cited as a sort of "existence proof" that implicit biases are at least *capable* of change, but this research is just as widely dismissed as lacking direct import for the broader struggle against prejudice, discrimination, and inequality. I find this puzzling. Why aren't these training procedures on the table as *one* important thing that those of us concerned to combat discrimination should be doing, and making available to everyone on a large scale? Why aren't researchers testing these procedures in the field? Policymakers already "spend billions of dollars annually on interventions aimed at prejudice reduction in schools, workplaces, neighborhoods, and regions beset

8. The effects of non-subliminal training on social behavior were not tested. See also Phills, Kawakami, Tabi, Nadolny, and Inzlicht (2011). See van Dessel, de Houwer, Roets, and Gast (2016) for additional references, useful theoretical discussion about underlying mechanisms, and some notable failures to replicate the effects of purely subliminal approach training on IAT scores. Wennekers, Holland, Wigboldus, and van Knippenberg (2012) reduced prejudice by having participants nod in response to typical Moroccan names and shake their heads in response to typical Dutch names. Notably, Wennekers (2013) found that nodding in response to only 50% of the Moroccan stimuli, instead of 100%, failed to significantly reduce implicit prejudice. This suggests that consistency in responses is important (see Olson & Fazio 2006: 431), which is a reason to be skeptical about how effectively we can replicate these lab-based interventions in daily life (Section 3). We cannot expect to approach or have universally positive interactions with every member of every social group, stigmatized or otherwise.

by intergroup conflict" (Paluck & Green 2009: 340). Much of this money and time is spent on interventions, such as diversity training, the effectiveness of which "remains largely unknown" (Paluck & Green 2009: 359). Yet nobody, to my knowledge, has seriously advocated implementing these training procedures in these contexts, not even on an exploratory, experimental basis.

Instead, Kawakami's supposedly "laborious 480-trial procedure" (Olson & Fazio 2006: 431), which requires "many, many repetitions to learn nonstereotypical responses" (Stewart & Payne 2008: 1343), is often cited as a point of contrast when researchers discover a less intensive, less demanding intervention (see also Mendoza et al. 2010: 512–3, 521). Many continue to assume that implicit biases are, despite the aforementioned evidence for their partial malleability, still a little too rigid, inaccessible, and unwieldy for changing them directly to be a viable strategy. I will, in Sections 4–7, explore their grounds for pessimism about counterstereotype and affirmation training in greater depth. As it stands, researchers are committed to finding interventions that require less time and effort, and which work primarily by leaving the biases in place but mitigating their effects on judgment and behavior.

Nevertheless, many acknowledge that Kawakami's research might have *indirect* practical import. The trend is to assume that these studies are relevant *only insofar* as they can be replicated in applied, "real-world" contexts. Social scientists and philosophers evidently take for granted that individuals will not actually engage in these very training procedures (or close variants of them), and that we must therefore figure out how individuals can mimic these procedures in everyday life, for example, by approaching and having positive interactions with counterstereotypical individuals.[9] Even Phills and colleagues seem to assume that these training procedures are not themselves good candidates for actual interventions:

> The next step for this research, however, would be to test these procedures in a more applied setting. For example, one possible strategy is to have schools implement morning welcome activities in which students from different ethnic/racial groups approach one another. (2011: 208)

This sort of welcoming activity may be beneficial, but it strikes me as odd to construe it as somehow in competition with the computer-based training proce-

9. Wennekers writes that "repeatedly approaching out-group members and noticing that nothing bad happens may make you less likely to avoid them," (2013: 85) and Schneider concludes that, "Obviously, in everyday life people are not likely to get such deliberate training, but it is certainly possible that those who routinely have positive and nonstereotypic experiences from people with stereotyped groups will replace a cultural stereotype with one that is more individual and generally less negative." (2004: 423). Philosophers such as Alcoff (2010: 131–132), Anderson (2010: 152; 2012: 167–170), and Huebner (2016: Section 4.2) draw similar conclusions.

dures. It would probably be wise for students to work through those procedures *before* the relevant social activity, to nudge those interactions toward going off on the right foot. That is, if there are ways to promote or replicate these interventions in everyday life, so much the better. But why assume that the "next step" is simply to pursue the everyday-life strategies to the exclusion of the training procedures? Why not pursue them in conjunction? The assumption that these procedures are relevant only insofar as they can be replicated in everyday life strikes me as roughly analogous with, say, making the discovery that a certain targeted sports drill improves athletic importance, and then concluding from this discovery that the thing to do is scrimmage more. The evidence discussed in Section 2 suggests that a certain sort of intervention *does* reduce prejudice, but it is interpreted as signifying that *another* sort of intervention—which is for some reason taken to be more authentic and mutually exclusive with the first—*might* reduce prejudice.

What I find additionally puzzling about this pattern of inference is that we already have decades of evidence about attempts to debias through everyday social activities. Further indirect, lab-based evidence does not seem especially informative. Attempts to change attitudes through social interaction (the contact hypothesis) have a long history, and evidence for their success is substantial but complicated.[10] For example, Henry and Hardin (2006) found that intergroup contact generally reduced *explicit* reports of prejudice, but that its effects on *implicit* prejudice were mediated by the social status of the participants. Contact reduced the implicit prejudice of black Americans toward white, but not of white toward black, and it reduced the implicit prejudice of Lebanese Muslims toward Lebanese Christians, but not of Christians toward Muslims. In these and other cases, the higher-status group's implicit biases were unaffected. Moreover, the conditions conducive to effective contact (namely, cooperating toward a common goal, on terms of equal social status) are in some contexts extremely difficult to construct and maintain. A rival "conflict" hypothesis seeks to explain how contact often *amplifies* intergroup animosity. Thus the morning welcome activities envisioned by Phills and colleagues might, across a range of contexts, exacerbate rather than undermine bias. Even when contact reduces prejudice, the effect sizes tend to be relatively small.

I by no means wish to discount the importance of actual intergroup interaction, whether for the specific aim of debiasing or for broader concerns related to social justice. I strongly agree with theorists such as Danielle Allen (2004) and Elizabeth Anderson (2010) that ongoing *de facto* segregation between social groups is a major cause of inequality and an impediment to just democratic

10. See, e.g., Dixon et al. (2012), Kelly, Faucher, and Machery (2010), Pettigrew and Tropp (2006), Putnam (2007), and Shook and Fazio (2008).

decision-making, and that, therefore, integration and interaction between members of diverse social groups are deeply important aims. However, given the large body of evidence on the limitations and occasionally counterproductive effects of intergroup contact, it seems to me that we should actively pursue complementary strategies that nudge these interactions in the right direction, and perhaps amplify their debiasing effects. Maybe if individuals volunteered for a little approach training beforehand, intergroup encounters would be more likely to start off on the right foot and unfold in more positive ways. What really puzzles me, in other words, is that these training procedures are interpreted as evidence for the debiasing power of social contact rather than as evidence for *ways of improving* the debiasing power of social contact.

That there are important connections between approach training and social contact is clear. Phills and colleagues (2011: 198) found that approach training led to "psychological closeness" of a distinctive sort, by strengthening white and Asian participants' associations between blacks and self-related words ("I," "me," "self," etc.). This increased self-identification with blacks evidently played a significant role in reducing anti-black bias. In a sense, then, this research is in keeping with the age-old strategy of reducing prejudice by breaking down "us" versus "them" dichotomies. Approach training may be, in effect, *the contact hypothesis in a bottle*. This is not to say that approach training or its effects are equivalent to actual intergroup contact or its effects. Like most "distillations" or "lab-designed imitations" of naturally occurring phenomena, there are important differences between the bottled version and the "real thing," which usually means that the bottled version is worse in many respects—and better in others. There are myriad potential benefits of having cooperative, respectful intergroup interactions that cannot be achieved merely by moving a joystick back and forth. There is, however, at least one considerable advantage to the computer-based training procedures: we can guarantee that 100% of the trials are counter-biasing in the procedures, but not in unscripted interactions in everyday environments.[11]

It seems to me that these very training procedures, or variants of them that emerge in response to further empirical developments, are themselves among the activities we should all be engaged in to work to undermine the biases we harbor that can do harm to others and ourselves. Rather than portray everyday-life strategies as the more authentic alternative to counterstereotype and affirmation training, I suspect these two sorts of intervention are apt to be mutually reinforcing; I will say more about why in Sections 4 and 7 (see also Madva 2016b). In addition to the everyday-life strategies, I believe we should eventually make versions of these training procedures widely available, and begin to consider how institutions might incorporate them into broader antidiscrimina-

11. On the importance of consistency, see Footnote 8.

tion strategies. Of course, before making serious investments of hope and resources, further research could test the effects of these procedures in the field. I highlight outstanding empirical questions in what follows. However, researchers are not seriously considering such research. So far, investigations of how prejudice-reduction techniques affect behavior outside the lab, such as Devine and colleagues (2012), have focused *only* on how to mimic these procedures in everyday-life interactions, rather than using these procedures themselves. Why?

4. 1st Empirical Concern: The Relearning Worry

The most commonly cited concern, about these and almost every other individual-level debiasing strategy, is how long the effects last.[12] While evidence suggests that these procedures reduce bias for *at least* a day (Kawakami et al. 2000; Forbes & Schmader 2010) or a week (Hu et al. 2015), nobody has, to my knowledge, tested just how long people stay debiased after counterstereotype or approach training. The durability of these procedures is fundamentally an open empirical question. The failure to perform these studies is partly explained by the fact that longitudinal interventions are expensive and unwieldy.[13] However, based on personal correspondence with several researchers, I worry that pessimism about the durability of debiasing may be another contributing factor. Moreover, some write as if the absence of evidence is evidence of absence. For example, Mendoza and colleagues (2010: 520) insinuate that certain studies that detected effects lasting days also *failed* to detect effects lasting any longer, whereas tests on the long-term durability of these training procedures have simply not been done.

The basic conjecture underlying the relearning worry is that as soon as people step outside of the lab, they will be bombarded with stereotypes all over again, and reacquire (or learn anew) all of their biases. For example, Mendoza and colleagues write that attempts "to change underlying representations of racial groups . . . may be more difficult to maintain upon reexposure to societal stereotypes outside the laboratory" (2010: 520; cf. Huebner 2016: 71). Call this the *bombardment basis* for the relearning worry. This conjecture seems to be premised upon a certain commonsensical view of prejudices and stereotypes, according to which we initially acquire these undesirable attitudes through repeated exposure to negative representations of social groups. This is intuitively a gradual process, whereby our biases slowly get stronger, reinforced by ever more prejudice-promoting experiences. Intuitively, the outcome of this gradual

12. See, e.g., Devine et al. (2012: 1267–1268), Lai, Hoffman, and Nosek (2013: 320–321), Mendoza et al. (2010: 520–521) and Wennekers (2013: 130–131).

13. In conversation, Brandon Stewart suggested that another contributing factor is an academic stigma against excessively applied and insufficiently theoretical research.

process is that prejudices will become deeply ingrained in our minds and subsequently be difficult to change. So, the thought goes, won't this process just repeat itself after debiasing?

Since the relevant studies have not been done (for example, Mendoza and colleagues do not cite evidence to support their conjecture about the effects of reexposure to stereotypes after debiasing), pessimists must look elsewhere for indirect support. One source of pessimism might be evidence from developmental psychology that implicit biases tend to form early in childhood and persist through adulthood (Olson & Dunham 2010). While explicit biases improve as children get older—adults are less likely to report racial preferences than 10-year-olds, and 10-year-olds are less likely to report such preferences than 6-year-olds—implicit biases remain surprisingly stable.[14] This might suggest that debiasing effects are likely temporary: whatever causal forces are keeping implicit biases stable over time (presumably some combination of psychological and environmental factors) will still be there after debiasing, and will lead individuals to revert back to their prior biased state.

This research, however, consists of longitudinal observation without experimental intervention. It suggests that, in the ordinary course of things, implicit biases typically don't change in lasting ways; it is silent about whether they can. The developmental research is, moreover, ultimately inconsistent with the commonsense view of prejudice. Infants seem to pick up these biases very quickly *without* years of being bombarded with stereotypes. Kawakami and others' research, in turn, undermines the commonsense view about the resilience of bias in adulthood, suggesting that individuals *can* reduce these biases, at least temporarily. The question is whether the changes will last. So on these points the commonsense view of prejudice, which underlies the relearning worry, is completely off-base. Why, then, should we be so worried about the additional commonsensical pronouncement that getting bombarded with stereotypes outside the lab will undo the effects of debiasing? It is common for social scientists, philosophers, and activists nowadays to speak about how much we've learned about prejudice in recent decades, but I wonder if pessimism about the durability of debiasing is itself a holdover of the *old-fashioned* views that all this research is supposed to have debunked.

In addition to developmental studies, another source of pessimism is evidence that adult exposure to certain forms of "mass media" increases bias. For example, implicit racial biases increase after listening to violent rap music (but not pop; Rudman & Lee 2002), and after watching television clips in which white

14. The stability I mean here is not within-individual test-retest reliability, but overall demographic (between-individual) trends. A wide variety of transient, context-specific variables affect how particular individuals score on measures of implicit bias on particular occasions, such as the "mass-media" effects I describe shortly. See also Section 5 and Madva (2016a).

characters display subtle, nonverbal bias toward black characters (Weisbuch, Pauker, & Ambady 2009). Suppose that, in keeping with the bombardment basis, individuals will encounter many more of these stereotype-promoting than stereotype-disconfirming phenomena once they leave the lab. The prediction that individuals will inevitably relearn their biases depends on a further assumption: that their biases will, over time, come to reflect whatever bombards them most. But we know that this picture of the human mind—as an empty head that simply gets filled with the preponderance of information it encounters—is utterly false. If it were true, it would mean that the mind was an extremely accurate mirror of nature, in the sense that our inductively grounded beliefs and expectations would be closely calibrated to the actual regularities we encounter. It is old news that we don't work like that.[15] We suffer from a profound confirmation bias, being more likely to seek out and attend to evidence that reinforces what we already believe than to consider contravening evidence. And our beliefs often persevere in the face of the contravening evidence that we do happen to consider. It is just false that our biases depend primarily on the mere preponderance of "evidence" we take in, in the form of magazine covers, news stories, or what have you. Typically, belief perseverance, the confirmation bias, and a host of other cognitive dispositions help to create and sustain our biases, but there is reason to think that these dispositions can also be recruited to serve more egalitarian ends.

Rather than being empty heads with no filters on incoming information, what we notice and how we interpret it is profoundly shaped by our implicit and explicit goals.[16] Aims that typically work in *favor* of stereotyping include the desire to protect one's self-esteem (e.g., by putting down another group) and to see the world as a fundamentally just place where people deserve their lot. Aims that often work *against* stereotyping include a desire to be egalitarian, to treat a person as an individual, and to take an outsider's perspective on things. Which goals we have makes all the difference to what we notice and how we interpret whatever bombards us. If we respond to a stereotypical representation by thinking, "There's a grain of truth in that," then we might just be trying to feel better about ourselves—and reinforcing our biases. If, instead, we respond by shouting, "*No! That's Wrong!"* then *that very same exposure* to a stereotypical representation could weaken our biases and reinforce our egalitarianism.

Once we become sufficiently debiased, then, and insofar as we are motivated to stay that way, many of these psychological dispositions might now operate to maintain our *de*biases. Even if we encounter disproportionately more stereotypical than counterstereotypical representations, we might pay disproportionately

15. See Madva (2016b: Section 3) for further discussion.

16. See Kunda and Spencer (2003), Moskowitz (2010), and Uhlmann, Brescoll, and Machery (2010).

less attention to the stereotypes, and perhaps "meaningfully negate" or otherwise discount them when we notice them. Of course, this is speculative. My aim is not to convince you through a priori speculation that debiased individuals will never relearn their biases, but to emphasize that, in the absence of any direct evidence to the contrary, the burden is on the pessimist to explain why the relearning worry is daunting enough to support the widespread perception that these training procedures lack direct, practical import. None of this is to say that we won't also have to work at being egalitarian, or that retraining our biases through these simplistic procedures will instantly endow us with all the right cognitive dispositions—but why shouldn't these procedures be part of this overall process? One simple thing we can do to stay debiased is form concrete plans for how to react to stereotype bombardment. For example, "When I see a stereotypical representation, I will go to my window and shout, *I'm mad as hell and I'm not going to take it anymore!*" and, "When I see a counterstereotypical exemplar, I will cheer, *Shine on, you crazy diamond!*"

Moreover, although there is no *direct* evidence demonstrating the durability of debiasing, *indirect* evidence to support optimism is growing. Devine and colleagues (2012) taught participants five strategies they could employ in daily life to reduce their racial biases. This intervention led to partial reductions of bias that lasted at least eight weeks.[17] Evidence also suggests that counterstereotypical teachers can reduce their students' biases. Dasgupta and Asgari (2004) found that first-year undergraduate women who took multiple classes with woman math and science professors showed less implicit gender bias after one year (see also Rudman Ashmore, and Gary 2001; Stout, Dasgupta, Hunsinger, & McManus 2011). Presumably, the participants in these studies were simultaneously being bombarded with gender stereotypes every time they turned on the television, or read a *New York Times* obituary of a woman rocket scientist that foregrounded her reputation as the world's best Mom and an expert at making beef stroganoff.[18] Yet their salient classroom experiences evidently "won out" over the media bombardment. So while I certainly agree that exposure to stereotypes in mass media tends to have pernicious effects, especially in the ordinary course of events that leads many of us to acquire biases in the first place, I find it highly unlikely that being bombarded with stereotypes is sufficient for individuals to

17. This intervention, which did not include any targeted approach or counterstereotype training of the kind described in Section 2, tended to reduce but not eliminate bias. In light of the studies on alcoholism recovery that I discuss in the next paragraph, there is reason to think that such training procedures might *enhance* the effectiveness of Devine et al.'s (2012) daily-life strategies. I asked one of the authors of this study why they did not include any approach or counterstereotype training, and the author didn't know. My impression was that it simply hadn't occurred to the researchers as a serious option.

18. See Sullivan (2013, April 1) for discussion and links to Martin's (2013, March 30) obituary of Yvonne Brill.

either learn or relearn their biases. A further, crucial set of variables revolves around *how individuals react* to what bombards them: their short- and long-term motivations and goals, their conscious thoughts, their background store of experiences, and their unreflective patterns of attention, feeling, and action.

Perhaps the strongest evidence—albeit still indirect—for the durability of prejudice-reducing interventions comes from clinical research. Wiers, Eberl, Rinck, Becker, and Lindenmeyer (2011) found that patients recovering from alcoholism who, immediately prior to undergoing standard treatment, were trained to avoid images of alcohol (in four sessions lasting only 15 minutes each) were significantly less likely to relapse *one year* after being discharged, in comparison to patients who underwent no training or sham training.[19] Of course, no one is making the absurd claim that moving a joystick back and forth will, all by itself, cure alcoholism.[20] The point is that alcohol-avoidance training *together* with other forms of therapy tended to have much more durable effects than therapy alone. By the same token, I am not claiming that approach training will, all by itself, solve racism and end inequality. I am claiming that it is one thing we should be doing, together with everything else that we should be doing. The most commonly cited reason to write off these training procedures is pessimism about durability, but such pessimism is, at this time, unwarranted. Wiers and colleagues' research suggests that, if anything, these procedures *enhance* the durability of standard interventions (thus serving as another reminder that the question raised in Section 3, whether to engage in these very training procedures versus try to replicate them in daily life, represents a false choice). While I doubt that social prejudices are more difficult to dislodge than addictive impulses, we obviously cannot assume that approach training's long-term effects on prejudice will be comparable to its demonstrated effects on alcoholism recovery. The long-term effects on prejudice are pressing empirical questions, which continue to go untested.

But suppose that the effects of debiasing are not permanent. How long would they have to last in order to be worthwhile? Suppose debiasing worked like dental cleanings, and it was best to debias ourselves once or twice a year. Would a biannual trip to the debiaser be too much to ask? Would it be a counterproductive waste of time to debias ourselves once in a while even if we didn't re-up quite as often as recommended? What if we can debias ourselves *subliminally* while engaged in other tasks?

In response to my question about visiting the debiaser twice a year, a referee for *Ergo* points out that I offer "no discussion of the points that would make this question answerable: Who would run these de-biasing clinics? How would

19. These impressive findings were replicated by Eberl et al. (2013).
20. See, e.g., Lindgren et al. (2015).

they be funded? How would people be made to go to them?" These are very important structural-institutional questions, to which I will return in Sections 6 and 7, but I should clarify that I do not actually advocate coercing people into debiasing themselves if they don't want to. I am concerned here with the many people, including myself, who harbor biases that we reflectively disavow—those of us who already want to be less biased than we are. As the relearning worry is the most commonly cited reason for skepticism about these training procedures (and about all individual-level debiasing strategies), I believe it deserves to be considered somewhat independently from complex questions about funding and implementation. My aim in the foregoing paragraph was to raise questions regarding just how long-lasting the effects have to be in order to be worthwhile, by appealing to an analogy with another sphere of life (dental visits) where we take for granted that a certain amount of upkeep is preferable. Many of us (including those who have accessible, high-quality dental care) do not visit the dentist as often as we know we should. Other activities that we may not get around to often enough include cleaning the interior of the fridge, cleaning out closets, replacing smoke-detector batteries, etc. But it would be bizarre to infer from the fact that we don't do them frequently enough to the conclusion that they are not worth doing at all, or to infer from the fact that the best effects are achieved by doing them a few times a year to the conclusion that there is no ethical or prudential reason to do them *ever*. I do not purport to have an answer for exactly how long-lasting the effects of debiasing procedures would need to be in order to be worth doing, let alone to be considered morally obligatory, although it seems clear to me that permanence is an unreasonably high bar. All that said, I take the referee's point that we cannot give fully satisfactory answers to these questions in the absence of a fuller picture of what implementing these interventions would entail. I will hazard some conjectures about institutional implementation in Section 7, but the need for a fuller picture strikes me as all the more reason to initiate field studies to determine precisely what would be required in order to make these interventions as effective and durable as possible.

5. 2nd Empirical Concern: The "Context-Specificity" Worry

Another pervasive concern, which is more serious than the relearning worry insofar as it has substantial, if still indirect, empirical support, is that the effects of debiasing might be highly *context-specific*. Might the effects only be visible in this particular lab, or on that particular test? Rather than unlearning their implicit biases, participants might just be learning to *subtype*—picking up on distinctive features of a specific type of individual (or context) within the larger group, such that their default impression of the group remains unchanged. Bouton (2002:

976) helpfully compares this phenomenon to learning that a familiar word has multiple meanings. For example, when we learn that the exclamation "Fire!" has a different meaning in a movie theater from in a shooting gallery, we do not unlearn the first-learned meaning; we learn that the meaning of "Fire!" depends on its context. Subsequently, our default reaction to hearing someone shout "Fire!" will, in novel contexts, likely reflect the first-learned meaning rather than the second. Researchers can test whether debiasing has similarly context-dependent effects by exposing participants to novel exemplars of a social group in novel contexts, and seeing whether their automatic responses reflect their first impressions of the group or their more recently learned counter-impressions.

Rydell and colleagues have done just this, in a series of studies using a different implicit learning paradigm from Kawakami's, and seem to have confirmed all of our worst fears.[21] Generally speaking, it looks like first impressions are incredibly important: people's initial salient exposure to a category member forms the backdrop for their future encounters with other category members. People can pick up quickly on the fact that novel category members do not fit the original mold, but rather than revising their overall impression of the category, they glom onto specific, individuating features of the novel exemplar or its context. In Rydell and colleagues' experiments, participants might read information about a person named Bob, seeing his photo against a blue computer screen. Suppose the information depicts Bob in a positive light and they form a positive impression of him. If they subsequently learn negative facts about Bob against a *yellow* computer screen, then they will eventually learn to automatically respond negatively to Bob—but only when they encounter him against a yellow background. If they later see him against a blue background, or some novel color, their automatic response will reflect their initial positive impression. Maybe what we have been interpreting as attitude malleability just reflects a kind of "fine-tuning" where people's default attitudes toward groups remain stable but they learn about particular subtypes who don't fit the mold.

The context-specificity worry has substantial empirical support, and is consistent with decades of research on animal learning. As far as I can tell, however, the context-specificity of training in Kawakami's paradigm has not been tested. And there is straightforward evidence internal to Kawakami and others' studies to support the hypothesis that these sorts of debiasing will be less susceptible to those sorts of context effects. Their potential for context-generality is, in fact, a primary reason that I have honed in on these particular procedures out of the hundreds of studies that purport to reduce bias.

First, a number of these studies demonstrate how training in one "mode" or context can have effects on tests in a very different "mode." Retraining implicit

21. E.g., Rydell and Gawronski (2009). See Gawronski and Cesario (2013) for a review.

stereotypes led to changes in implicit *prejudice*, even though the stimuli during training and testing were completely different (Gawronski et al. 2008). Subliminal approach training influenced participants in the context of taking an IAT but also in the context of interacting with another human being, with a face they had never seen before (Kawakami, Phills, Steele, & Dovidio 2007). Different *sorts* of approach training, which share nothing in common except their conceptual "approach-iness" led to reductions in bias, across an array of different measures (e.g., Phills et al. 2011). Math-gender counterstereotype training improved women's performance on tests of working memory and math at least a day later (Forbes & Schmader 2010). Avoiding images of alcohol influenced implicit attitudes but also reduced the likelihood of relapse into alcoholism for at least one year (Wiers et al. 2011; Eberl et al. 2013). There seems to be substantial evidence that these procedures generalize to precisely those novel contexts we're most interested in, including open-ended social interactions, tests, and actual decisions about whether to consume alcohol.

Second, it bears emphasizing that significant effects do not appear in Kawakami's debiasing paradigm until after participants have already worked through *80 trials*, and it takes a few hundred more trials before participants approach a ceiling past which they cannot improve. It takes a reasonable amount of practice and effort over a significant number of trials. This suggests that the psychological forces at play are not quite so fast-learning (and perhaps context-specific or surface-level) as those involved in other interventions that have been found to reduce bias, such as Olson and Fazio's (2006) finding that just 24 subliminal exposures to counterstereotypical pairings could reduce bias on one measure, or Blair, Ma, and Lenton's (2001) finding that 5 minutes of imagining a counterstereotypical woman could reduce gender bias on several measures. There is good reason to think that something *more*, or at least something *different*, is going on in Kawakami's paradigm.

Third, in addition to the total number of trials necessary to reach significant effects, it is noteworthy that these forms of training involve robust (if rote) *actions* on the part of the participants. They are not just passively taking in information (as if watching TV), but engaging in embodied counter-biasing behaviors. This contrasts with, say, Dasgupta and Greenwald's (2001) paradigm of reminding participants of admired black individuals and infamous white individuals.[22] Recalling counterstereotypical exemplars may be a valuable exercise, but it is, plausibly, just making certain positive subtypes of the categories more accessible, without actually changing participants' attitudes about these categories. For that, more direct actions that actually challenge those attitudes might be necessary, and they might have to be repeated a few hundred times.

22. A large-scale attempt to replicate this study did not find similarly strong effects (Joy-Gaba & Nosek 2012).

My final response to the context-specificity worry is more nuanced, and I develop it in greater length elsewhere (Madva 2016a). Our aim is not the total erasure of "stereotypical associations" from our minds. There are many contexts where stereotypes ought to spring immediately to mind: in particular, when people are being treated in stereotypical ways and we must swiftly respond "*No! That's wrong!*" We need to know about stereotypes in order to challenge them. I take this to mean that a certain sort of context-specificity is a good thing. We want to not use or think about stereotypes when they are irrelevant, and we want to think about them when they are relevant. Thus, evidence for the context-specificity of these sorts of interventions is not, just as such, a bad thing. It remains to be seen, of course, whether the sort of context-specificity that implicit biases actually exhibit maps onto the sort of context-specificity that would be cognitively ideal. But research on the goal-dependence of stereotyping (Section 4) suggests that if we adopt the right sorts of goals, we can make significant progress toward regulating our knowledge of stereotypes so that they are activated in the right contexts, and inhibited in the wrong ones.

6. Practical Unfeasibility

Suppose we grant, for the sake of argument, that these training procedures lead to reasonably durable, context-general reductions in bias. Critics of debiasing further justify their skepticism, in part, by referring to the fact that these supposedly "laborious" procedures require "many, many repetitions" to be effective, thereby implying that they are somehow unfeasible. As if the sheer fact that they involve *hundreds of trials* is sufficient to establish that they are too onerous and labor-intensive to figure as a legitimate component of the larger struggle against prejudice and discrimination.

How labor-intensive are they? Reliably significant effects start appearing after about 160 trials, and many studies include just 200.[23] The benefits of additional training are still visible from 200 to 300 trials, but, following a classic "learning curve," start to tail off around 400 (Kawakami et al. 2000: Study 3). At most, participants work through 480 trials. One upshot might be that even if we were only to work through 200 trials, we could become significantly less biased than we are, although we would not reach our maximally debiased potential. I take it that becoming *less* biased would presumably still be desirable, even if, for whatever reason, becoming as unbiased as psychologically possible were unfeasible. In any case, working through these hundreds of trials can be done on any

23. Gawronski et al. (2008), Johnson, Kopp, and Petty (2016), and Wennekers et al. (2012; 2013).

personal computer, and done subliminally, perhaps merely by "liking" things on social media or playing Angry Birds. Working through all 480 trials takes about 45 minutes. Given the stakes, 45 minutes is *nothing*.[24]

I cannot seriously entertain the possibility that three-quarters of an hour of counterconditioning is too much to ask of ourselves. Maybe if we had to *constantly* countercondition ourselves, this would become burdensome, but, in light of my responses to the relearning worry, I doubt this is an insurmountable threat. It is simply false that these training procedures are prohibitively laborious or time-consuming. The widespread conviction that implicit biases are too deeply ingrained to uproot in any practically feasible way is undermined by these very findings.

This leads me to suspect that the prevalent misperception of debiasing as unfeasible may, ironically, be explained in part by a number of well-known social and cognitive biases, including, for example, the framing effect. Working through 480 trials to countercondition a bias, described in one context or "frame," sounds like a lot ("many, many repetitions"). Yet the 45 minutes it takes to do so is miniscule in comparison to the tremendous resources that individuals, governments, schools, and businesses already devote to diversity initiatives and prejudice reduction, to say nothing of the time and resources devoted to the education of democratic citizens, and to teaching students foreign languages, musical instruments, sports, typing skills, and calculus. Compare it to the investments we make in dieting, therapy, and breaking bad habits and addictions. 45 minutes is less time than many people spend *per day* on exercise and the honing of other skills. American children spend an average of four hours a day watching television, and an average of 135 hours a year learning foreign languages. They can't give up one afternoon to try out a prejudice reduction strategy that has significant empirical support?

However, one might still think that debiasing is unfeasible because there are a *lot* of biases out there, and if it takes 20–45 minutes to significantly reduce each of them, then how many hours will it take to fix them all? This is an important question to explore empirically, but it seems unfair and misguided to suggest

24. See Section 1 for brief discussion of why the stakes here are high for everyone. An *Ergo* referee suggests that, like the question of durability (Section 4), the question whether 45 minutes of training is unfeasible cannot be answered unless the details of institutional implementation are filled in. As before, I grant that having a fuller picture in view will ultimately be necessary for assessing whether these procedures represent effective and morally acceptable tools to combat discrimination and inequality (I will say more about institutional implementation in Section 7). Nevertheless, I think that questions regarding how time-consuming and laborious the training procedures are deserve to be considered in their own right. For example, even if we stipulate for the sake of argument that these procedures can be institutionally sponsored in unobjectionable ways, a further question is whether they are so intensive that it would be unreasonable (or even just unlikely) for people to do them.

that it poses a problem for the practical feasibility of debiasing. First, there seems to be another framing effect afoot, such that all implicit biases are being grouped together as the *same* problem—Implicit Bias—sharing a single underlying cause and requiring a single solution. But we would not, for example, rule out particular proposals for *institutional* interventions on the grounds that they won't be equally effective at countering all possible forms of discrimination. The institutional interventions that best address disadvantages for women in STEM fields may not overlap perfectly with those that best address racial discrimination in the criminal justice system, nor with those that best address the exclusion of individuals with disabilities from public spaces. Second, if debiasing ourselves in all relevant respects proves too laborious or time-consuming, then individuals can simply prioritize those biases that are more directly relevant to their daily lives, occupations, career goals, ethical commitments, or idiosyncratic hang-ups. We don't, as it happens, all share exactly the same biases. The essential debiasing procedures for medical doctors, high-school guidance counselors, and airport security employees might differ greatly (or they might not). Third, if we can do this training subliminally, while wholly absorbed in other unrelated tasks (surfing the internet, social media, video games?), then it might be more or less irrelevant how many hours it would take to countercondition all the relevant biases. Fourth, evidence suggests that some debiasing procedures might generalize in important respects (Section 5). It then becomes a crucial empirical question which specific training procedures most efficiently achieve the broadest range of relevant effects. For example, perhaps we can train ourselves to automatically "avoid prejudice" and "approach egalitarianism" in general. Glaser and Knowles (2008) found that individuals with implicit negative attitudes *toward prejudice per se* showed less racial bias. Perhaps we should practice approaching the voting booth and avoiding the status quo. In fact, I would argue that for every broad structural reform we ought to prioritize in the struggle for social justice, there exist some individual-level reforms that we ought to prioritize *because these individual reforms promote that structural reform* (Madva 2016b: Section 1). To my mind, then, the question is not *whether* to employ individual-level training procedures. The (straightforwardly empirical) question is simply *which* specific attitudes or habits to target, given our broader social-political aims.

At this point, I can only speculate about additional factors that might drive an aversion to debiasing. In discussions with colleagues, students, or acquaintances, it sometimes seems as though people have a *kneejerk* negative response to the very idea, and thereafter confabulate reasons that justify their aversion.[25]

25. Cf. Haidt (2001). My consideration of how social and cognitive biases might contribute to skepticism about debiasing also draws from speculations made about the role of cognitive biases in, e.g., the widespread indifference or failure to act in response to climate change, global poverty and hunger (Gifford 2011) and mass incarceration (Alexander 2012: 198). It is plausible that perva-

To many the whole business seems *creepy*. It sounds like "thought police" and brainwashing. Talking seriously about counterconditioning inevitably calls up images of *A Clockwork Orange*, with Malcolm McDowell strapped to a chair, eyelids peeled back, being injected with giant needles full of nausea-inducing chemicals while he watches an endless stream of graphic violence. I hope it goes without saying that there is a lot to object to in the counterconditioning of *A Clockwork Orange* that I am not advocating here.

Of course, nobody is made uneasy by the prospects of having to actually go through the motions of training or retraining themselves in other contexts—memorizing flashcards, working through problem sets, practicing sports drills and musical scales. We might be instinctively averse to these activities because of their *tedium*, but not because of their creepiness. Many people also readily acknowledge the importance of cultivating good habits to living an ethically desirable life. In this way, the creepiness worry about debiasing might reflect a misunderstanding of the phenomenon in question. Perhaps counterconditioning would be problematic if it involved indoctrinating alien beliefs and values. But the aim of debiasing is to help us better live up to and embody the commitments we already have, not to instill new ones. We are trying to *fight back against* the alien beliefs, values, and habits of thinking, feeling, and acting that we absorb from our systemically racist and sexist environments. That's why genuine, full-blooded retraining has to be part of the discussion. Just like unlearning bad habits and learning new skills or languages, there simply has to be a central role for *practice*.[26]

What if debiasing has other, unforeseen consequences on our beliefs, values, and habits? Another worry associated with *A Clockwork Orange* is that debiasing interventions could have unexpected effects apart from bias reduction. Perhaps an intervention that effectively reduces a person's biases will also make him chronically depressed or low in self-esteem, or maybe it will make him smugly self-satisfied and complacent. Whether there are problematic unintended consequences of debiasing is an empirical question like any other, and should be explored, but I suspect that the benefits of reducing widespread discrimination will outweigh any such unforeseen costs. Of course, the potential for unforeseen costs is a risk for any intervention aimed at prejudice reduction, including the

sive biases toward indifference, ignorance, and avoidance of the people and problems that are geographically, socially, or temporally "distant" from us contribute to the tendency to cast debiasing as unfeasible and ineffective. Of course, these are also precisely the sorts of biases that approach training might help us overcome.

26. In conversation, Michael Brownstein and Manuel Vargas suggested that there might be some additional factors that explain (without really justifying) our kneejerk reluctance to debiasing, such as the alienating perception that the training requires using myself (or my mind or body) as a mere means to an end. Or our specific reluctance to debiasing might be due to how loaded racism, sexism, and prejudice are with ethical, political, and emotional baggage (in contrast to practicing problem sets and musical scales). Both seem highly plausible.

interventions that do not strike people as creepy, and, indeed, is a ubiquitous risk for every kind of intervention in every kind of system—whether the system is psychological, biological, technological, social, ecological, etc. There does not seem to be a special problem of unanticipated side-effects for Kawakami's procedures. Moreover, if, say, reducing white men's biases will also lower their self-esteem, then the solution, I think, is to find alternative sources of self-esteem. Alternatively, if these interventions risk making us complacent, they should be coupled with strategies to resist complacency.[27]

In fact, objections about the creepiness of debiasing seem to seriously underappreciate the extent to which politicians and businesses are already trying to brainwash us using these very tools. In "How to Like Yourself Better, or Chocolate Less" (2009), Irena Ebert and colleagues found that even well-established implicit preferences for *Haribo* gummy bears versus *Milka* chocolate could be reversed—through a training procedure that, using different stimuli, was also found to enhance implicit self-esteem (see also Gibson 2008). Perhaps research on approach training partly inspired an MSNBC commercial campaign in 2010, which featured ads that paired the progressive-sounding slogan "Lean Forward" with photos of its leading personalities, such as Rachel Maddow. In other words, the cat is already out of the bag. To object to debiasing on the grounds that it has a weird whiff of brainwashing is to fail to appreciate the extent to which massive resources are devoted to brainwashing us through precisely these means all the time. Why would we want big business to have a monopoly on brainwashing![28]

In this vein, the creepiness worry seems especially dissonant with the bombardment basis for the relearning worry. There seems to be a straightforward tension in arguing both that debiasing is pointless because we will just relearn the biases upon leaving the lab and that debiasing is creepy because it is like brainwashing. The anticipated relearning is presumably supposed to occur as a result of similarly brainwashing-esque procedures, such that our external environments imbue us with prejudiced beliefs, values, and habits that we would rather not have. It is puzzling that we would let ourselves become inured to the reality of powerful external forces brainwashing us all the time, but feel queasy about the opportunity to resist these forces and take matters into our own hands by debiasing ourselves.

I have also found, especially in discussions with students, that these procedures may seem problematic because of the ways they use photos of real people. Specifically, these procedures strike some as *using* people as mere means to help make ourselves less biased rather than treating these individuals as ends in themselves. However, the same could be said of most of the other interventions

27. I say more about the unintended consequences of debiasing in Madva (2016b: Section 5).
28. Thanks to Katie Gasdaglis for discussion on this point.

on offer, e.g., reflecting on infamous white individuals to drive down a preference for whites. Perhaps using fictional individuals or computer-generated faces could circumvent this worry. (Wennekers, 2013, found that approaching faces and avoiding images of *closets* reduced prejudice.) But perhaps such procedures would still seem objectionable because they "use" racial whiteness and blackness to reduce our prejudices. If we take this concern seriously, however, then efforts to replicate these interventions in everyday life are far more troubling than lab-based versions. Bringing whites into contact with blacks *for the sake of* reducing prejudice seems a much clearer case of using people as means. Frankly, I doubt that any prejudice-reduction technique with a fighting chance at efficacy will completely avoid raising this concern.[29] Perhaps the best I can say here is that anyone who finds these procedures objectionable on these grounds should not be required to do them (this is easy for me to say, as I believe that no one should be required to do them).

I suspect that one of the most significant biases driving kneejerk pessimism about debiasing is the extent to which these studies *implicate us as individuals*. If individuals can really take their implicit biases into their own hands, that means *I* can do so, and if I can, then, other things being equal, I probably should. But if I can tell myself a plausible story about how it is a massive social-structural problem that cannot be solved at the individual level, then I do not have to feel bad for failing to take steps to improve myself. The primary oversight in this sort of self-deflecting response is the failure to appreciate that, even if changing ourselves as individuals won't directly change the whole world, these biases are nevertheless leading us to treat the *other individuals* we encounter (and ourselves) in morally problematic ways. It is imperative that each of us ask ourselves, as Barack Obama implored in response to Trayvon Martin's shooting, "Am I wringing as much bias out of myself as I can? Am I judging people as much as I can, based on not the color of their skin, but the content of their character?" Implicit bias is as much a genuinely *ethical* problem as it is a *political* one; we as individuals are regularly failing to treat the other individuals with whom we interact as we ought. The problem is not just "out there" in the sociopolitical ether, but embodied and enacted in the myriad subtle and not-so-subtle ways we treat each other. Calling it political can be a way of forgetting that it's ethical, too.

Another source of kneejerk pessimism might have to do with how *stupid* or *brainless* these interventions seem.[30] There is a certain fantasy that the hard

29. See, e.g., Anderson (2010: Chapter 8) on the bleak prospects of colorblind strategies to reduce racial prejudice and inequality.

30. "Indeed," write Forbes and Schmader about their counterstereotype training, "it is almost shocking to think that having someone pair a basic activity, such as walking, with math would be sufficient to both alter the nature of a stereotype and free up subsequent working memory resources when performing in the domain" (2010: 13).

problems in our lives must be overcome by some deep, cathartic experience, or via some profound insight into human nature. In personal correspondence, Miranda Fricker made the similar suggestion that these studies might be perceived as a threat to our moral depth and stability. We like to think that our virtues as well as our vices "run deep." I wonder whether this sort of desire for depth isn't responsible, in part, for the continued resistance to accepting that less sophisticated habits of thinking, feeling, and acting make significant causal contributions to many of our personal and social ills, including prejudice and discrimination, and that these habits will have to be changed in order to remedy those ills.

In the context of fighting sexism and racism, the desire-for-deep-answers might manifest in the conviction that we must understand Marx's critique of capitalism, Foucault's analysis of power, and MacKinnon's account of discrimination before we can get serious about combating large-scale social ills. I agree that we must understand these analyses. We must take a hard look at the underlying structures of power and oppression, and work to change them, but there is no inconsistency in combating prejudice on personal and political fronts *concurrently*. The desire-for-deep-answers may partly inspire the critique of debiasing, which I discuss in the next section, as too simplistic and individualistic. How could a simple thing like changing an individual's prejudices combat this incredibly complex power structure? (The framing effect may be at work here as well.)

Perhaps another concern, roughly to do with intersectionality, is that just approaching blacks and avoiding whites with a joystick is problematically oversimplified in contrast to the inherent complexity of social identity. I think this point is basically right. In a similar vein, the pervasive, racially biased habits explored by theorists such as Allen (2004), Alcoff (2006), and Sullivan (2006) and are far richer and more complex—psychologically, socially, historically, and symbolically—than those involved in Kawakami's debiasing procedures. My response is to invoke an analogy with linguistic fluency (Madva 2012). Memorizing vocabulary and grammar rules is not the same as becoming fluent in a second language. But we need to memorize vocabulary and learn a bunch of rules before becoming truly fluent. Kawakami's simplistic procedures may be the anti-prejudicial equivalent of memorizing flashcards. These are the *basics*, which put us in a better position to actually *act* in unbiased ways in the everyday world, with all its inherent complexity.

7. Individualism

Although many philosophers, social scientists, and activists agree that the pervasion of biased "microbehaviors" contributes to macro-level injustices, many are skeptical of interventions that seek to change these microbehaviors by coun-

terconditioning individuals' implicit biases. We should, they argue, instead focus on the substantive, structural factors that perpetuate discrimination and inequality. I agree that social structures such as *de facto* segregation (Section 3) and pervasively biased mass media (Section 4) are significant sources of injustice, which, at least in the ordinary course of things, tend to reinforce our biases. I wholeheartedly agree, therefore, that profound structural interventions are necessary, and I take up concerns about the putatively individualistic focus of debiasing efforts in greater depth in a companion essay (2016b). Far from being in competition, however, training procedures like Kawakami's may be integral to the successful implementation of broader structural reforms.

Institutional efforts to combat discrimination and promote diversity, such as race-based affirmative action and the integrationist rezoning of school and voting districts, continue to be controversial. Support for these policies tends to be deeply divided along racial lines (Drake 2014). In American courts, the overarching pattern in recent years has been to roll back existing structural interventions because they purportedly amount to "reverse" discrimination (e.g., *Parents Involved v. Seattle* 2007; *Schuette v. Coalition* 2013; *Wal-Mart v. Dukes* 2011). I frankly fail to see how, in the contemporary political climate, structural interventions for addressing bias, discrimination, and inequality have cornered the market on brass-tacks pragmatism. Apart from asking how effective debiasing will be, we might also ask, *how much opposition will there be*? We can make counterstereotype and approach training widely available to individuals without overhauling social structures in potentially contentious ways. While we can weave these forms of debiasing into our institutions, we need not. These procedures will not live or die on the whims of lawmakers and judges. If we are speaking practically about the current state of US politics, then the individualist strand in debiasing might be a virtue rather than a vice. These training procedures can be defended in terms of the values and political ideals of those who object to institutional interventions as paternalistic or reverse-discriminatory: by making these procedures widely available, we can give individuals the free choice to take responsibility for debiasing themselves.

As opposition to affirmative action has grown (or at least held steady), and as the courts have struck down some historically influential defenses of the practice (e.g., by discounting the justification of affirmative action as a compensation for past injustice), theorists and activists have sought out alternative ethical and legal grounds for it. Debiasing figures prominently among "new" justifications for affirmative action, as many claim that promoting members of underrepresented groups to positions of prominence will produce "debiasing agents," counterstereotypical exemplars who debias their peers.[31] Matters are likely not so simple.

31. See Alcoff (2010), Anderson (2010), Jolls and Sunstein (2006), and Kang and Banaji (2006).

If coworkers *believe* that others have been promoted ahead of them simply to satisfy a quota, they may resent what they (perhaps wrongly) perceive to be undue benefits, under-evaluate their performances in the future, and so on. For example, Kaiser and colleagues (2013) found that the mere presence of diversity-promoting structures can ironically lead some privileged individuals to become more discriminatory. Given such findings, we cannot assume that institutional interventions will have debiasing effects. Implementing them without sufficient attention to the motivations, interpretations, and biases of the individuals involved can backfire, begetting heightened prejudice and discrimination. Fortunately, we do not have to look far for psychological interventions that could mutually reinforce institutional change. Kawakami's training procedures could provide the necessary psychological scaffolding to implement antidiscrimination initiatives without amplifying hostility; at the same time, integrationist initiatives like affirmative action might provide the necessary environmental scaffolding to reinforce the effects of training procedures (e.g., people will encounter counterstereotypes both during training and in the workplace, and have opportunities to have their debiased expectations confirmed). The fundamental answer to the individualist criticism is simple: implement debiasing on an institutional scale.

However, the prospect of institutional sponsorship of debiasing raises worries of its own—again calling up images of "thought police" and mandatory brainwashing—but these worries are, again, misguided and unfair. They are misguided because they fail to appreciate the extent to which debiasing is a *response* to objectionable forms of brainwashing that are already operative, and because they wrongly construe the aim of debiasing to be the manipulation of our beliefs, or the implantation in our minds of external goals and values (Section 6). Instead, the aim of debiasing is ultimately to bring our unreflective habits of thinking, feeling, and acting into accord with the beliefs and values we already endorse, or at least claim to. These worries are also unfair, because institutional sponsorship need not take the controversial form of, e.g., a universal debiasing mandate. There are numerous "nudges" that institutions can employ to encourage debiasing without making it obligatory, such as by auto-enrolling employees in a debiasing program and allowing them to opt out.

Before concluding, let me return to the series of questions about institutional implementation raised by an *Ergo* referee at the end of Section 4. The referee argues that questions about durability, feasibility, and the scope of our normative obligations to reduce our biases are unanswerable without a better sense of how people would be compelled to debias themselves, how debiasing would be funded, who would be in charge, etc. First, I do not think people should be compelled to engage in these procedures, nor do I think individuals or institutions should be permitted to subliminally debias employees or consumers without

their knowing. Perhaps laws will need to be passed to outlaw such practices. Given that politicians and corporations may already be allocating significant resources toward brainwashing us through comparable means, the need for such regulations, or for the enforcement of existing regulations, may be overdue.[32] Second, regarding funding, I believe that the same individuals and institutions (schools, charities, businesses, governments) who now spend billions of dollars on diversity training and other prejudice-reduction techniques (i.e., on interventions that already aim to influence us but have not generally been shown to do so effectively) should allocate some of those resources to studying these training procedures. I don't find anything in-principle problematic about members of these institutions funding and spearheading such studies. What I find problematic is that many existing interventions are pursued with no serious concern for or assessment of their effectiveness, often just to "check a box" to mark completion and thereby prevent lawsuits. Third, regarding who runs the debiasing, I believe that much of the precise nature of these interventions should be driven by the results of empirical research, but the decision-making should also be, as much as possible, transparent to and within the power of participants themselves. Individuals should be fully aware what they are signing up for, and even free to choose exactly which of their own attitudes, if any, they'd like to change.

Here are some concrete examples, which are highly speculative in terms of details but exemplify the kinds of implementation I believe would be appropriate. First, some internet browsers allow users to install "add-ons," e.g., programs that block advertisements. Perhaps we could develop debiasing add-ons which individuals choose to install, and then freely toggle on and off, such that whenever they click on a link or "like" something online, they are (perhaps subliminally, perhaps consciously) presented with certain images (members of other social groups, healthy foods, etc.). Individuals could presumably even choose or upload the precise sets of images involved. (Again, it would be wrong and should be illegal to flash such subliminal images without individuals' knowing, voluntary participation.) Second, it might be possible to develop and make freely available otherwise fun, ordinary video games that include debiasing components. Consider, for example, a dancing game, in which players practice approach-oriented dance moves in response to images of individuals from other social groups; or a word-unscrambling game (like Boggle) in which players see images of outgroup members and then have to unscramble letter-strings like "flyndire" into "friendly." Third, many institutions require their members to take online courses about, for example, sexual harassment. Perhaps there could

32. The UK and Australia ban subliminal advertising (BBC News 2009, September 28). Another major obstacle to subliminal advertising is fear of public-relations backlash, as Facebook experienced after publishing a study showing that varying the content of users' news feeds could influence their emotions (Albergotti 2014).

be a (completely voluntary!) option to subliminally debias oneself during such online training, or, after the training is over, to play a debiasing game. Fourth, many occupations require individuals to engage in continuing education, but individuals have a range of options about precisely which courses to take. Perhaps one of the options could be a mini-course that first explains the nature and causes of contemporary prejudice, then explains the research behind these debiasing procedures, and then gives participants the option to engage in them—again, allowing participants to choose which sets of stimuli they'd like to work with, perhaps after they have tested themselves to uncover their own idiosyncratic biases. Of course, whether these interventions would be feasible or effective are open-ended empirical questions. I see no good reason why we are not exploring them, and nothing to suggest they'd be inherently unjust.

When I give lectures or training sessions on implicit bias, some individuals remain resistant to the idea that they might be biased, but many others accept it, and often their first and most persistent question is what to do in response. They're motivated to be less biased but lack concrete strategies. I tell them about some of the most promising institutional interventions and daily-life debiasing tricks, but it's readily apparent to audiences that these interventions will not get us all the way there. One conspicuously missing piece of the puzzle, I submit, is the availability of humdrum ways to *practice*, which we recognize as necessary in a variety of other spheres of life, including learning musical instruments, languages, quantitative skills, and even ethical dispositions like courage and compassion. Kawakami's training procedures strike me as a tangible, straightforward way to practice being less biased, yet they are widely discounted by social scientists, activists, policymakers, and philosophers. In this essay, I have tried to speak to critics' most prominent concerns, regarding empirical efficacy, practical feasibility, and individualism. I appreciate that not everyone will be as bullish about the prospects of these specific procedures as I am, but to dismiss them wholesale is a mistake, which consigns to the shelf some tools that may have a distinctive role to play in the struggle for social justice.

Acknowledgments

For extensive feedback on several drafts, I am especially indebted to Michael Brownstein, Katherine Gasdaglis, and Jennifer Saul. I am also grateful for constructive comments on drafts or presentations of this material from Irene Blair, Tinu Cornish, Josh Ebiner, Miranda Fricker, Matthew Goren, Jules Holroyd, Derick Hughes, Erik Girvan, Pete Jones, Tim Kenyon, Lorencia Martinez, Erich Hatala Matthes, Marco Nathan, Victoria Plaut, Michelle Rheinschmidt-Same, Katherine Spencer, Christina Stephens Carbone, Daniel Silvermint, Brandon

Stewart, Manuel Vargas, anonymous referees at *Ergo*, and many others at the Implicit Bias, Philosophy, and Psychology conference at the University of Sheffield in April, 2013; Victoria Plaut's Culture, Diversity, and Intergroup Relations Lab in September, 2013; the Rocky Mountain Ethics Congress in August, 2013; my *Feminist Philosophy* class at UC-Berkeley in August, 2014; my *Feminist Philosophy of Science* class at Vassar in May, 2015; and my *Philosophy & Psychology of Implicit Bias* class at Cal Poly Pomona in May, 2016.

References

Albergotti, Robert (2014, June 30). Furor Erupts Over Facebook's Experiment on Users. *The Wall Street Journal*. Retrieved from http://www.wsj.com/articles/furor-erupts-over-facebook-experiment-on-users-1404085840

Alcoff, Linda M. (2006). *Visible Identities: Race, Gender, and the Self*. Oxford University Press. https://doi.org/10.1093/0195137345.001.0001

Alcoff, Linda M. (2010). Epistemic Identities. *Episteme*, *7*(02), 128–137. https://doi.org/10.3366/epi.2010.0003

Alexander, Michelle. (2012). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press.

Allen, Danielle S. (2004). *Talking to Strangers: Anxieties of Citizenship since Brown v. Board of Education*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226014685.001.0001

Anderson, Elizabeth (2010). *The Imperative of Integration*. Princeton University Press. https://doi.org/10.1515/9781400836826

Anderson, Elizabeth (2012). Epistemic Justice as a Virtue of Social Institutions. *Social Epistemology*, *26*(2), 163–173. https://doi.org/10.1080/02691728.2011.652211

Antony, Louise (2002). Quine as Feminist: The Radical Import of Naturalized Epistemology. In Louise M. Antony and Charlotte Witt (Eds.), *A Mind of One's Own*, *Feminist Essays on Reason and Objectivity* (2nd ed., 110–153).Westview Press.

Banks, Ralph. R. and Richard T. Ford (2008). (How) Does Unconscious Bias Matter: Law, Politics, and Racial Inequality. *Emory Law Journal*, *58*(5), 1053–1122.

Bargh, John. A. (1999). The Cognitive Monster: The Case against the Controllability of Automatic Stereotype Effects. In Shelly Chaiken and Yaacov Trope (Eds.), *Dual-Process Theories in Social Psychology* (361–382). Guilford Press.

BBC News (2009, September 28). Negative Subliminal Messages Work. Retrieved from http://news.bbc.co.uk/2/hi/health/8274773.stm

Beaulac, Guillaume and Tim Kenyon (2014). Critical Thinking Education and Debiasing (AILACT Essay Prize Winner 2013). *Informal Logic*, *34*(4), 341–363. https://doi.org/10.22329/il.v34i4.4203

Beeghly, Erin (2015). What is a Stereotype? What is Stereotyping? *Hypatia*, *30*(4), 675–691. https://doi.org/10.1111/hypa.12170

Blair, Irene V., Nilanjana Dasgupta, and Jack Glaser (2015). Implicit Attitudes. In Mario Mikulincer, Phillip R. Shaver, Eugene Borgida, and John A. Bargh (Eds), *APA Handbook of Personality and Social Psychology, Volume 1: Attitudes and Social Cognition* (665–691). American Psychological Association. https://doi.org/10.1037/14341-021

Blair, Irene V., J. E. Ma, and A. P. Lenton (2001). Imagining Stereotypes Away: The Moderation of Implicit Stereotypes through Mental Imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841. https://doi.org/10.1037/0022-3514.81.5.828

Bouton, Mark E. (2002). Context, Ambiguity, and Unlearning: Sources of Relapse after Behavioral Extinction. *Biological Psychiatry*, *52*(10), 976–986. https://doi.org/10.1016/S0006-3223(02)01546-9

Brownstein, Michael, and Jennifer Saul (Eds.) (2016). *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198766179.001.0001

Calanchini, Jimmy, Karen Gonsalkorale, Jeffrey W. Sherman, and Karl C. Klauer (2013). Counter-Prejudicial Training Reduces Activation of Biased Associations and Enhances Response Monitoring. *European Journal of Social Psychology*, *43*(5), 321–325. https://doi.org/10.1002/ejsp.1941

Dasgupta, Nilanjana and Shaki Asgari (2004). Seeing Is Believing: Exposure to Counterstereotypic Women Leaders and its Effect on Automatic Gender Stereotyping. *Journal of Experimental Social Psychology*, *40*(5), 642–658. https://doi.org/10.1016/j.jesp.2004.02.003

Dasgupta, Nilanjana and Anthony G. Greenwald (2001). On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals. *Journal of Personality and Social Psychology*, *81*(5), 800–814. https://doi.org/10.1037/0022-3514.81.5.800

Devine, Patricia G., Patrick S. Forscher, Anthony J. Austin, and William T. Cox (2012). Long-Term Reduction in Implicit Race Bias: A Prejudice Habit-Breaking Intervention. *Journal of Experimental Social Psychology*, *48*(6), 1267–1278. https://doi.org/10.1016/j.jesp.2012.06.003

Dixon, John, Mark Levine, Steve Reicher, and Kevin Durrheim (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), 411–425. https://doi.org/10.1017/S0140525X11002214

Drake, Bruce (2014, April 22). Public Strongly Backs Affirmative Action Programs on Campus. *Pew Research Center*. Retrieved from http://pewrsr.ch/1gPklxV

Eberl, Carolin, Reinout W. Wiers, Steffen Pawelczack, Mike Rinck, Eni S. Becker, and Johannes Lindenmeyer (2013). Approach Bias Modification in Alcohol Dependence: Do Clinical Effects Replicate and for Whom Does It Work Best? *Developmental Cognitive Neuroscience*, *4*, 38–51. https://doi.org/10.1016/j.dcn.2012.11.002

Ebert, Irena D., Melanie C. Steffens, Rul Von Stülpnagel, and Petra Jelenec (2009). How to Like Yourself Better, or Chocolate Less: Changing Implicit Attitudes with One IAT Task. *Journal of Experimental Social Psychology*, *45*(5), 1098–1104. https://doi.org/10.1016/j.jesp.2009.06.008

Forbes, Chad E. and Toni Schmader (2010). Retraining Attitudes and Stereotypes to Affect Motivation and Cognitive Capacity under Stereotype Threat. *Journal of Personality and Social Psychology*, *99*(5), 740–754. https://doi.org/10.1037/a0020971

Gawronski, Bertram and Joseph Cesario (2013). Of Mice and Men: What Animal Research Can Tell Us about Context Effects on Automatic Responses in Humans. *Personality and Social Psychology Review*, *17*(2), 187–215. https://doi.org/10.1177/1088868313480096

Gawronski, Bertram, Roland Deutsch, Sawsan Mbirkou, Beate Seibt, and Fritz Strack (2008). When "Just Say No" is Not Enough: Affirmation versus Negation Training

and the Reduction of Automatic Stereotype Activation. *Journal of Experimental Social Psychology*, *44*(2), 370–377. https://doi.org/10.1016/j.jesp.2006.12.004

Gibson, Bryan (2008). Can Evaluative Conditioning Change Attitudes toward Mature Brands? New Evidence from the Implicit Association Test. *Journal of Consumer Research*, *35*(1), 178–188. https://doi.org/10.1086/527341

Gifford, Robert (2011). The Dragons of Inaction: Psychological Barriers that Limit Climate Change Mitigation and Adaptation. *American Psychologist*, *66*(4), 290–302. https://doi.org/10.1037/a0023566

Glaser, Jack, and Eric D. Knowles (2008). Implicit Motivation to Control Prejudice. *Journal of Experimental Social Psychology*, *44*(1), 164–172. https://doi.org/10.1016/j.jesp.2007.01.002

Haidt, Jonathan (2001). The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haslanger, Sally (2015). Social Structure, Narrative, and Explanation. *Canadian Journal of Philosophy*, *45*(1), 1–15. https://doi.org/10.1080/00455091.2015.1019176

Henry, P.J. and Curtis D. Hardin (2006). The Contact Hypothesis Revisited: Status Bias in the Reduction of Implicit Prejudice in the United States and Lebanon. *Psychological Science 17*(10), 862–868. https://doi.org/10.1111/j.1467-9280.2006.01795.x

Hu, Xiaoqing., James W. Antony, Jessica D. Creery, Ilana M. Vargas, Galen V. Bodenhausen, and Ken A. Paller (2015). Unlearning Implicit Social Biases during Sleep. *Science*, *348*(6238), 1013–1015. https://doi.org/10.1126/science.aaa3841

Huebner, Bryce (2016). Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. *Implicit Bias and Philosophy: Volume I: Metaphysics and Epistemology* (47–79). Oxford University Press.

Johnson, India R., Brandon M. Kopp, and Richard E. Petty (2016). Just Say No! (and Mean It): Meaningful Negation as a Tool to Modify Automatic Racial Attitudes. *Group Processes & Intergroup Relations*. Advance online publication. doi:10.1177/1368430216647189

Jolls, Christine, and Cass R. Sunstein (2006). The Law of Implicit Bias. *California Law Review*, *94*(4), 969–996. https://doi.org/10.2307/20439057

Joy-Gaba, Jennifer A. and Brian A. Nosek (2010). The Surprisingly Limited Malleability of Implicit Racial Evaluations. *Social Psychology*, *41*(3), 137–146. https://doi.org/10.1027/1864-9335/a000020

Kaiser, Cheryl R., Brenda Major, Ines Jurcevic, Tessa L. Dover, Laura M. Brady, and Jenessa R. Shapiro (2013). Presumed Fair: Ironic Effects of Organizational Diversity Structures. *Journal of Personality and Social Psychology*, *104*(3), 504–519. https://doi.org/10.1037/a0030838

Kang, Jerry and Mahzarin R. Banaji (2006). Fair Measures: A Behavioral Realist Revision of "Affirmative Action". *California Law Review*, *94*(4), 1063–1118. https://doi.org/10.2307/20439059

Kawakami, Kerry, John F. Dovidio, and Simone van Kamp (2005). Kicking the Habit: Effects of Nonstereotypic Association Training and Correction Processes on Hiring Decisions. *Journal of Experimental Social Psychology*, *41*(1), 68–75. https://doi.org/10.1016/j.jesp.2004.05.004

Kawakami, Kerry, John F. Dovidio, and Simone van Kamp (2007). The Impact of Counterstereotypic Training and Related Correction Processes on the Application of Stereotypes. *Group Processes and Intergroup Relations*, *10*(2), 139–156. https://doi.org/10.1177/1368430207074725

Kawakami, Kerry, Curtis E. Phills, Jennifer R. Steele, and John F. Dovidio (2007). (Close) Distance Makes the Heart Grow Fonder: Improving Implicit Racial Attitudes and Interracial Interactions through Approach Behaviors. *Journal of Personality and Social Psychology*, *92*(6), 957–971. https://doi.org/10.1037/0022-3514.92.6.957

Kawakami, Kerry, John F. Dovidio, Jasper Moll, Sander Hermsen, and Abby Russin (2000). Just Say No (to Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation. *Journal of Personality and Social Psychology 78*(5), 871–888. https://doi.org/10.1037/0022-3514.78.5.871

Kawakami, Kerry, Jennifer R. Steele, Claudi Cifa, Curtis E. Phills, and John F. Dovidio (2008). Approaching Math Increases Math = Me and Math = Pleasant. *Journal of Experimental Social Psychology*, *44*(3), 818–825. https://doi.org/10.1016/j.jesp.2007.07.009

Kelly, Daniel, Luc Faucher, and Edouard Machery (2010) Getting Rid of Racism: Assessing Three Proposals in Light of Psychological Evidence. *Journal of Social Philosophy*, *41*(3), 293–322. https://doi.org/10.1111/j.1467-9833.2010.01495.x

Kunda, Ziva and Steven J. Spencer (2003) When Do Stereotypes Come to Mind and When Do They Color Judgment? A Goal-Based Theoretical Framework for Stereotype Activation and Application. *Psychological Bulletin*, *129*(4), 522–544. https://doi.org/10.1037/0033-2909.129.4.522

Lai, Calvin K., Kelly M. Hoffman, and Brian A. Nosek (2013). Reducing Implicit Prejudice. *Social and Personality Psychology Compass*, *7*(5), 315–330. https://doi.org/10.1111/spc3.12023

Lai, Calvin K., A. L. Skinner, E. Cooley, S. Murrar, M. Brauer, T. Devos, . . . Nosek, B. A. (2016). Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

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

Lindgren, Kristen P., Reinout W. Wiers, Bethany A. Teachman, Melissa L. Gasser, Erin C. Westgate, Janna Cousijn, . . . Clayton Neighbors (2015). Attempted Training of Alcohol Approach and Drinking Identity Associations in US Undergraduate Drinkers: Null Results from Two Studies. *PLOS ONE*, *10*(8). doi:10.1371/journal.pone.0134642.

Madva, Alex (2012). The Hidden Mechanisms of Prejudice: Implicit Bias and Interpersonal Fluency (Doctoral dissertation). Retrieved from https://academiccommons.columbia.edu/catalog/ac%3A153250

Madva, Alex (2016a). Virtue, Social Knowledge, and Implicit Bias. In Michael Brownstein and Jennifer Saul (Eds.), *Implicit Bias & Philosophy: Metaphysics and Epistemology* (Vol. 1, 191–215). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198713241.003.0008

Madva, Alex (2016b). A Plea for Anti-Anti-Individualism: How Oversimple Psychology Misleads Social Policy. *Ergo*, *3*(27), 701–728. doi:10.3998/ergo.12405314.0003.027

Martin, Douglas (2013, March 30) Yvonne Brill, a Pioneering Rocket Scientist, Dies at 88. *The New York Times*. Retrieved from http://www.nytimes.com/2013/03/31/science/space/yvonne-brill-rocket-scientist-dies-at-88.html

Mendoza, Saaid A., Peter. M. Gollwitzer, and David M. Amodio (2010) Reducing the Expression of Implicit Stereotypes: Reflexive Control through Implementation Intentions. *Personality and Social Psychology Bulletin*, *36*(4), 512–523. doi:10.1177/0146167210362789.

Morewedge, Carey K., Haewon Yoon, Irene Scopelliti, Carl W. Symborski, James H. Korris, and Karim S. Kassam (2015). Debiasing Decisions: Improved Decision Making with a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences, 2(1),* 129–140. https://doi.org/10.1177/2372732215600886

Moskowitz, Gordon B. (2010). On the Control over Stereotype Activation and Stereotype Inhibition. *Social and Personality Psychology Compass*, *4*(2), 140–158. https://doi.org/10.1111/j.1751-9004.2009.00251.x

Olson, Kristina R. and Yarrow Dunham (2010). The Development of Implicit Social Cognition. In Bertram Gawronski and B. Keith Payne (Eds.), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications* (241–254).

Olson, Michael A. and Russell H. Fazio (2006) Reducing Automatically Activated Racial Prejudice through Implicit Evaluative Conditioning. *Personality and Social Psychology Bulletin 32*, 421–433. https://doi.org/10.1177/0146167205284004

Paluck, Elizabeth Levy and Donald P. Green (2009). Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. *Annual Review of Psychology*, *60*, 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607

Parents Involved in Community Schools v. Seattle School District No. 1, 551 U.S. 701 (2007).

Pettigrew, Thomas F. and Linda R. Tropp (2006). A Meta-Analytic Test of Intergroup Contact Theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Phills, Curtis E., Kerry Kawakami, Emmanuel Tabi, Daniel Nadolny, and Michael Inzlicht (2011). Mind the Gap: Increasing Associations between the Self and Blacks with approach Behaviors. *Journal of Personality and Social Psychology*, *100*(2), 197–210. https://doi.org/10.1037/a0022159

Putnam, Robert D. (2007) *E Pluribus Unum*: Diversity and Community in the Twenty-First Century – The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, *30*(2), 137–174. https://doi.org/10.1111/j.1467-9477.2007.00176.x

Rudman, Laurie A., Richard D. Ashmore, and Melvin L. Gary (2001). "Unlearning" Automatic Biases: The Malleability of Implicit Prejudice and Stereotypes. *Journal of Personality and Social Psychology*, *81*(5), 856–868. https://doi.org/10.1037/0022-3514.81.5.856

Rudman, Laurie A. and Matthew R. Lee (2002). Implicit and Explicit Consequences of Exposure to Violent and Misogynous Rap Music. *Group Processes & Intergroup Relations*, *5*(2), 133–150. https://doi.org/10.1177/1368430202005002541

Rydell, Robert J. and Bertram Gawronski (2009). I Like You, I Like You Not: Understanding the Formation of Context-Dependent Automatic Attitudes. *Cognition and Emotion*, *23*(6), 1118–1152. doi:10.1080/02699930802355255

Schneider, David J. (2004). *The Psychology of Stereotyping*. Guilford Press.

Schuette v. Coalition to Defend Affirmative Action, 133 S. Ct. 1633, 568 U.S., 185 L. Ed. 2d 615 (2013).

Shook, Natalie J. and Russell H. Fazio (2008). Interracial Roommate Relationships: An Experimental Field Test of the Contact Hypothesis. *Psychological Science*, *19*(7), 717–723. https://doi.org/10.1111/j.1467-9280.2008.02147.x

Stewart, Brandon D. and Keith B. Payne (2008) Bringing Automatic Stereotyping under Control: Implementation Intentions as Efficient Means of Thought Control. *Personality and Social Psychology Bulletin*, *34*, 1332–1345. https://doi.org/10.1177/0146167208321269

Stout, Jane G., Nilanjana Dasgupta, Matthew Hunsinger, and Melissa A. McManus (2011). STEMing the Tide: Using Ingroup Experts to Inoculate Women's Self-Concept in Sci-

ence, Technology, Engineering, and Mathematics (STEM). *Journal of Personality and Social Psychology*, *100*(2), 255–270. https://doi.org/10.1037/a0021385

Sullivan, Margaret (2013, April 1). Gender Questions Arise in Obituary of Rocket Scientist and Her Beef Stroganoff. *The New York Times* Public Editor's Journal. Retrieved from http://publiceditor.blogs.nytimes.com/2013/04/01/gender-questions-arise-in-obituary-of-rocket-scientist-and-her-beef-stroganoff/

Sullivan, Shannon (2006). *Revealing Whiteness: The Unconscious Habits of Racial Privilege*. Indiana University Press.

Uhlmann, Eric L., Victoria L. Brescoll, and Edouard Machery (2010). The Motives Underlying Stereotype-Based Discrimination against Members of Stigmatized Groups. *Social Justice Research*, *23*(1), 1–16. https://doi.org/10.1007/s11211-010-0110-7

van Dessel, Pieter, Jan de Houwer, Arne Roets, and Anne Gast (2016). Failures to Change Stimulus Evaluations by Means of Subliminal Approach and Avoidance Training. *Journal of Personality and Social Psychology*, *110*(1), e1–e15. doi:10.1037/pspa0000039

Wal-Mart Stores, Inc. v. Dukes, 131 S. 2541, 564 U.S. 277, 180 L. 2d 374 (2011).

Weisbuch, Max, Kristin Pauker, and Nalini Ambady (2009). The Subtle Transmission of Race Bias via Televised Nonverbal Behavior. *Science*, *326*(5960), 1711–1714. https://doi.org/10.1126/science.1178358

Wennekers, Annemarie M. (2013). Embodiment of Prejudice: The Role of the Environment and Bodily States (Doctoral dissertation). Radboud University Nijmegen, Netherlands.

Wennekers, Annemarie M., Rob W. Holland, Daniel H. Wigboldus, and Ad van Knippenberg (2012). First See, Then Nod: The Role of Temporal Contiguity in Embodied Evaluative Conditioning of Social Attitudes. *Social Psychological and Personality Science*, *3*(4), 455–461. https://doi.org/10.1177/1948550611425862

Wiers, Reinout W., Carolin Eberl, Mike Rinck, Eni S. Becker, and Johannes Lindenmeyer (2011). Retraining Automatic Action Tendencies Changes Alcoholic Patients' Approach Bias for Alcohol and Improves Treatment Outcome. *Psychological Science*, *22*(4), 490–497. https://doi.org/10.1177/0956797611400615