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


## Abstract

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 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 shown that "counterstereotype training," which involves repeatedly affirming counterstereotypes, and "approach training," which involves practicing approach-oriented behaviors toward stigmatized words and images, lead to significant reductions in implicit prejudice, stereotype accessibility, and even susceptibility to stereotype threat.[1] These training procedures don't just influence scores on indirect measures like the Implicit Association Test (IAT); they debias unreflective social behaviors (leading white and Asian participants to instinctively sit closer to a black interlocutor) and deliberative decisions

[1] On reducing stereotype accessibility and implicit prejudice, see Kawakami et al. (2000), Kawakami, Dovidio, and van Kamp (2005, 2007), Kawakami et al. (2007), Gawronski et al. (2008), Johnson (2009), Stewart et al. (2010), Phills et al. (2011), Wennekers et al. (2012), Wennekers (2013), and Calanchini et al. (2013). On reducing stereotype threat, see Kawakami et al. (2008), Forbes and Schmader (2010), and Stout et al. (2011). For further debiasing procedures and quasi-experimental demonstrations, see Dasgupta and Greenwald (2001), Rudman et al. (2001), Blair (2002) Dasgupta and Asgari (2004), Plant et al. (2005), Olson and Fazio (2006), Dasgupta and Rivera (2008), Joy-Gaba and Nosek (2010), and French et al. (2013). For a meta-analysis of prejudice reduction strategies, see Paluck and Green (2009).

about job candidates (reducing the likelihood that participants choose a man over an equally qualified woman), and improve performance on math tests. 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 (see, e.g., Alcoff 2010; Anderson 2012; Bargh 1999; Huebner forthcoming; Mandelbaum 2014; Mendoza et al. 2010; Stewart and Payne 2008). For example, David Schneider's opus on social cognition, *The Psychology of Stereotyping* (which, not including references, totals 568 pages), 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).

Why are these "debiasing" procedures so readily written off? There are a handful of frequently cited reasons, which 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) Many critics simply don't believe that these interventions will really work. Many suspect that individuals will quickly "relearn" their biases upon leaving the lab, or that the effects of debiasing will hold only in highly specific contexts—effective in the lab but not the "real world" (Alcoff 2010; Anderson 2012; Devine et al. 2012; Gawronski and Cesario 2013; Huebner forthcoming; Mandelbaum 2014; Mendoza et al. 2010; Stewart and Payne 2008; Olson and Fazio 2006; Wennekers 2013).

(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 (*ibid*.).

(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 (Alcoff 2010; Anderson 2010, 2012; Banks and Ford 2008; Dixon et al. 2012; Haslanger 2012, 2013; Huebner forthcoming). 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.

Here I reply to these criticisms of debiasing, although I give a more wide-ranging response to the concerns about individualism in a companion paper.[2] I begin by surveying the relevant research (§2), going into some depth where the details are important for addressing concerns about debiasing.[3] I briefly discuss, and express puzzlement over, how theorists tend to construe the "real-world implications" of these findings (§3). I then address each of the major concerns about debiasing (§§4-7), and conclude by sketching a few strategies individuals can employ in daily life to debias themselves (§8). Ultimately, I argue, a comprehensive strategy for combating prejudice and discrimination should include a central role for training our biases away.

## 2. Research survey

In Kawakami and colleagues' seminal 2000 paper, "Just Say No (to Stereotyping)," participants repeatedly "negated" stereotypical associations and "affirmed" counterstereotypical associations. They saw images of racially typical black and white male faces paired with potentially

[2] "A Plea for Anti-Anti-Individualism," in preparation.
[3] Another reason for describing in considerable depth the extant research on counterstereotype and approach training is that no comparable survey (let alone meta-analysis) focusing specifically on these retraining procedures has been published.

stereotypical traits. If they saw a stereotypical face-word pairing, such as a black face paired with the word "athletic," they pressed a button labeled "NO." If they saw a counterstereotypical pairing, such as a white face paired with "athletic," they pressed a button labeled "YES."[4] Participants worked through 4 blocks of 96 face-word pairings, totaling at 384 "trials." Including time to rest between blocks, this took under 45 minutes.

The procedure, dubbed "negation training," was sandwiched between a pretest and a posttest of automatic stereotype activation. Unlike participants who repeatedly *affirmed stereotypes* or underwent no training at all, those who underwent negation training went from being biased to unbiased on this measure—no longer showing any significant influence of stereotypes on behavior.[5] The researchers also found that negation training eliminated the automatic activation of skinhead stereotypes.[6] These effects persisted when they were tested again after 2, 4, 6, and 24 hours. In fact, participants were even *less* biased the next day (presumably because they weren't cognitively burned out from all the training).[7] Kawakami and

[4] One commentator expressed concerns about whether the buttons were actually "labeled" or not. In the first two studies, the participants actually pressed buttons that had the words "YES" and "NO" on them. In a third study, they just pressed the M and Z keys on a computer keyboard. Participants can just be told that pressing some arbitrary button *means* "affirming" or "negating" something (in other words, like all language users, they can learn the references of arbitrary symbols or actions). There is, nevertheless, a significant question about what these actions of pressing buttons really signify to the participants—are they *really* negating stereotypes when they go through these motions?—a question which is raised in one follow-up study by Johnson (2009) that I discuss below.

[5] There were no significant differences between pre- and posttests for participants who were trained to affirm stereotypes or had no training at all. In the third study, the test of stereotype activation involved priming participants with potentially stereotypical words and then measuring how long they took to identify a face as black or white. In the first two studies, the test of stereotype activation was a variant of the Stroop task, wherein "participants, following the presentation of [social category primes, like SKINHEAD or ELDERLY], were instructed to name the ink color of skinhead stereotypes (e.g., criminal) or elderly stereotypes (e.g., afraid) as quickly as possible. If stereotype activation is automatic in the pretest of the primed Stroop task and participants have not yet learned to inhibit this activation, participants will be slower at color-naming stereotypes because they are unable to ignore their content and focus on the naming of the ink colors" (872). The idea is that if you see the word "skinhead" and then the word "vandal" in green, stereotype activation will prime you to read the word, making you slower to identify the green color than if you see "skinhead" followed by "forgetful." They also found that novel stereotypical traits, which were not part of the training, did not activate stereotypes posttest.

[6] In the skinhead training, they only saw the word "skinhead," rather than images of skinheads' faces.

[7] The researchers had also intended to train away stereotypes about the elderly, but they failed to find evidence for automatic stereotype activation against the elderly during pretest. While negation training completely eliminated

colleagues wrote, "In short, practice does make perfect—or at least very good—stereotype negators" (884).

Jimmy Calanchini and colleagues (2013) replicated Kawakami's findings using generic positive and negative words, rather than racial stereotypes, in the training task. They also found that this counter-prejudice training simultaneously reduced implicit racial bias and increased the capacity to control the expression of bias. Two additional studies outside of Kawakami's lab have partially replicated but partially qualified these findings. First, Bertram Gawronski and colleagues (2008) observed that the original studies confounded two sorts of training—the repeated *affirmation* of counterstereotypes and the repeated *negation* of stereotypes. Gawronski and colleagues 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 counterstereotypical pairings and instructed others to simply negate the stereotypical pairings. After 200 trials, participants who repeatedly affirmed counterstereotypes showed significant reductions in implicit biases, while those who negated stereotypes showed exacerbated implicit biases. The upshot according to the researchers is to just say "Yes" to counterstereotypes, rather than "No" to stereotypes:

> More precisely, the present findings suggest that thinking about stereotyped groups or individuals in counterstereotypical terms (e.g., "old people are good drivers") is more effective in reducing unwanted stereotyping than attempts to negate an existing stereotype (e.g., "it is not true that old people are bad drivers") (376).

Gawronski and colleagues found that affirmation training reduced implicit *gender stereotyping* and implicit *racial prejudice*.[8] Moreover, retraining racial *stereotypes* led to changes in racial

---

skinhead stereotype activation, it is important to bear in mind that not all groups and stereotypes are created equal. There is much more research to be done regarding which stereotypes are and are not automatically activated in which contexts, what the effects of their activation are, and so on.

[8] The training in Study 1, on gender stereotyping, involved pairing typical male and female names with traits relating to strength ("mighty") vs. weakness ("dainty"). The stereotyping measure was a sequential priming task, wherein participants saw the names immediately followed by the traits, pressing one button if they saw a strength word, and another if they saw a weakness word. The training and the measure were highly similar in this particular

*prejudice*. For those who make a lot of hay out of the distinction between semantic, belief-like attitudes (aka stereotypes) and affective, motivational attitudes (aka prejudices), it is noteworthy that retraining putatively "cold" cognitive attitudes led to changes in affective responses.[9]

Gawronski and colleagues hypothesize that the primary factor driving changes in implicit bias was simply "enhanced attention" to one rather than another set of stimuli (375). However, India Johnson raises the possibility that the negation training was too *half-hearted*: "not strong enough, or not meaningful enough" (2009, 12). Gawronski's training just involved pressing the space bar, and, depending on the condition, participants were simply told that pressing it "meant" negating stereotypes or affirming counterstereotypes. So Johnson put some pizzazz in the space-bar pressing by instructing participants that the action was equivalent to saying "*No! That's wrong*!"[10] Johnson found that 200 trials of this "meaningful negation" of stereotypes did in fact

---

study, raising the possibility that participants were getting better at producing unbiased responses on this particular measure while leaving their underlying biases intact. This concern is addressed in Study 2, in which the training paired black and white faces with positive, stereotypically white traits ("intelligent," "wealthy") vs. negative, stereotypically black traits ("poor," "lazy"). The measure was a subliminal affective priming task, with the masked words "black" or "white" followed by generic positive or negative words (such as "paradise" or "rotten"). Study 2 is thus much more impressive than Study 1, because the training and the measure used completely different stimuli: face-with-stereotypical-trait pairings during training versus subliminal-race-word-with-generic-evaluative-word pairings during testing.

[9] This is, in other words, evidence that these two types of implicit "association" might not be so clearly distinct as some have argued (e.g., Amodio and Devine, 2006; Holroyd and Sweetman, forthcoming). Admittedly, many of the stereotype-related words used in training were clearly affect-laden: "Trait words related to the negative stereotype of Black people: *poor, dishonest, complaining, violent, shiftless, superstitious, lazy, threatening, dumb, hostile*… Trait words related to the positive stereotype of White people: *intelligent, successful, ambitious, industrious, educated, responsible, wealthy, ethical, smart, friendly*" (376, original emphasis). But the important point is that the words used to test automatic evaluation were (almost) entirely *unrelated*, semantically speaking, to the relevant racial stereotypes: "Positive target words: *paradise, summer, harmony, freedom, honesty, honor, health, cheer, pleasure, heaven, friend, sunrise, love, relaxation, peace, vacation, happy, lucky, miracle, gift*… Negative targets words: *evil, sickness, vomit, bomb, murder, abuse, prison, death, assault, cancer, rotten, accident, grief, poison, stink, cockroach, virus, disaster, ugly, terror*" (376, original emphasis). I argue elsewhere that most objectionable stereotypes are *inherently* evaluative and affect-laden, which is a reason to doubt the legitimacy of sharp distinctions between stereotyping and prejudice.

[10] Of course, it's still reasonable to wonder what this performance really means to the participants. Johnson's label for the activity—"meaningful negation"—might be misleading. Objecting "That's wrong!" in response to something is not equivalent to objecting "That's false!" Saying or thinking that something is wrong might be effective by virtue of repeatedly generating more palpable, salient negative affect, in the manner of moral and emotional indignation or outrage, rather than by virtue of repeatedly thinking that a stereotype is false or misleading and should be "negated." Participants might be cultivating negative affective responses to stereotypes (i.e.,

reduce implicit prejudice.[11] Interestingly, she also found that meaningful negation was most effective for individuals who reported being strongly motivated to be unprejudiced.

In addition to reducing implicit stereotyping and prejudice on various indirect, computer-based measures, these debiasing procedures seem to influence "real-world" social behaviors.

In Kawakami, Dovidio, and van Kamp (2007), participants first underwent gender counterstereotype training, by pairing men's faces with words like "sensitive" and women's faces with words like "strong."[12] They next evaluated four applications (résumés and cover letters) ostensibly for a position as "chairperson of a District Doctor's Association" (143). All of the applicants were qualified, but two had male names and two had female names (counterbalanced so that half the participants saw a particular résumé with a man's name and the other half saw that same résumé with a woman's name). The evaluation of applicants involved two separate stages: judging the applicants along 16 different dimensions (8 stereotypically masculine traits like "risk-taker" and 8 feminine traits like "helpful") and then simply choosing the best candidate. Some participants made the trait judgments first and chose the best candidate second, while other participants completed the two tasks in the opposite order.

Among participants who had received no training, only 35% chose a woman for the job.

---

evaluative conditioning), rather than "convincing" themselves that stereotypes are false. In this way, the invocation of affect in "meaningful negation" might be importantly similar to the "approach/avoid" training I discuss below.

[11] Johnson also found that meaningfully negating *counter*stereotypes led to *increases* in automatic prejudice, and failed to replicate Gawronski's finding that non-meaningfully negating stereotypes increased automatic stereotyping and prejudice. I'm not inclined to read too much into the inconsistent findings regarding whether variants of these training procedures can *increase* bias. Most adults' automatic dispositions toward stereotyping and prejudice are well-established and probably close to ceiling, so we should generally expect to see weaker effects when it comes to exacerbating biases than when it comes to reducing them. This is in keeping with almost all of Kawakami and colleagues' articles on counterstereotype and approach training, which include conditions wherein participants repeatedly practice the stereotypical or prejudicial responses. There's often just a non-significant or marginally significant trend toward exacerbation in those conditions.

[12] Following-up on Kawakami, Dovidio, and van Kamp (2005), participants were repeatedly shown photos of men or women above pairs of words, such that one word was stereotypically associated with the gender of the face, and the other word was not (e.g., a woman's face above the words "sensitive" and "strong"). Participants in the relevant experimental condition had to consistently choose the trait that was *not* stereotypically associated with the face.

Bearing in mind that the gendered names and résumés were randomly mixed and matched for different participants, this can pretty much only be interpreted as evidence for a majority preference for giving the leadership position to a man. Yet among participants who *had* undergone counterstereotype training, 61% chose a woman. These are striking data; however, there is an equally striking catch. These effects were only observed when the task of choosing the best candidate came *second*, after the trait evaluation. When this choice task was first, only 37% of those who had undergone the training chose a female candidate. A similar pattern emerged when the order of the tasks was switched, in that participants were consistently *biased* on the first task and *debiased* on the second, regardless of which task actually came first.

What's going on here? Participants seem to recognize that the researchers are trying to debias them, and then try to correct for this perceived influence by deliberately responding in more stereotypical ways, at least at first. Once they have an opportunity to explicitly counteract the debiasing, they stop trying to resist the training and *then* the effects emerge. Subsequently, they respond in *counter*stereotypical ways. The psychological mechanisms underlying all of this are up for grabs, but the researchers take these findings to "have direct implications for the effectiveness of certain types of anti-bias programs. Strong interventions to reduce bias which appear 'heavy-handed' may arouse correction motivations, at least initially, to control for these influences" (151). An implication seems to be that even though, say, a white male employer might express resentment or discomfort *during* or *immediately after* an intervention aimed at reducing discrimination in the workplace, the effects of that intervention might nevertheless emerge later. Given that Johnson (2009) found differential effects for counterstereotype training depending on participants' reported concerns about prejudice, further research should examine whether these initial "correction" effects are more likely to occur for participants *not* strongly

motivated to be unprejudiced, or from more privileged backgrounds, and so on.

Using an altogether different procedure, Kawakami, Phills, Steele, and Dovidio (2007) found that participants can change their implicit biases and unreflective social behaviors by practicing "approach" and "avoidance" behaviors. 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. This training significantly reduced participants' implicit racial prejudice on the IAT.[13] In some cases, participants were explicitly told that moving the joystick would metaphorically signify either approaching or avoiding the images of faces, while in other cases they were merely instructed how to move the joystick, without any explanation of why. In still further cases, the images of the faces were "masked" and shown so quickly participants didn't notice them, and instead believed that they were just moving the joystick when they saw the words "approach" or "avoid." Significant effects were found in each of these cases, regardless whether the meaning of the training was explicit or subliminal. Subjects were also interviewed regarding whether they knew what the point of the experiment was; in the subliminal condition, they didn't. Perhaps this subliminal training precludes the temporary backlash observed in the previous study (although I expect that subliminal training

[13] It bears mentioning that the IAT, which was also used by Calanchini et al. (2013), is a different measure of implicit prejudice from the affective priming measure used in Gawronski et al. (2008) and Johnson (2009). That is, a variety of different indirect measures demonstrate significant effects of these training procedures. In Study 1, participants in a control condition who just moved the joystick left or right showed less posttest bias (*D* was 0.43) than participants who approached whites and avoided blacks (*D* was 0.52), and more posttest bias than those who approached blacks and avoided whites (*D* was 0.23). In Study 3, participants saw Asian faces instead of white faces, but still showed reduced implicit racial bias on the standard black-white race IAT, suggesting that approaching blacks is effective somewhat independently of avoiding whites. Wennekers (2013) also found that approaching faces and avoiding images of *closets* reduced prejudice. These findings are, prima facie, in conceptual tension with recent work on other forms of "counterstereotype exposure." Joy-Gaba and Nosek (2010) found that exposure to admired black exemplars doesn't reduce implicit prejudice without exposure to disliked white exemplars, and even then the effect sizes are much smaller than originally reported by Dasgupta and Greenwald (2001).

will strike some as a decidedly creepier variant of an already spooky approach to prejudice reduction). Most important, Kawakami and colleagues found that subliminal approach training influenced *actual social behavior*, leading white and Asian participants to sit closer to a black interlocutor (a confederate posing as a fellow student) and face him head-on, rather than at an indirect angle.

In the Netherlands, Annemarie Wennekers and colleagues (2012, 2013) replicated the effects of embodied retraining techniques on implicit prejudice by having participants nod in response to typical Moroccan names and shake their heads in response to typical Dutch names, and also found that the "effects seem to be stronger for people who did not report to be aware of the goal of the study" (116). Wennekers (2013) also 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 and Fazio 2006, 431, for further discussion), which is a reason to be skeptical about how effectively we can replicate these lab-based interventions in daily life (§3). We cannot expect to approach or have positive social interactions with every member of a particular social group, stigmatized or otherwise.[14]

These debiasing procedures can also be employed to *help ourselves* cope with the stereotypes that might negatively affect *us*. Kawakami and colleagues (2008) reported the beneficial effects for undergraduate women of repeatedly approaching math-related images ("e.g., calculators, equations"). Those who initially reported that they did not like math and were not good at it tended, after the training, to identify with and prefer math on computer-based measures, as well as to answer more questions on a math test. A series of follow-up studies by

[14] Wennekers and colleagues also found that nodding *after* seeing the stigmatized stimulus (as if nodding in *response* to something) effectively reduced bias, but that nodding *before* seeing the stimulus did not. The spatiotemporal ordering and "bodily meaning" of stimuli and behaviors in these training procedures seems important.

Forbes and Schmader (2010) replicated these findings using a different training procedure, and with a 24-30 hour delay between the training and the math test.[15] They also found that gender-math *counterstereotype* training seemed more effective than approach training. Women subtly trained to associate the phrase "women are good at" with math-related words exhibited increased working memory as well as improved performance on math questions from the GRE.

Taken together, counterstereotype and approach training seem to be effective procedures for debiasing ourselves along a number of key dimensions, influencing a host of measures of cognitive and emotional attitudes, as well as unreflective social behavior, deliberative decision-making, and test-taking. This seems like it should be a big deal.

## 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 debiasing 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 particular procedures in the field? Policymakers already "spend billions of dollars annually on interventions aimed at prejudice reduction in schools, workplaces, neighborhoods, and regions beset by intergroup conflict"

[15] In lieu of approach training, Forbes and Schmader (2010) adapted the "personalized IAT" to train women to associate the phrase "I like" with math-related words. Their counterstereotype training was adapted from the standard IAT. One of the notable features of these studies, the results of which are fascinating and shed a great deal of light on the complexity of stereotype threat, is that they show that the IAT is not just a measure of the mind but can be used to *influence* attitudes and stereotypes. The training procedures were sufficiently indirect that "few participants revealed any awareness" of the aims of the study.

(Paluck and Green 2009, 340). Yet nobody, to my knowledge, has seriously advocated implementing these debiasing procedures in these contexts, not even on an exploratory, experimental basis.

Instead, Kawakami's supposedly "laborious 480-trial procedure" (Olson and Fazio, 2006, 431), which requires "many, many repetitions to learn nonstereotypical responses" (Stewart and Payne, 2008, 1343), is often cited as a point of contrast when researchers discover a less intensive, demanding intervention.[16] Many continue to assume that implicit biases are, despite evidence for their partial malleability, still a little too rigid, inaccessible, and unwieldy for changing them directly to be a viable strategy, and so are committed to finding interventions that require less time and effort, and which work primarily by leaving the biases in place and enhancing individual self-control over them. As Keith Payne explained in an interview, "If you boil it down, the solution sounds kind of easy: just maximize control. But how do you do that? As it plays out in the real world, it's not so easy."[17]

Nevertheless, many acknowledge that Kawakami's research might have *indirect* practical

[16] A striking example of a less intensive intervention is Stewart and Payne's (2008) finding that a race-weapon bias could be reduced simply by rehearsing an implementation intention ("Whenever I see a Black face on the screen, I will think the word, safe"). Stewart and Payne, in contrast to Kawakami, claim to be "providing participants with a specific control strategy that required little effort and that they could employ on demand" (1343). See Mendoza et al. (2010, 521) for a similar contrast between implementation intentions and more intensive debiasing procedures. I am all for employing implementation intentions in the struggle against discrimination. If-then plans like this have proven to be incredibly effective in a wide variety of domains (Gollwitzer and Sheeran, 2006). In fact, we should use them in the instructions for debiasing procedures. Individuals can also be encouraged to rehearse the relevant implementation intentions *after* they undergo the training procedures described above, to help them stay debiased.

[17] Reported by Carpenter (2008). See Mendoza et al. (2010, 512-3) for similar sentiments. I read Payne here as talking about what I refer to elsewhere (2012) as *local control*, regulating expressions of bias on particular occasions, but he might also be talking about control in very general terms, to include *long-term control*, which includes strategies that change our underlying biases, in which case I agree with him. Any strategy to overcome biases is, in that sense, a strategy for controlling them. But Payne seems to think it obvious that we should figure out ways to maximize local control in particular. What puzzles me is that we are not also looking as seriously at strategies that change the underlying biases themselves, and thereby render control over them superfluous. Self-control is and will ever be a limited cognitive resource, which gets depleted when we use it, and which we can only use when we realize we are in a context that requires it. As I see it, these limitations mean that self-control should be the stopgap measure we rely on when other options are unavailable. If we just get rid of the biases, there's nothing to control. By contrast, Mendoza et al. (2010, 512-3) seem to suggest that the primary reason to explore debiasing procedures like Kawakami's at all is that immediate, local control might seem unavailable.

import: there is a strange trend of assuming that these studies must be "translated" somehow out of their artificial laboratory context into an applied, "real-world" setting, as if they are only relevant if we can figure out how to mimic them in our everyday social lives. Wennekers (2013, 85) writes that "repeatedly approaching out-group members and noticing that nothing bad happens may make you less likely to avoid them." Schneider (2004, 423) 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." Researchers thus seem to think that the practical upshot of these findings is simply that they indicate how those who happen to be lucky enough to have lots of positive experiences with counterstereotypical exemplars in daily life might become less biased. Philosophers such as Linda Martín Alcoff (2010, 131-2), Elizabeth Anderson (2010, 152; 2012, 167-70), and Bryce Huebner (forthcoming, §4.2) draw similar conclusions.

Even Phills, Kawakami, and colleagues (2011) seem to assume that these debiasing 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. These activities not only may strengthen the extent to which students identify with members of other social categories but also may increase their sense of belonging and academic achievement. (208)

This sort of welcoming activity may be beneficial, but it strikes me as odd to construe it as somehow in competition with the *actual* debiasing procedures studied in the lab. It would probably be a good idea to debias students *before* the relevant social activity.

Real-world attempts to change attitudes through social interaction (the "contact hypothesis") have a long history, and evidence for their success is substantial but complicated

(see, e.g., Dixon et al. 2012; Kelly, Faucher, and Machery 2010; Pettigrew and Tropp 2006; and Putnam 2007). 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. Social 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 implicit biases of the higher-status group remained unaffected. Even when contact reduces prejudice, the effect sizes tend to be relatively small, and the conditions conducive to effective social contact (namely, cooperating toward a common goal, on terms of equal social status) are difficult to construct and maintain. A rival "conflict" hypothesis seeks to explain how social contact often *amplifies* intergroup animosity.

I by no means wish to discount the importance of actual intergroup interaction. 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 impediment to just democratic decision-making, and that, therefore, integration and interaction between members of diverse social groups are deeply important aims. As Allen (2004, 182-3) writes, "The development of new norms for the interaction of strangers within the polis requires public discussion *among* strangers. Trust grows only through experience; habits of citizenship are fashioned only through actual interaction." Similarly, Anderson (2010, 152) claims that "there is one particular kind of know-how that can only be constructed in integrated groups: knowledge of how to effectively cooperate and communicate on terms of equality across group lines, in a relaxed and comfortable way."

At the same time, Allen recognizes that, "Experience can, however, undo the basis of

trust as easily as it establishes it" (56). As I explain in §4, putting people in a room together, or simply initiating a spontaneous conversation with a member of a different social group, could as easily amplify prejudice as it could mollify it. I do not believe that we should pin much hope on getting white students to unlearn their implicit prejudices toward their black and Latino classmates *simply* by having them shake hands in homeroom. We should actively pursue complementary strategies to nudge such intergroup interactions in the right direction. Maybe if these students volunteered for a little approach training beforehand, these encounters would be more likely to start off on the right foot and unfold in more positive ways.

There are, nevertheless, important connections between approach training and the contact hypothesis. Phills, Kawakami, and colleagues (2011) found that these embodied approach behaviors 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.).[18] In fact, increases in self-black associations seemed to *mediate* the reduction of implicit bias. Approach training evidently increased self-identification with blacks, and this self-identification in turn reduced bias. In a certain sense, then, this research is in keeping with the age-old strategy of reducing prejudice by breaking down "us" vs. "them" dichotomies.[19] Approach training may be, in effect, *the contact hypothesis in a bottle*. This is not to say that approach training or its

[18] "In particular, because approach behaviors imply a decrease in distance and increased physical closeness between the self and an object, approach orientations will result in accentuated psychological closeness between the self and the target" (198). One study even found neural evidence for increased self-black associations on an EEG. Phills and colleagues found similar effects using a novel sort of approach-training computer program, wherein white and Asian participants repeatedly moved circles containing their own names so that they overlapped with circles containing images of black faces (10 blocks of 24 trials, totaling at 240). They also found that participants in this approach training condition formed a stronger association with blacks and self-ascribed *traits*. The experimenters had polled them a week earlier asking which positive and negative traits they were most likely to self-ascribe. "Participants trained to approach Blacks ($D$ = .02, $SD$ = .17) were faster to associate the specific traits that they ascribed to the self with Blacks than participants trained to avoid Blacks ($D$ = -.13, $SD$ = .20)" (202).

[19] For more on the importance of implicit self-identification and sharing similarities for intergroup relations and stereotype threat, see Stout et al. (2011) and Mallett, Wilson, and Gilbert (2008).

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 lab-based debiasing procedures: we can guarantee that 100% of the trials are counterstereotypical in the lab, but not that 100% of social encounters are counterstereotypical (or cooperative, respectful, etc.) in the "real world." (I do, nevertheless, believe that there are a number of complementary debiasing strategies we can implement in daily life; I discuss a few examples in §8.)

It seems to me that these very debiasing procedures, or close variants of them, 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 looking solely to "real-world" and imperfect translations of these procedures, we should be making these procedures widely available (e.g., on the Project Implicit website[20]), and considering ways in which institutions might incorporate debiasing into broader antidiscrimination strategies. Of course, before we make serious investments of hope and resources, further research could test the effects of these procedures in the field. I highlight several outstanding empirical questions in what follows. However, even researchers like Kawakami and Wennekers 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 on how to translate or mimic these

[20] This website currently has a litany of IATs that anyone can take (projectimplicit.net). Since Forbes and Schmader (2010) used variants of the IAT for debiasing, it would seem to be incredibly straightforward to make some potentially debiasing IATs widely available (and to test their effects on large populations).

procedures in daily-life interactions, rather than using these procedures themselves. Why?

## 4. 1st empirical concern: the "relearning" worry

Far and away, the most commonly cited concern, about these and pretty much every other individual-level strategy for reducing prejudice, is how long the effects last.[21] To my knowledge, nobody has tested how long people stay debiased after counterstereotype or approach training. The durability of debiasing 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.[22] I worry, however, that pessimism about the durability of debiasing is another contributing factor, in which case this pessimism becomes a self-fulfilling prophecy: nobody actually tests it because everybody expects it to come out a certain way.

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, Saaid Mendoza and colleagues (2010, 520) write that attempts "to change underlying representations of racial groups… may be more difficult to maintain upon reexposure to societal stereotypes outside the laboratory."[23] And Bryce Huebner (forthcoming) writes, "as we watch or read the news, watch films, rely on tacit assumptions

[21] See, e.g., Devine et al. (2012, 1267-8), Mendoza et al. (2010, 520-1) and Wennekers (2013, 130-1), who also cites clinical research using similar procedures, which "show strong effects, but also high levels of relapse in the long run."

[22] In conversation, psychologist Brandon Stewart suggested that another contributing factor is a stigma in academic psychology against doing work that is too applied and insufficiently theoretical. Anecdotally, I have also heard rumblings of unpublished, unsuccessful replications of these debiasing techniques.

[23] Mendoza et al. cite no evidence for this claim, however, because there is none. They do, however, cite studies of debiasing interventions with "effects lasting a day or two," in contrast to studies on implementation intentions, which showed effects lasting weeks or months (520). This paragraph comes dangerously close to inferring evidence of absence from absence of evidence, because they cite the studies that showed effects lasting 24 hours as if they also *failed* to show effects lasting longer. But these studies simply did not test longer-term effects.

about what is likely to happen in particular neighborhoods, or draw elicit inferences on the basis of the way in which a person is dressed, we cause ourselves to backslide into our implicit biases." 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 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, 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 remain stable through adulthood (Dunham et al. 2008; for reviews, see Olson and Dunham 2010; Ziv and Banaji 2012). 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. 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 relearn or 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.[24] 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 and stereotyping over the past few 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.[25]

Another source of pessimism is evidence that exposure to certain forms of "mass media" enhances implicit bias. For example, implicit racial biases increase after listening to violent rap

[24] Infants start picking out social categories and acquiring biases about category members in their first months (as measured by looking time). One part of the explanation for the rapid formation of group biases seems to be an "automatic ingroup-related positivity" and another part seems to be the "rapid internalization of (directional) group status," such that individuals quickly form positive attitudes toward high-status groups and negative attitudes toward low-status groups. Children seem to acquire both implicit and explicit biases very quickly.

The relevance of group status to implicit bias is also visible in social contact research. Intergroup contact seems more likely to reduce the implicit biases of the disadvantaged than the advantaged (Henry and Hardin 2006). If borne out, these findings on the effects of group status might constitute important examples of *how* unjust social structures *per se* support implicit biases. It is not just that kids become biased by seeing too many stereotypes on TV; their biases reflect real-world power disparities. Score one for the revolutionaries who think we can't change implicit biases without overhauling social structures, although another response to this research (which seems consistent with a politically radical outlook but does not require it) might be to emphasize to our kids and to ourselves, in ways that I discuss below, that these differences in group status are *wrong* and unfair and ought to be changed.

[25] It also seems to be the case that, if you really take the relearning worry seriously, you should be pessimistic about *a lot more* than just these specific debiasing strategies. For example, we shouldn't bother with implementing the school-based social-contact activity suggested by Phills et al. (2011) above, because as soon as the students leave school and turn on the radio or the television, or open a newspaper, they are going to get bombarded with stereotypes and "lose" all the egalitarian psychological currency they just acquired. (Those who think real prejudice reduction can only be wrought through a thoroughgoing social revolution should be nodding their heads at this point.)

music (but not pop; Rudman and Lee 2002), and after watching television clips in which white characters display subtle, nonverbal bias toward black characters (Weisbuch, Pauker, and 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. 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.[26] Typically, belief perseverance, the confirmation bias, and a host of other cognitive dispositions help to create and sustain our implicit 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

[26] Indeed, it is obviously the case that, for at least some stereotypes and prejudices, we acquire them without sufficient evidence and maintain them despite the good evidence against them (even if we "count" exposure to distorted media representations as evidence). For example, one of the disheartening findings from developmental psychology is that children's acquisition of biases is *accelerated* simply by virtue of learning the names for certain social groups. Often, all children need to do is learn the name to acquire the bias—no gradual accrual of evidence required (Leslie 2009).

and how we interpret it is profoundly shaped by our implicit and explicit goals.[27] 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 make 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 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 less attention to the stereotypes, and perhaps "meaningfully negate" or otherwise discount them when we notice them. Of course, this is clearly speculative. My aim is not to convince you through a priori speculation that debiased individuals will never relearn their implicit 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 debiasing 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 in the lab will instantly endow us with all the right

[27] Building on Kunda and Spencer (2003), Moskowitz (2010) reviews a wide array of ways in which implicit social cognition depends on an agent's goals. See Uhlmann, Brescoll, and Machery (2010) for evidence that stereotyping is driven by epistemically questionable aims, rather than by the aim *to be accurate*. These are good candidates for goals we should teach children *not* to have.

cognitive dispositions—but debiasing should clearly 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, evidence for the potential durability of debiasing is growing. Patricia Devine and colleagues (2012) taught participants five strategies they could employ in daily life to reduce their racial biases. This intervention led to reductions of bias that lasted at least 8 weeks (implicit biases were even slightly lower after 8 weeks than after 4). Notably, participants' reported concerns about discrimination also increased, and this increased concern seemed to significantly enhance bias reduction. Evidence also suggests that counterstereotypical teachers can reduce their students' implicit 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 et al. 2001 and Stout et al. 2011). Presumably, the participants in this study were simultaneously being bombarded with stereotypical representations of women as nurturing and men as assertive 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.[28] Yet their salient classroom experiences evidently "won out" over the media bombardment.

Perhaps the strongest evidence for the durability of these interventions comes from clinical research. Reinout Wiers and colleagues (2011) found that patients recovering from alcoholism who, immediately prior to undergoing standard treatment, were trained to avoid

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

images of alcohol (in 4 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 prior to standard treatment. These impressive findings were replicated by Carolin Eberl and colleagues (2013). Of course, no one is making the absurd claim that moving a joystick back and forth will, all by itself, cure alcoholism. 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 debiasing procedures is pessimism about durability, but such pessimism is, at this time, unwarranted. Wiers and Eberl's studies suggest that, if anything, these procedures *enhance* the durability of standard interventions. I doubt that implicit social prejudices are more difficult to dislodge than addictive impulses, but we obviously cannot assume that approach training has long-term effects on prejudice reduction comparable to those on alcoholism recovery. The long-term effects of these procedures on prejudice reduction 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?

## 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 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.[29] 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.

Robert 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 pretty much confirmed all of our worst fears.[30] 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

[29] If the formulation of this toy example offends your sensibilities about semantics and pragmatics, please ignore it.
[30] See Gawronski and Cesario (2013) for a review.

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 a bunch of 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 pretty 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 debiasing interventions out of the hundreds of studies that purport to reduce implicit 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 racial *stereotypes* led to changes in implicit racial *prejudice*, even though all the stimuli during training and testing were

completely different (Gawronski et al. 2008).[31] 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, et al. 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 measures of implicit stereotyping as well as women's performance on tests of working memory and math at least a day later (Forbes and 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).[32] There seems to be substantial evidence that these procedures generalize to at least some novel contexts, and, indeed, to precisely those contexts we're most interested in.

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

[31] See note #8. In Study 2, the counterstereotype training paired black and white faces respectively with positive, stereotypically white traits ("intelligent," "wealthy") and negative, stereotypically black traits ("poor," "lazy"), whereas the implicit prejudice measure tested associations between the masked words "black" or "white" and generic positive or negative words (such as "paradise" or "rotten"). The stimuli of face-with-stereotypical-trait pairings during training are quite different from the subliminal-race-word-with-generic-evaluative-word pairings during testing.

[32] Kawakami's studies on "correction" effects may also bear on the context-specificity worry. In the context of evaluating job candidates, participants initially responded in more stereotypical ways and then, after satisfying the goal of correcting for the perceived influence, they responded in counterstereotypical ways (Kawakami, Dovidio, and van Kamp 2007). This pattern may suggest that participants' new "default" is substantially more counterstereotypical, and that they had to *exert effort* to continue to be stereotypical. This is also reflected in another study by Kawakami, Dovidio, and van Kamp (2005), which found that participants under cognitive load made less stereotypical *initial* post-training judgments. Insofar as cognitive load disrupts controlled rather than automatic processing, this finding suggests that counter-stereotyping wasis their new automatic response. This is the opposite pattern from what one would expect if the effects were problematically context-specific. Normally, the findings of context-specificity are that people seem to be debiased immediately after the intervention (perhaps before the transient priming effects wear off), just so long as they are highly motivated and in the same lab, etc. For example, Smith and de Houwer (2014) found that that a persuasive message (of the kind emphasized by Mandelbaum, 2014) influenced implicit attitudes on one measure administered immediately after participants read the message, but not on a second measure, which was administered immediately after the first measure. One possibility is that the effects of the persuasive message on implicit attitudes were particularly fragile and short-lived, i.e., extremely context-specific. By contrast, in Kawakami's studies, participants respond with *more* bias at first, perhaps reacting against the training, but then demonstrate reduced bias on subsequent measures.

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 on implicit measures, 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 different 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 also noteworthy that these forms of training involve pretty robust (if rote) *actions* on the part of the participants. They are not just passively taking in information (as if watching TV[33]), but engaging in embodied performances of counterstereotypical and approach behaviors. This contrasts with, say, Dasgupta and Greenwald's (2001) paradigm of exposing participants to images of admired black individuals and infamous white individuals. In that study, which found significant reductions in implicit bias but has not been replicated with similarly strong effects (Joy-Gaba and Nosek 2012), participants had to choose which of two descriptions accurately applied to the person represented. The example they offer is Martin Luther King Jr. paired with the descriptions "Leader of the Black Civil Rights movement in the 1960s" and "Former Vice President of the United States." Choosing the correct option here might help to *remind* participants of the counterstereotypical nature of the individual in question, but it is not as if they

[33] Or merely watching a screensaver with counterstereotypical exemplars. Mazarin Banaji made a photo screensaver that would cycle through counterstereotypical exemplars. A file full of such images—e.g., of prominent women in the military—is available for download from the website for National Center for State Courts (http://www.ncsc.org/ibeducation).

are actually endorsing or affirming the counterstereotype. This sort of intervention is, plausibly, just making certain positive *subtypes* of the categories more accessible, without actually changing participants' attitudes about these categories.[34] For that, more direct actions that actually challenge those attitudes might be necessary, and they might have to be repeated a few hundred times.

My final response to the context-specificity worry is more nuanced, and I develop it in greater length elsewhere.[35] We should not, I argue, aim for the total erasure of "stereotypical associations" from our minds. There are many contexts where stereotypes *ought* to spring immediately to mind: in particular, we need to be able to automatically detect when people are being treated in stereotypical ways and 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, especially when other people are using them in an objectionable way. In this vein, 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 (§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.

[34] I interpret similarly Blair, Ma, and Lenton's (2001) study on imagining a counterstereotypical exemplar for 5 minutes. The possibility that these studies primarily work by briefly enhancing subtype accessibility relates to Han et al.'s (2010) contention that many interventions that induce immediate changes on the IAT might not lead to actual changes in implicit biases. I would bet, however, that "many, many repetitions" of these interventions would do so.
[35] Madva, "Virtue, Social Knowledge, and Implicit Bias" (forthcoming).

## 6. Practical unfeasibility

Suppose we grant, for the sake of argument, that these debiasing procedures lead to reasonably durable, context-general reductions in implicit 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 of the studies cited above include just 200.[36] 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. Of course, becoming *less* biased would presumably still be desirable, even if, for whatever reason, becoming completely unbiased, or becoming as unbiased as humanly possible, were unfeasible. In any case, working through these hundreds of trials can be done on any 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. 45 minutes is *nothing*.

[36] Gawronski et al. (2008), Johnson (2009), and Wennekers et al. (2012, 2013) reached significant effects with just 200 trials.

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 debiasing 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. Perhaps one way of seeing how impressed psychologists are by the magnitude of trials involved in these procedures is that the specific number is often reported differently. Johnson (2009, 8) puts the number in Kawakami's original 2000 studies at "a total of 160 trials," and Bargh (1999, 377) puts it at 240. It as if the actual number of trials doesn't matter. What matters is only that it's *really high*—it's over a hundred, it's hundreds, it's so many! Yet the 45 minutes it takes to get through 480 trials 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.[37] 45 minutes is

[37] Other biases may be involved in the search for "quick and easy fixes" besides extensive training. But we should be just as skeptical about "quick and easy fixes" for individuals trying to overcome their prejudices as we are already are about quick and easy fixes in other domains. The impulse and continuing search for quick fixes may be

less time than many people spend *per day* on exercise and the honing of other skills. American children spend an average of 4 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 implicit 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 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. This seems unfair. 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 interventions that best address systemic disadvantages for women in STEM fields will not overlap perfectly with those that best address systemic racial discrimination in the criminal justice system, nor with those that best address the systemic 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 employees in airport security might differ

---

part of the problem, a way to put off investing the work we know we need to do. (Of course, in a certain frame, 45 minutes of debiasing sounds like a quick-and-easy fix if ever there was one, so maybe this criticism can be raised against my advocacy of Kawakami's debiasing as well.)

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, the finding that, e.g., retraining racial *stereotypes* reduces racial *prejudice* suggests that some debiasing procedures might generalize in important respects (as I argued in §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 effectively train ourselves to automatically "avoid prejudice" and "approach egalitarianism" in general. Glaser and Knowles (2008) found that individuals who have implicit negative attitudes *toward prejudice per se* showed less racial bias on some measures. Another example might be adopting an "approach-oriented" mindset to social interactions. Trawalter and Shelton (2006) induced either approach-oriented or avoidance-oriented mindsets in participants before engaging in an interracial conversation:

> Specifically, participants in the prevention-focused condition were told, "It is important to the study that you avoid appearing prejudiced in any way during the interaction." By contrast, participants in the promotion-focused condition were told, "It is important to the study that you approach the interaction as an opportunity to have an enjoyable intercultural dialogue." (409)

Participants who adopted an approach mindset to the conversation were less cognitively depleted by the interaction than those who had adopted an avoidance mindset. Perhaps we should train ourselves to automatically approach diversity and dialogue and thereby "make interracial contact rewarding rather than depleting" (411). Perhaps we should practice approaching the voting booth and avoiding the status quo.[38] These are crucial empirical questions to pursue.

At this point, I can only speculate about additional factors that might drive an aversion to

[38] See "A Plea for Anti-Anti-Individualism" (in prep) for further discussion of precisely which sorts of attitudes we should aim to retrain.

debiasing. In discussions with colleagues, students, or acquaintances, it sometimes seems as though people just have a *kneejerk* negative response to the very idea, and thereafter confabulate reasons that justify their aversion (a la Haidt 2001). I suspect that a number of factors contribute to making the whole business seem *creepy*. It sounds like "thought police" and brainwashing. Talking seriously about counterconditioning inevitably calls up images of *The Manchurian Candidate* and *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 and values 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*.[39]

[39] In conversation, Michael Brownstein and Manuel Vargas suggested that there might be some additional factors

Here I hope to bracket the theoretical debate whether implicit prejudices and stereotypes are, at bottom, a matter of beliefs, as opposed to a matter of habits, mere associations, or aliefs. (If implicit biases are beliefs, and debiasing changes implicit biases, then, yes, debiasing changes beliefs, but not obviously in a troubling way.) Nevertheless, in many cases, there may be some further senses in which debiasing requires or leads to changes in what we believe and value. For example, Devine and colleagues (2012) found that their anti-prejudice intervention tended to increase participants' concern about discrimination, and this increase in concern moderated reductions in implicit bias, echoing Johnson's (2009) finding that debiasing procedures were most effective among participants who reported being strongly motivated to be unprejudiced. Perhaps debiasing procedures will increase the extent to which we view discrimination as a problem, or perhaps getting ourselves to care more about discrimination will heighten the effects of debiasing. If so, then changes in explicit beliefs, motivations, and ethical commitments may be involved, but I take it that many of us would not perceive such changes as alien implantations. I believe I *ought* to be concerned about discrimination, and perhaps I ought to be more deeply concerned than I am presently. It may be that my explicit concerns about discrimination are not as strong and salient as they should be.

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. Part of the opposition to debiasing might be an *attachment* to the explicit attitudes that implicit biases support, e.g., a person's self-esteem and

---

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 strike me as highly plausible contributing factors.

conviction that the world is fair might be causally supported in part by negative implicit attitudes toward members of low-status social groups. Perhaps an effective intervention that reduces this person's biases will also make him chronically depressed, low in self-esteem, or angry about global injustice. This is an empirical question like any other, which should be explored, but I suspect that the benefits of reducing widespread discrimination and oppression 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 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 debiasing procedures. Finally, 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.

In any event, 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. Bryan Gibson's (2008) article in *The Journal of Consumer Research* reported that an unobtrusive conditioning procedure changed implicit preferences for such "mature brands" as Coke versus Pepsi (but only for participants who did not already have a strong preference). Gibson proposed that these findings should contribute to further inquiry into ideal strategies for *product placement*. 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 debiasing procedure that, using different stimuli, was also found to enhance implicit self-esteem. 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![40]

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 and values 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.[41]

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

[40] Thanks to Katie Gasdaglis for helping me appreciate this point.

[41] Another source of perceived creepiness (similar to Brownstein and Vargas' suggestions two footnotes earlier) might be that these training procedures often involve using photos of real black and white men: perhaps this feels like *using* people as mere means to help make ourselves less biased rather than treating these individuals as ends in themselves. Of course, much the same could be said of most of the other interventions on offer, e.g., reflecting on infamous white individuals to help drive down an implicit preference for whites. If this were the real source of the worry, there would seem to be straightforward ways around it—just use lifelike computer-generated images of faces in debiasing procedures rather than images of real people. Maybe these strategies would still be objectionable insofar as they involve "using" racial whiteness and blackness as means to reduce our prejudices. If we are to take this concern seriously, however, then the "real life" applications of these ideas are far more troubling than the lab-based versions. Bringing whites into social contact with blacks *for the sake of* removing their prejudices seems to be a much clearer case of using actual people as means to achieve some further end, unlike the lab-based training, which need not *actually* involve interacting with other people in potentially objectionable ways.

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-institutional 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.

My consideration of how social and cognitive biases might contribute to skepticism about debiasing 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 and global poverty and hunger. A commonly cited bias is a sense of "distance" that we feel toward people and problems that are physically far away or very different from us in some salient respect. For example, Robert Gifford (2011) explains that people around the globe tend to discount environmental risks perceived to be in the distant future, and even tend to think that environmental damage is worse in other, distant countries than it is in one's own country. Allen (2004, 101) argues that "our

best-taught habit of citizenship is 'don't talk to strangers.'" We tend to avoid, ignore, and maintain distance from those we don't already know and those we perceive to be different from ourselves, with harmful consequences for intergroup communication and democratic cooperation. Similarly, Michelle Alexander, in describing the current system of mass incarceration in the US, writes:

> Race plays a major role—indeed, a defining role—in the current system, but not because of what is commonly understood as old-fashioned, hostile bigotry. This system of control depends far more on *racial indifference* (defined as a lack of compassion and caring about race and racial groups) than racial hostility… (2010, 198, original emphasis)

It is plausible that pervasive biases toward indifference, ignorance, and avoidance of the people and problems that are geographically, socially, or temporally "distant" 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.

Another source of kneejerk pessimism might have to do with how *stupid* or *brainless* these interventions seem. "Indeed," write Forbes and Schmader about their counterstereotype training (2010, 13), "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." There is a kind of fantasy that the hard 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.)

I believe there is another important concern at play here, roughly to do with intersectionality, that just approaching blacks and avoiding whites with a joystick is problematically over-simple 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) and Sullivan (2006) 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 (see my (2012) for more on the analogy). Memorizing vocabulary and grammar rules is not the same as becoming fluent in a second language. But we do need to memorize vocabulary and learn a bunch of rules before becoming truly fluent. Kawakami's simplistic debiasing procedures may be the anti-prejudicial equivalent of memorizing flashcards. These are just the *basics*, which will put us in a better position to actually *act* in unbiased ways in the "real world," with all its

inherent complexity.

## 7. Individualistic versus social-institutional approaches to discrimination

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 counterconditioning individuals' implicit biases (Alcoff 2010; Anderson 2010, 2012; Banks and Ford 2008; Dixon et al. 2012; Haslanger 2012, 2013; Huebner forthcoming). We should, they argue, instead focus on setting up institutional structures that preclude the operation of implicit biases in advance (such as anonymous reviewing) or counteract their operation after the fact (such as structures of affirmative action). I wholeheartedly support many of these structural interventions, and I take up concerns about the putatively individualistic focus of debiasing research in greater depth in a companion paper.[42] Far from being in *competition*, however, debiasing may be integral to the successful implementation of broader systematic changes.

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. In *Parents*

[42] I am also sympathetic with the criticism that philosophers have been especially prone to lose sight of the social and structural forest in the individualist trees (Haslanger 2013). Legal theorists, for example, are way ahead of philosophers on considering the political-institutional context of implicit bias. See Lawrence (1987) and the collection of papers edited by Levinson and Smith (2012) (none of which, however, mentions Kawakami's research, although Dasgupta and Greenwald (2001) and Blair's (2002) review are cited in passing).

*Involved v. Seattle* (2007), the Supreme Court ruled that districts could not classify students by race in order to integrate schools; in *Schuette v. Coalition* (2013), the Court upheld Michigan's ban on affirmative action in state university admissions; in *Wal-Mart v. Dukes* (2011), a 5-4 majority could not have cared less about the claim that Wal-Mart managers were allowed too much discretion in hiring and promoting, which allegedly allowed for implicit bias to distort their decision-making. I frankly fail to see how, in the contemporary political climate, institutional interventions for addressing bias have cornered the market on brass-tacks pragmatism. Apart from asking how effective debiasing will be, then, we might ask *how much opposition will there be*? We can make counterstereotype or approach training widely available to individuals without overhauling institutional structures in potentially contentious ways. While we can (and, I believe, should) weave these forms of debiasing into our institutions, we need not. Debiasing strategies 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 debiasing 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 these "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 (Alcoff 2010; Anderson 2010; Huebner forthcoming; Jolls and Sunstein 2006; Kang and Banaji 2006).[43] Matters are likely not so simple. 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 (2012) 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 could easily backfire, begetting heightened prejudice and discrimination. Fortunately, we do not have to look far for psychological interventions that could *mutually reinforce* institutional change. Debiasing procedures could provide the necessary *psychological* scaffolding to implement antidiscrimination initiatives without amplifying hostility; at the same time, initiatives like affirmative action might provide the necessary environmental scaffolding to reinforce the effects of debiasing 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, unfair and misguided. They are unfair because institutional sponsorship of debiasing need not take the (potentially) objectionable form of a universal debiasing mandate.

[43] Anyone moved by Kantian objections to debiasing, i.e., that retraining itself involves using others or ourselves as mere means (see notes #38 and #40), should be pretty alarmed by this proposal.

There are myriad "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 freely opt out. These worries are also 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. 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.

## 8. Appendix: "real-world" applications and "translations" of debiasing strategies

What are some examples of real-world applications of counterstereotype or approach training? Even if you yourself don't undergo counterstereotype training, Carr, Dweck, and Pauker (2012) found that simply *believing* that prejudice is malleable rather than fixed can make individuals' behavior significantly less biased. Mallett, Wilson, and Gilbert (2008, 271) found that focusing on similarities with outgroup members can improve social interactions, even if the similarities are as humdrum as shared preferences for "apples versus oranges and carpets versus hardwood floors." The debiasing strategies that Devine and colleagues (2012) taught their participants are all excellent examples of how we might "translate" counterstereotype and approach training into the real world: (1) stereotype replacement (noticing and replacing a stereotypical response with a counterstereotypical one), (2) imagining a counterstereotypical exemplar, (3) focusing on "individuating" rather than "group-based" features of others, (4) taking the perspective of a stereotyped group member, and (5) increasing opportunities for positive social contact.

Another example seems to be adopting an "approach-oriented" mindset to one's interactions (Trawalter and Shelton 2006). Then again, maybe sometimes we should *approach egalitarianism* while at other times we should *avoid prejudice*. Phills and colleagues (2011) found that taking approach-equality strategies to positive images (e.g., photos of smiling, racially diverse groups of people) reduced implicit racial bias, as did taking avoid-prejudice strategies to negative images (e.g., of Ku Klux Klan members burning crosses). But it is far less effective to say "Yes" to equality while faced with images of the Klan and to say "No" to prejudice while faced with positive images. The idea is that, "under certain conditions, both approach and avoidance motivations can successfully decrease implicit prejudice" (972). There are, then, myriad ways in which we can take these debiasing lessons to heart, and apply them broadly in our daily lives. But I still don't see why we shouldn't also *just do the training*.[44]

**References**

Alcoff, L. M. (2010). Epistemic identities. *Episteme, 7*(02), 128-137.

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

Allen, D. S. (2004). *Talking to strangers: Anxieties of citizenship since Brown v. Board of Education*. University of Chicago Press.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of personality and social psychology*, *91*(4), 652.

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

Anderson, E. (2012). Epistemic justice as a virtue of social institutions. *Social Epistemology*, *26*(2), 163-173.

[44] Acknowledgments.

Banks, R. R., & Ford, R. T. (2008). (How) Does Unconscious Bias Matter: law, Politics, and Racial Inequality. *Emory LJ*, *58*, 1053.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), Dual-process theories in social psychology (pp.361-382). New York: Guilford Press.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of personality and social psychology*, *81*(5), 828-841.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*(3), 242-261.

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological psychiatry*, *52*(10), 976-986.

Calanchini, J., Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. *European Journal of Social Psychology*, *43*(5), 321-325.

Carpenter, S. May 1st 2008: Buried Prejudice. *Scientific American Mind*, 32-39.

Carr, P. B., Dweck, C. S., & Pauker, K. (2012). "Prejudiced" behavior without prejudice? Beliefs about the malleability of prejudice affect interracial interactions.

Dasgupta, N., and Asgari, S. 2004: Seeing is believing: Exposure to counterstereotypic women leaders and its effect on automatic gender stereotyping. *Journal of Experimental Social Psychology* 40, 642-658.

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

Dasgupta, N., & Rivera, L. M. (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition*, *26*(1), 112-123.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*.

Dixon, J., Levine, M., Reicher, S., & 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.

Drake, B. (April 22, 2014). Public strongly backs affirmative action programs on campus. *Pew Research Center*. URL = < http://pewrsr.ch/1gPklxV >

Dunham, Y. December, 2011: The development of implicit bias. Presentation for the *Implicit Bias & Philosophy Workshop: The Nature of Implicit Bias*. University of Sheffield, UK.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences*, *12*(7), 248-253.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316-326.

Eberl, C., Wiers, R. W., Pawelczack, S., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2013). Approach bias modification in alcohol dependence: Do clinical effects replicate and for whom does it work best?. *Developmental cognitive neuroscience*, *4*, 38-51.

Forbes, C. E., & Schmader, T. (2010). Retraining attitudes and stereotypes to affect motivation and cognitive capacity under stereotype threat. *Journal of personality and social psychology*, *99*(5), 740.

Gawronski, B., & Cesario, J. (2013). Of Mice and Men What Animal Research Can Tell Us About Context Effects on Automatic Responses in Humans.*Personality and Social Psychology Review*.

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

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research*, *35*(1), 178-188.

Gifford, R. (2011). The dragons of inaction: Psychological barriers that limit climate change mitigation and adaptation. *American Psychologist*, *66*(4), 290.

Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, *44*(1), 164-172.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology*, *38*, 69-119.

Haidt J (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psvchol Rev 108:814-834

Han, H.A., Czellar, S., Olson, M.A., Fazio, R.H. 2010: Malleability of attitudes or malleability of the IAT? *Journal of Experimental Social Psychology* 46, 286-298.

Haslanger, S. 2013: Social Meaning and Philosophical Method. Presidential Address at the Eastern Division Meeting of the American Philosophical Association.

Haslanger, S. 2012: *Resisting Reality: Social Construction and Social Critique*. Oxford.

Henry, P.J., and Hardin, C.D. 2006: The Contact Hypothesis Revisited: Status Bias in the Reduction of Implicit Prejudice in the United States and Lebanon. *Psychological Science* 17, 862-68.

Holroyd, J. & Sweetman, J. Forthcoming. The Heterogeneity of Implicit Biases. Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume I, Metaphysics and Epistemology*. Oxford: Oxford University Press.

Huebner, B. Forthcoming. Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. *Implicit Bias and Philosophy: Volume I: Metaphysics and Epistemology*. Oxford University Press.

Johnson, I. R. (2009). *Just say "No"(and mean it): Meaningful negation as a tool to modify automatic racial prejudice* (Doctoral dissertation, Ohio State University).

Jolls, Christine, and Cass R. Sunstein. "The law of implicit bias." *California Law Review* (2006): 969-996.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology*, *41*(3), 137-146.

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. (2012). Presumed fair: Ironic effects of organizational diversity structures.

Kang, J., & Banaji, M. R. (2006). Fair Measures: A Behavioral Realist Revision of" Affirmative Action". *California Law Review*, *94*(4), 1063-1118.

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology*, *41*(1), 68-75.

Kawakami, K., Dovidio, J.F., and van Kamp, S. 2007: The Impact of Counterstereotypic Training and Related Correction Processes on the Application of Stereotypes. *Group Processes and Intergroup Relations* 10 (2), 139-156.

Kawakami, K., Dovidio, J.F., Moll, J., Hermsen, S. and Russin, A. 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 , 871–888 .

Kawakami, K., Phills, C.E., Steele, J.R., and Dovidio, J.F. 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.

Kawakami, K., Steele, J. R., Cifa, C., Phills, C. E., & Dovidio, J. F. (2008). Approaching math increases math= me and math= pleasant. *Journal of Experimental Social Psychology*, *44*(3), 818-825.

Kelly, D., Faucher, L., and Machery, E. 2010: Getting Rid of Racism: Assessing Three Proposals in Light of Psychological Evidence. *Journal of Social Philosophy* 41 (3), 293-322.

Kunda, Z., and Spencer, S.J. 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.

Lawrence III, C. R. (1987). The id, the ego, and equal protection: Reckoning with unconscious racism. *Stanford Law Review*, 317-388.

Leslie, S.J. 2009: The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*.

Levinson, J.D. & Smith, R.J. (2012) *Implicit Racial Bias Across the Law*, Cambridge.

Madva, A.M. (Forthcoming). Virtue, Social Knowledge, and Implicit Bias. *Implicit Bias & Philosophy*, eds. Jennifer Saul & Michael Brownstein. Oxford.

Madva, A.M. (In prep). A Plea for Anti-Anti-Individualism: How Oversimple Psychology Misleads Social Policy.

Madva, A.M. (2012). *The Hidden Mechanisms of Prejudice: Implicit Bias & Interpersonal Fluency*, PhD Dissertation, Columbia University.

Mallett, R. K., Wilson, T. D., & Gilbert, D. T. (2008). Expect the unexpected: Failure to anticipate similarities leads to an intergroup forecasting error. *Journal of personality and social psychology*, *94*(2), 265.

Mandelbaum, E. 2014: Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Nous*.

Martin, D. March 30, 2013: Yvonne Brill, a Pioneering Rocket Scientist, Dies at 88. *The New York Times*. URL = < http://www.nytimes.com/2013/03/31/science/space/yvonne-brill-rocket-scientist-dies-at-88.html >

Moskowitz, G.B. 2010: On the Control Over Stereotype Activation and Stereotype Inhibition. *Social and Personality Psychology Compass* 4 (2), 140-158.

Olson, K. R., & Dunham, Y. (2010). The development of implicit social cognition. *Handbook of*

*implicit social cognition: Measurement, theory, and applications*, 241-254.

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

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual review of psychology,60*, 339-367.

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

Park, S. H., Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice moderates the effect of cognitive depletion on unintended discrimination. *Social Cognition*, *26*(4), 401-419.

Pearson, A.Dovidio, J.F., Gaertner, S.L. 2009: The nature of contemporary prejudice: insights from aversive racism. *Social and Personality Psychology Compass* 3, 1-25.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, *90*(5), 751.

Phills, C. E., Kawakami, K., Tabi, E., Nadolny, D., & Inzlicht, M. (2011). Mind the gap: Increasing associations between the self and Blacks with approach behaviors. *Journal of Personality and Social Psychology*, *100*(2), 197-210.

Phills, C. E., Santelli, A. G., Kawakami, K., Struthers, C. W., & Higgins, E. T. (2011). Reducing implicit prejudice: Matching approach/avoidance strategies to contextual valence and regulatory focus. *Journal of Experimental Social Psychology*, *47*(5), 968-973.

Plant, E. A., Peruche, B. M., & Butz, D. A. (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental Social Psychology*, *41*(2), 141-156.

Putnam, R.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.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). " Unlearning" Automatic Biases: The Malleability of Implicit Prejudice and Stereotypes. *Journal of personality and social psychology*, *81*(5), 856-868.

Rudman, L. A., & Lee, M. R. (2002). Implicit and explicit consequences of exposure to violent and misogynous rap music. *Group Processes & Intergroup Relations*, *5*(2), 133-150.

Schneider, D.J. 2004: *The Psychology of Stereotyping*. New York: Guilford Press.

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

Smith, C. T., & De Houwer, J. (2014). The impact of persuasive messages on IAT performance is moderated by source attractiveness and likeability. *Social Psychology*, 45(6):437–448.

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

Stout, J. G., Dasgupta, N., Hunsinger, M., & McManus, M. A. (2011). STEMing the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of personality and social psychology*, *100*(2), 255.

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

Sullivan, S. 2006: *Revealing Whiteness: The Unconscious Habits of Racial Privilege*.

Uhlmann, E. L., Brescoll, V. L., & Machery, E. (2010). The motives underlying stereotype-based discrimination against members of stigmatized groups. *Social Justice Research*, *23*(1), 1-16.

Valian, V. 1998: *Why So Slow? The Advancement of Women*. MIT Press.

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

Weisbuch, M., Pauker, K., & Ambady, N. (2009). The subtle transmission of race bias via televised nonverbal behavior. *Science*, *326*(5960), 1711-1714.

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

Wennekers, A. M., Holland, R. W., Wigboldus, D. H., & van Knippenberg, A. (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.

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological Science*, *22*(4), 490-497.

Ziv, T., & Banaji, M. R. (2012). Representations of Social Groups in the Early Years of Life. *The SAGE Handbook of Social Cognition*, 372.