Alex Madva September 2013

Biased Against De-Biasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle Against Prejudice[1]

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 and discrimination. 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.

I. 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.[2] These training procedures don't just influence scores on indirect measures like the

[1] Thanks to Michael Brownstein and Katie Gasdaglis for extensive comments on earlier drafts, to everyone at the Implicit Bias, Philosophy, and Psychology workshop at the University of Sheffield in April 2013, and to the audience at the Rocky Mountain Ethics Congress at CU-Boulder in August 2013, where I also benefited greatly from Daniel Silvermint's excellent comments.

[2] 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), and Wennekers (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

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 about job candidates (making participants less likely to 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 psychologists and activists as lacking direct, practical import for the broader struggle against prejudice and discrimination.[3] 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."[4]

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."[5]

(PRACTICAL UNFEASIBILITY) Many allege (typically in passing) that, even if these debiasing procedures prove to be effective, they would still be too laborious and time-consuming

---

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).

[3] See, e.g., Bargh (1999 , 377), Stewart and Payne (2008), Mendoza et al. (2010), and an interview of Keith Payne in Carpenter (2008).

[4] Schneider (2004, 423).

[5] For discussion, see Olson and Fazio (2006, 431-2), Devine (2012, 1277-8), Gawronski and Cesario (2013), Mandelbaum (manuscript), and Mendoza et al. (2010, 521).

to be practically feasible.[6]

(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.[7]

Here I reply to these criticisms of debiasing. Ultimately, a comprehensive strategy for combating prejudice and discrimination should include a central role for training our biases away. First, I survey the relevant research. I go into some depth because the details are important for answering concerns about the efficacy and feasibility of debiasing.

II. 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 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."[8]

[6] See the same authors cited in notes 3, 4, and 5.

[7] See Alcoff (2010), Anderson (2012), and Dixon et al. (2012) on the limits of individualist approaches to bias. One might reasonably describe these concerns as broadly "Marxist" or "Foucauldian;" the idea is that we're wasting our time unless we're talking about directly changing the underlying material conditions or radically restructuring power relations. This was also the thrust of Haslanger's (July 2012) lecture, where she points out that the turn away from individualism has been an integral part of feminist and anti-racist theorizing in recent decades.

[8] A commentator once 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 "NO" and "YES" on them. In a third study, they just pressed the M and Z buttons on a computer keyboard. Participants can, then, 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? Johnson (2009) raises this concern, as I discuss below. See also my (2012, ch.1).

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 pre-test and a post-test 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.[9] The researchers also found that negation training eliminated the automatic activation of skinhead stereotypes.[10] 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).[11] Kawakami and colleagues wrote, "In short, practice does make perfect—or at least very good—stereotype negators" (884).

Two follow-up studies outside of Kawakami's lab have partially replicated but partially qualified the original findings. First, Bertram Gawronski and colleagues (2008) observed that the original studies confounded two sorts of training—the repeated *affirmation* of

[9] There were no significant differences between pre- and post-tests 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 an unusual sort of 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 post-test.

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

[11] 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 pre-test. There was thus no bias to eliminate for those participants about that social group. While negation training completely eliminated skinhead stereotype activation, it's 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.

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.

Fig. 2. Mean scores of automatic preference for Whites over Blacks as a function of training task (affirmation of counterstereotypes vs. negation of stereotypes) and time of measurement (before vs. after training), Experiment 2.

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 had significant effects both on

automatic *gender stereotyping* and on automatic *racial prejudice*.[12] For those who make a lot of hay out of the distinction between semantic, belief-like associations (aka stereotypes) and emotional, motivational associations (aka prejudices), it is noteworthy that researchers found significant effects of training on *both* sorts of implicit association—and, moreover, that retraining automatic *racial stereotyping* led to changes in automatic *racial prejudice*. That is, retraining putatively "cognitive" stereotypical associations led to changes in automatic *affective* responses.[13]

Gawronski and colleagues ultimately 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 (2009) 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

[12] 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 a little too similar for my tastes in this particular study. The training in Study 2 (results are visible in the above figure) 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.

[13] This is, in other words, evidence that these two types of implicit "association" might not be so radically distinct as some (e.g., Amodio and Devine, 2006) have argued. 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). Michael Brownstein and I (in preparation) argue 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 at the level of implicit cognition.

some pizzazz in the space-bar pressing by instructing participants that the action was equivalent to saying "*NO! THAT'S WRONG*!"[14] Johnson found that 200 trials of this "meaningful negation" of stereotypes did in fact reduce automatic prejudice.[15] Interestingly, she also found that meaningful negation was most effective for individuals strongly motivated to be unprejudiced.

In addition to reducing automatic stereotyping and prejudice on indirect computerized measures, evidence also suggests that these debiasing procedures can influence "real world" social behaviors.

In Kawakami, Dovidio, and van Kamp (2007), participants first underwent gender counterstereotype training, by pairing male faces with words like "sensitive" and female faces with words like "strong."[16] 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

[14] 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., 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. It also calls to mind Glaser and colleagues' (2008) work on the "implicit motivation to control prejudice," whereby individuals who demonstrate strong automatic negative attitudes to words like "prejudice" on the IAT show, e.g., less Shooter Bias.

[15] 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 *exacerbated* 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* racial bias. Most adults' automatic dispositions toward stereotyping and prejudice are 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 studies wherein some participants repeatedly practice the stereotypical or prejudicial responses. There's often just a non-significant or marginally significant trend toward exacerbation in those conditions.

[16] 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.

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 male name and the other half saw that same résumé with a female 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. 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:

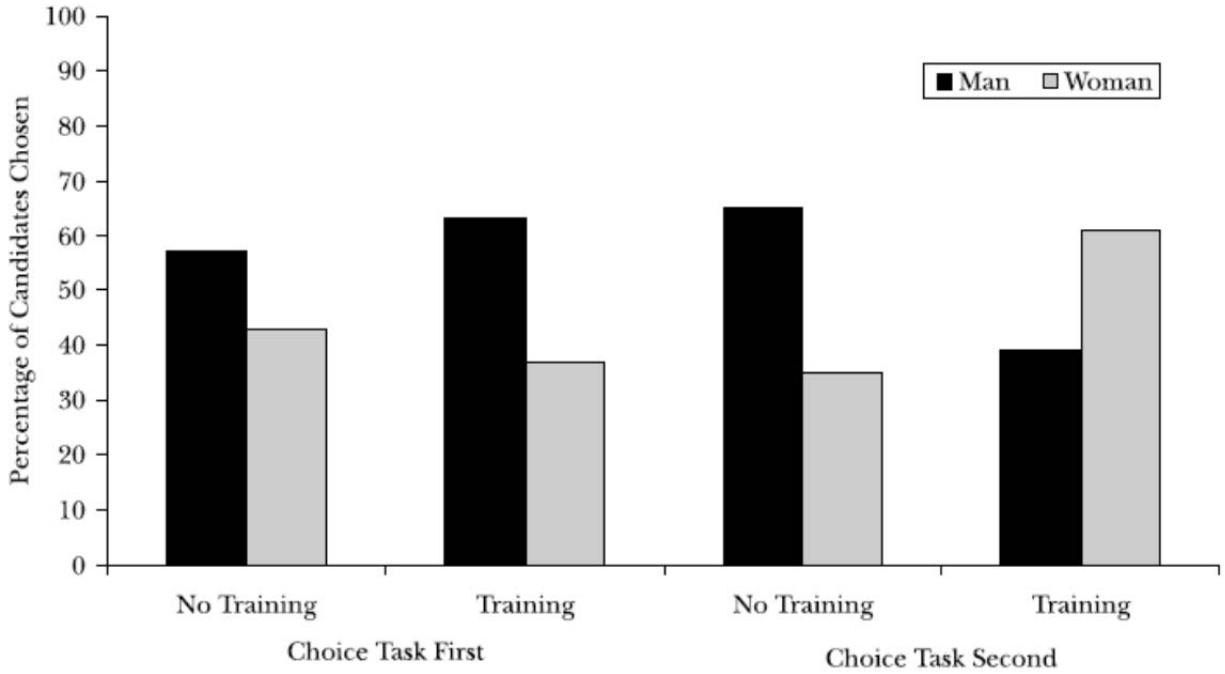

Figure 1. Effects of counterstereotypic training and order of tasks on choice of male and female candidates.

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.

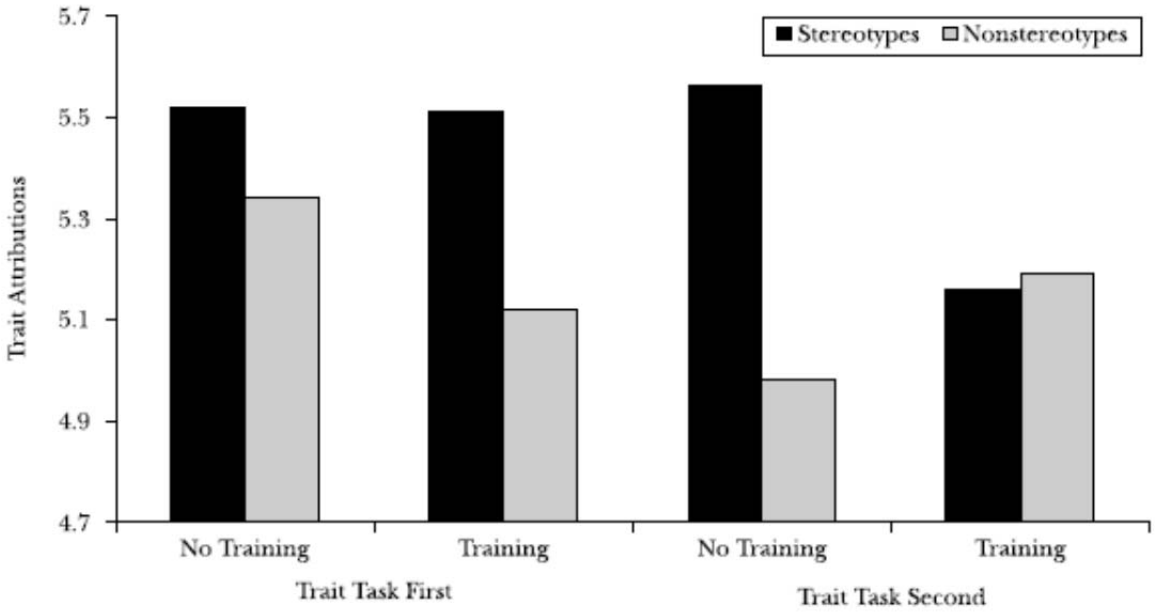

Figure 2. Effects of counterstereotypic training and order of tasks on attributions of stereotypic and nonstereotypic traits to male and female candidates.

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).[17]

Using an altogether different procedure, Kawakami, Phills, Steele, and Dovidio (2007)

[17] The implication seems to be that, even though, say, a white male employer might express resentment or discomfort *during* or *immediately after* an information session or other intervention aimed at reducing discrimination in the workplace, the effects of that intervention might show up later on. Given that Johnson (2009) found differential effects for counterstereotype training depending on participants' explicit concerns about prejudice, an important follow-up study 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, etc.

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 bias on the IAT. [18] 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 all conditions, regardless whether the meaning of the training was fully explicit or subliminal.[19] 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 will strike some as a decidedly creepier variant of an already spooky approach to prejudice reduction).

[18] It bears mentioning that the IAT is a different measure of implicit prejudice than the affective priming measure used in Gawronski et al. (2008) and Johnson (2009): we've got *more and more measures* demonstrating significant effects of training. In a control condition in Study 1, participants just moved the joystick left or right (*D* was 0.43); in another condition, participants approached whites and avoided blacks (*D* was 0.52). In the approach-black/avoid-white condition, *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).

[19] There were no significant effects of training for participants who repeatedly moved the joystick left or right, or who repeatedly avoided black faces and approached white ones.

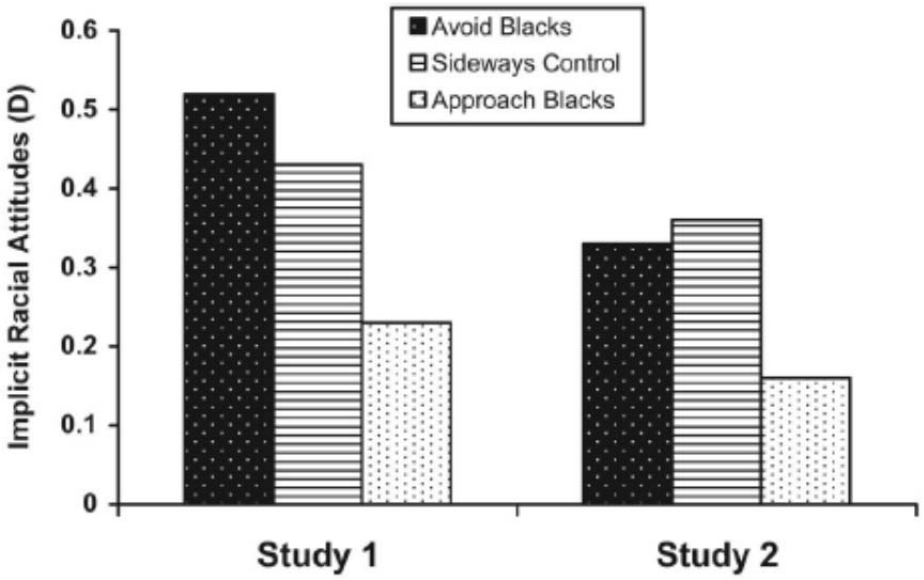

Figure 1. Effects of supraliminal (Study 1) and subliminal (Study 2) training procedures on implicit racial attitudes.

Annemarie Wennekers and colleagues (2012, 2013) replicated these effects by having participants nod versus shake their heads in response to typical Moroccan and 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).[20]

Moreover, Kawakami and colleagues found that subliminal approach training influenced *actual social behavior*, leading 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.

[20] Wennekers (2013) also found that nodding in response to only 50% of the Moroccan stimuli, instead of 100%, failed to have significant effects. 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 (see section III). We cannot expect to approach or have positive social interactions with every member of a particular social group, stigmatized or otherwise. Wennekers and colleagues also found that the nodding has to come *after* the stimulus (as if nodding in *response* to something), so spatiotemporal ordering seems important.

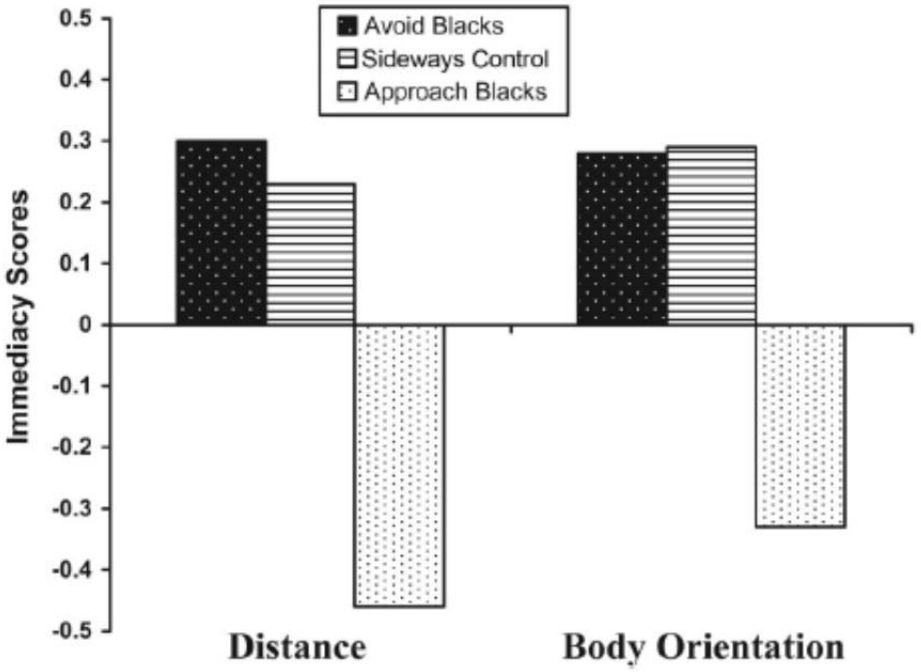

Figure 3. Effects of a subliminal training procedure in Study 4 on distance from and body orientation toward a Black confederate.

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 female undergraduates of repeatedly approaching math-related images ("e.g., calculators, equations").[21] 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 implicit measures, as well as to answer more questions on a math test. A series of follow-up studies by Forbes and Schmader (2010) replicated these findings using a different training procedure, and with a 24-30 hour delay between the debiasing procedure and the math test. 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

[21] While avoiding arts-related images, "e.g., guitars, poetry" (820).

increased working memory as well as improved performance on math questions from the GRE.[22]

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

III. 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 stereotyping, prejudice, and discrimination. 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? Policymakers already "spend billions of dollars annually on interventions aimed at prejudice reduction in schools, workplaces, neighborhoods, and regions beset by intergroup conflict."[23] Yet nobody, to my knowledge, has seriously advocated implementing these debiasing procedures in these contexts.

Instead, Kawakami's "laborious 480-trial procedure," which requires "many, many repetitions to learn nonstereotypical responses," is often cited as a point of contrast when

[22] 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.

[23] Paluck and Green (2009, 340).

researchers discover a less intensive, demanding intervention.[24] Many social psychologists 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 but 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."[25]

Nevertheless, many acknowledge that Kawakami's research might have *indirect* practical 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. Even Phills, Kawakami, and colleagues (2011) seem to assume that these debiasing procedures are not

[24] The first quote is from Olson and Fazio (2006, 431), the second from Stewart and Payne (2008, 1343), in an article reporting how 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 using implementation intentions in the struggle against discrimination, of course. 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, and we can encourage people to rehearse the relevant implementation intentions *after* they undergo the training, as icing on the cake to help their biases stay debiased after they leave the laboratory.

[25] 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, ch.3) 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 to control them. Payne thus takes it to be obvious that we should figure out ways to maximize local control. 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 pursue debiasing procedures at all is that immediate, local control seems out of reach.

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)[26]

I am all for implementing this sort of welcoming activity, but it is not in any 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 contact have a long history, and evidence for their success is mixed.[27] Henry and Hardin (2006) found that, while intergroup contact generally reduced *explicit* reports of prejudice, 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 remain unaffected. Even when contact reduces prejudice, the effect sizes tend to be relatively small, and the conditions conducive to effective social contact are difficult to construct and maintain. A rival "conflict" hypothesis seeks to explain how social contact often *amplifies* intergroup animosity. So I don't think we should not pin much hope on the prospects of getting white students to unlearn their implicit prejudices toward their black and Latino classmates

[26] Wennekers (2013, 85) also seems to think that her debiasing research should be translated: "repeatedly approaching out-group members and noticing that nothing bad happens may make you less likely to avoid them." And Schneider (2004, 423) writes, "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." These studies on retraining are taken to be relevant only insofar as they indicate how people might become less biased if they happen to be lucky enough to have lots of exposure to counterstereotypical exemplars in daily life.

[27] See Putnam (2007), Kelly, Faucher, and Machery (2010), Dixon et al. (2012), and my (2012, ch.4).

simply by having them shake hands in homeroom. Maybe if they 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 participants' associations between blacks and self-related words ("I," "me," "self," etc.).[28] In fact, increases in self-black associations seemed to *mediate* the reduction of implicit bias. Approach training evidently increased self-identification with the target group, 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.[29] Approach training is, in effect, *the contact hypothesis in a bottle*.[30]

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 to "real world" and

---

[28] "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 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).

[29] 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).

[30] This is not to say that approach training or its effects are *equivalent* to actual intergroup social contact or its effects. Like most "distillations" or "lab-designed replications" 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 some respects—and better in others. One salient difference is that we can guarantee that 100% of the trials are counterstereotypical in the lab, but not in the "real world." See note #20 on Wennekers et al. (2012, 2013).

imperfect translations of these procedures, we should be making these debiasing procedures widely available (e.g., on the Project Implicit website[31]), and considering ways in which institutions might incorporate debiasing into broader antidiscrimination strategies.

(I do, nevertheless, think that we should continue to explore effective "real-world applications" of these studies. I discuss a few examples in an appendix.)

IV. 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.[32] To my knowledge, nobody has tested how long people stay debiased after counterstereotype or approach training (Michael Brownstein, Brian Nosek, and I are collaborating with other psychologists and philosophers applying for funding to make these studies happen). 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.[33] I worry, however, that pessimism about the durability of debiasing is another contributing factor, in which case this pessimism becomes a self-fulfilling prophecy, where 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

[31] 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 debiasing IATs widely available.

[32] See, e.g., 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."

[33] 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.

anew) all of their biases. For example, 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."[34] Let's 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 empirical support.[35] One source of pessimism might be evidence from developmental

[34] 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. Mendoza et al. come dangerously close to inferring evidence of absence from absence of evidence, in that they cite the studies that showed effects lasting 24 hours as if they also *failed* to show effects lasting longer. But the relevant studies simply did not test for longer-term effects.

[35] One *theoretical* ground for pessimism has to do with the underlying nature of implicit biases. Eric Mandelbaum (manuscript) expresses skepticism about debiasing on the grounds that it reflects a misguided view of the psychological nature of implicit biases. He thinks interventions like Kawakami's reflect a misguided "associative" rather than "propositional" account of implicit biases. He refers to the Associative Interpretation of implicit Bias as AIB: "Much of the discussion is about how to extinguish implicit biases. If AIB really were the whole story, then we'd already know how deal with implicit racism: just put subjects in an extinction paradigm and poof, (at least temporary) implicit egalitarians we'd be. For what it's worth, I've heard very few people offer this solution. I suppose it's because deep-down most of us know that there is something wrong with AIB."

Mandelbaum thus interprets the widespread suspicion of direct debiasing strategies as evidence that we don't *really* believe that implicit biases are "mere associations," because, if we did, then we'd be actively trying to extinguish them. I agree with Mandelbaum about the maximally general claim that understanding the underlying nature of implicit attitudes is important for understanding what to do about them, but I intend to bracket these theoretical questions as much as possible in this paper. Mandelbaum's passage is puzzling for two reasons: first, because all the evidence I cited earlier suggests that we can become "at least temporary egalitarians" (the question here is whether these effects *last*—whether, e.g., media exposure outside the lab will condition us to *reacquire* the undesirable associations); second, because, although these sorts of debiasing procedures fit most "intuitively" with associative accounts, there is a long history of explaining conditioning in cognitive terms. The fall of behaviorism is attributed, in large part, to the fact that cognitive accounts could *better* explain the data—cognitive accounts, that is, of associative learning and extinction. I say more about the nature of implicit attitude change in (2012, ch.2).

psychology that implicit biases tend to form early in childhood and remain stable through adulthood.[36] 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.[37] Kawakami and others' research, in turn, undermines the commonsense view about the resilience of bias in adulthood, suggesting that individuals *can* unlearn 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

[36] Dunham et al. (2008). See Olson and Dunham (2010) and Ziv and Banaji (2012) for reviews.

[37] 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. Underprivileged groups tend to unlearn their implicit biases through social contact, while privileged groups do not. If these claims about perceptions of group status are vindicated, they might constitute an important example of *how* unjust social structures *per se* support implicit biases. It's not just that kids see too many stereotypes on TV; it's that they see real-world disparities between groups in social status. Score one for the revolutionaries who think we can't change implicit biases without overhauling social structures, although a less radical interpretation of this research is that it is emphasizing the importance of telling our kids and ourselves that these differences in group status are *wrong* and unfair and ought to be changed. This speaks to the goal-dependence of stereotyping that I will discuss shortly; people make these negative judgments about low-status groups *to make themselves feel better* in various ways.

be so worried about the additional commonsensical pronouncement that getting bombarded with stereotypes outside the lab will undo the effects of debiasing?[38]

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 music (but not pop), and after watching television clips in which white characters display subtle, nonverbal bias toward black characters.[39] 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

[38] It is common for psychologists and activists nowadays to speak about how much we've learned about prejudice and stereotyping over the past few decades, but I can't shake the sense that pessimism about the durability of debiasing is itself a holdover of the *old-fashioned* views that all this research is supposed to have debunked. 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.)

[39] See, respectively Rudman and Lee (2002) and Weisbuch, Pauker, and Ambady (2009). The rap music was by DMX, Dr. Dre, and Ice Cube, and the pop by Britney Spears and TLC. I interpret these studies on temporary increases in implicit bias as on a par with the studies on temporary decreases in bias that I discuss in section V.

consider.[40] It is just false that our biases depend primarily on the mere preponderance of "evidence" we take in, in the form of magazine covers, news stories, or what have you. Typically, belief perseverance, the confirmation bias, and a host of other cognitive dispositions help to create and sustain our 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 and how we interpret it is profoundly shaped by our implicit and explicit goals.[41] Aims that 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 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're motivated to stay that way, many of these psychological dispositions might now operate to maintain our *de*biases.

---

[40] Indeed, it's 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* or *facilitated* 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. See Leslie (forthcoming).

[41] 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 an array of evidence that stereotyping is driven by questionable aims (rather than by the aim *to be accurate*). These are good candidates for goals we should teach children *not* to have.

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 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.[42] This intervention led to reductions of bias that lasted at least 8 weeks.[43] Notably, participants' reported concerns about discrimination also increased, and this increased

[42] The strategies are excellent examples of how we might "translate" counterstereotype and approach training into the real world. See the appendix for further discussion.

[43] Participants implicit biases were even slightly lower after 8 weeks than after 4. 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 you had to debias yourself once or twice a year. Would an annual trip to the debiaser be too much to ask of ourselves? Suppose it was best to debias ourselves four times a year. Would that be too much? How much investment of time and effort is too much to ask? What if we can debias ourselves *subliminally*? Would it be a waste of time to debias ourselves once in a while even if we didn't re-up quite as often as recommended? How far from permanent does an intervention like this need to be in order to qualify as a counterproductive waste of time?

concern seemed to significantly enhance bias reduction. Evidence also suggests that counterstereotypical teachers can reduce their students' implicit biases.[44] Dasgupta and Asgari found that first-year female undergraduates who took multiple classes with female math and science professors showed less implicit gender bias after one year. 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 foregrounds her reputation as the world's best Mom and an expert at making beef stroganoff.[45] 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. Wiers et al. (2011) found that patients recovering from alcoholism who, immediately prior to undergoing standard treatment, were trained to avoid images of alcohol (in 4 sessions lasting 15 minutes each) were significantly less likely to relapse *one year* after being discharged.[46]

V. 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*.[47] Might the effects only be visible in this particular lab, or on that particular test?

---

[44] See Rudman et al. (2001), Dasgupta and Asgari (2004), and Stout et al. (2011).

[45] See Sullivan's (April 1, 2013) blog for discussion and links to Martin's (March 30, 2013) obituary of Yvonne Brill. URL = <http://publiceditor.blogs.nytimes.com/2013/04/01/gender-questions-arise-in-obituary-of-rocket-scientist-and-her-beef-stroganoff/>

[46] In comparison to patients who underwent no training or sham training prior to standard treatment. In a similar vein, Houben et al. (2011) found that practicing an inhibition or "stopping" response led to stronger implicit negative attitudes toward alcohol and decreased alcohol consumption during at least the following week.

[47] I say more about this issue in (2012, ch.2). The background to this worry lies in a series of findings that indirect

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. The worry might be, for example, that watching *The Cosby Show* won't necessarily "change the way you think" about black people in general, although it does change the way you think about well-to-do black fathers who wear baggy sweaters with colorful patterns. And it might change the way you think when in the context of watching a sitcom, but not in the context of actually entering a home. Researchers can test this 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.[48] Generally speaking, it looks like first impressions are incredibly important: people's initial salient exposure to a category member forms the backdrop for their future encounters with other category members. People can pick up quickly on the fact that novel category members don't 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

---

measures of bias are subject to striking context effects. For example, Barden et al. (2004) found that an image of a black man in a prison elicits negative responses if he is dressed like a prisoner, but positive responses if he is dressed like a lawyer.

[48] See Gawronski and Cesario (2013) for a recent review of this research. Thanks to Michael Brownstein for bringing this article to my attention.

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 see him against a blue background, or some novel color, their automatic response will reflect their initial positive impression. Maybe what we've 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 patterns in 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 many alternatives.)

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 automatic racial *stereotypes* led to changes in automatic racial *prejudice*, even though all the stimuli during training and testing were different (Gawronski et al. 2008). This training did not just influence people "in a context of potential stereotyping," but also "in a context of potential prejudice." 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" lead to reductions in bias, across an array of different measures

(e.g., Phills et al. 2011). Math-gender counterstereotype training improves 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 influences implicit measures but also reduces the likelihood of relapse into alcoholism for at least one year (Wiers et al. 2011).[49] 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 don't appear in Kawakami's debiasing paradigm until after participants have already worked through *80 trials*, and it takes a few hundred more trials before participants approach a ceiling past which they cannot improve. It takes a reasonable amount of 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 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.

[49] The studies that demonstrated "correction" effects, wherein participants try to correct for the perceived influence of the experiment, are also interesting to consider in relation to the context-specificity worry, because they demonstrated inconsistent behaviors *within the same context*. In the context of evaluating job candidates, participants initially responded in more stereotypical ways and *then*, after they satisfied their goal of correcting for the perceived influence, they responded in counterstereotypical ways (Kawakami, Dovidio, and van Kamp 2007). This pattern suggests 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; cognitive load disrupts controlled rather than automatic processing, so that suggests that counter-stereotyping is 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 (while they are still primed), and just so long as they are in the same lab, etc. In this case, participants acted *more* biased to compensate for the effects, and *thereafter* acted unbiased.

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[50]), 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 hasn't 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's not as if they're 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.[51] 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 (forthcoming). We should not, I argue, aim for the total erasure of

[50] 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).

[51] Ditto for Blair, Ma, and Lenton's (2001) study on imagining a counterstereotypical exemplar for 5 minutes. My concern that these studies primarily work by 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.

"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 stereotype activation (§IV above) 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.

VI. Practical unfeasibility

Critics of debiasing typically justify their skepticism, in part, by referring to the fact that the "laborious" procedure requires "many, many repetitions" to be effective, thereby implying that it is somehow unfeasible.[52] As if the sheer fact that it involves *hundreds of trials* is sufficient to establish that it's too labor-intensive to figure as a legitimate component of the larger struggle

[52] Olson and Fazio (2006) describe it as a "laborious 480-trial procedure" and Stewart and Payne (2008) emphasize that Kawakami's "extensive training," requires "many, many repetitions." One way of seeing how impressed psychologists are by the magnitude of trials involved 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!

against prejudice and discrimination.

How labor-intensive is it? Reliably significant effects start appearing after about 160 trials, and many of the studies cited above include just 200.[53] The benefits of additional training are still visible from 200 to 300 trials, but start to tail off around 400.

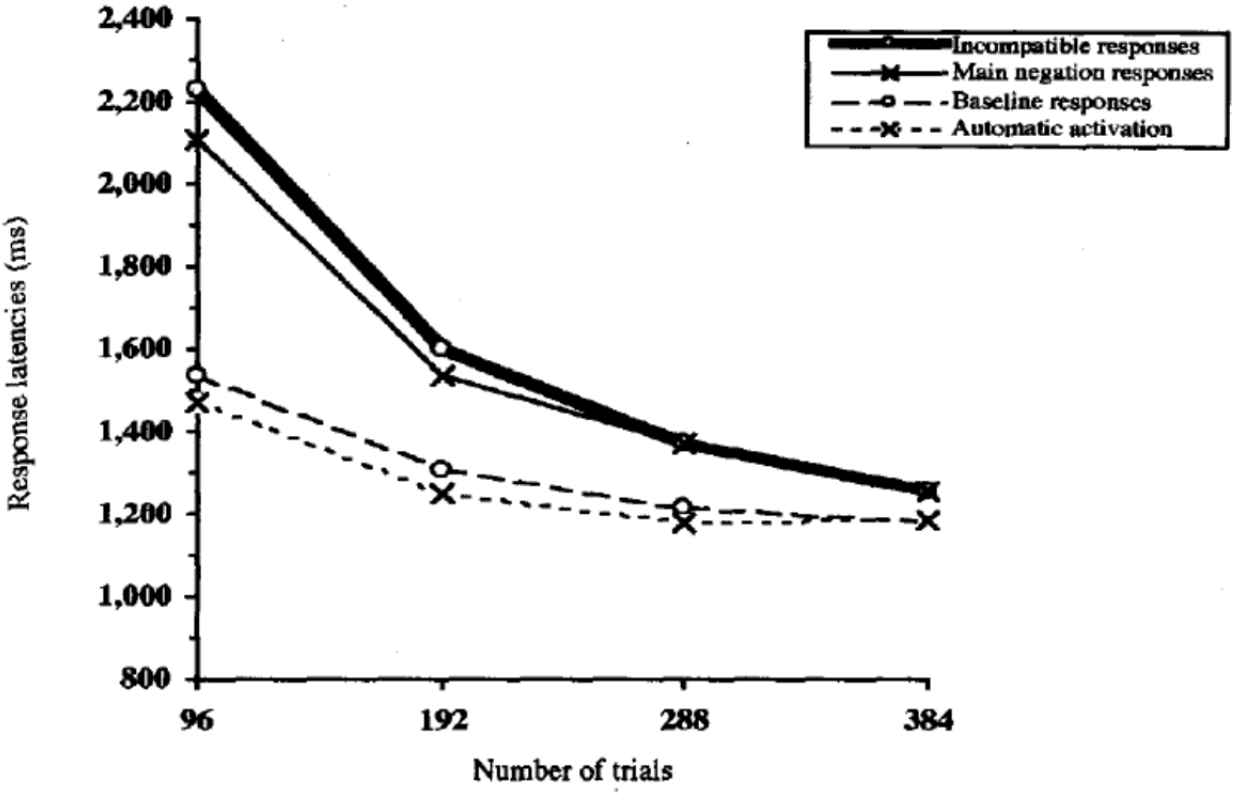

Figure 4. The effect of stereotype maintain and stereotype negation training on responses to consistent and inconsistent traits in the training task in Study 3.

In the above graph on response times from Kawakami et al. (2000):

The thick solid line indicates responding "YES" to counter-stereotypes.

The thin solid line indicates responding "NO" to stereotypes.

The thick dotted line indicates responding "NO" to counter-stereotypes.

The thin dotted line indicates responding "YES" to stereotypes.

[53] Gawronski et al. (2008), Johnson (2009), and Wennekers et al. (2012, 2013) reached significant effects with just 200 trials. One implication is that even if we were only to work through 200 trials, we might become significantly less biased than we are, but without reaching our full debiased potential. Of course, becoming *less* biased would still be desirable, even if, for whatever reason, becoming *completely unbiased* were practically unfeasible.

In other words, the bottom two curves represent default stereotypical responses, and the top two represent learned counterstereotypical responses. The steep drops of the top two curves show how participants became *faster and faster* at giving counterstereotypical responses from 100-200 trials, and from 200-300, following the "classic learning curve."[54]

At most, participants work through 480 trials. Yet 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 480 trials takes about 45 minutes. 45 minutes is *nothing*.

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. Yet the 45 minutes it takes to do so is miniscule in comparison to the tremendous resources that individuals, governments, schools, and businesses already devote to diversity initiatives and prejudice reduction, to say nothing of

[54] Of note: the initial large gap between the top two and the bottom two lines after 96 trials, the insignificant gap between the lines after 384 trials, and the significant distance between the top-two-after-384 trials and the bottom-two-after-96 trials.

the time and resources devoted to the education of democratic citizens, such as 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.[55] 45 minutes is less time than many people spend *per day* on exercise and the honing of other skills. American children spend an average of 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?[56]

At this point, I can only speculate about the sorts of biases that drive an automatic aversion to 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.[57] I suspect that a number of factors contribute to

[55] 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 part of the problem, a way to put off investing the work we know we need to. (Of course, 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.)

[56] Another common concern about feasibility is that there are a *lot* of implicit biases out there, and if it takes 45 minutes to fix each of them, then how many hours will it take to fix *all* of them? This is of course 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. I don't think, for example, that we should rule out particular proposals for *institutional* interventions on the grounds that they won't be equally effective at countering all possible forms of discrimination (e.g., 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). 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, or idiosyncratic hang-ups (we don't, in any case, all share exactly the same biases). The essential debiasing procedures for medical doctors, high-school guidance counselors, and employees in airport security might differ greatly (or they might not). Third, the finding that, e.g., debiasing racial *stereotypes* reduces racial *prejudice* suggests that some debiasing procedures might generalize in important respects (as I argued in section V). 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 (see the appendix). Fourth, 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 simply irrelevant how many hours it would take to eradicate every one of the relevant biases.

[57] A la Haidt (2001).

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 *A Clockwork Orange* that I am not advocating here.

Of course, nobody is weirded out 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.[58] That's why genuine, full-blooded retraining has to be part of the discussion. Just like unlearning

[58] Here I am bracketing the theoretical debate about whether implicit prejudices and stereotypes are, at bottom, a matter of beliefs, as opposed to a matter of habits, skills, know-how, mere associations, or aliefs. Nevertheless, in many cases, changing values might be at issue. Devine et al. (2012) found that increases in concern about discrimination seemed to moderate decreases in implicit bias. So changing the explicit stuff might be important, and it might be that retraining the implicit stuff is part of what changes the explicit stuff. But then these are explicit attitudes that many of us would at least acknowledge as valuable, even if these attitudes are not as strong and salient as they could be. That is, I believe I *ought* to be concerned about discrimination, and perhaps I ought to be more deeply concerned than I presently am. On the flip side, part of the opposition to debiasing might be an *attachment* to the attitudes that implicit biases support, e.g., by buttressing one's self-esteem and one's belief in a just world. This relates to a different sort of worry one might associate with *A Clockwork Orange*, that our debiasing interventions could have unexpected effects on us apart from bias reduction. Perhaps an effective intervention that reduces our biases will also make us chronically depressed or angry about global injustice. This is an empirical question like any other, which should be explored, but I think that if, say, removing white people's biases will also lower their self-esteem, then the solution is to find alternative sources of self-esteem. The benefits of reducing unfair behavior and systematic oppression will likely outweigh these unforeseen costs.

bad habits and learning new skills or languages, there simply has to be a central role for *practice*.[59]

In any event, these 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. 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,

[59] In conversation, Manuel Vargas and Michael Brownstein suggested that there might be some additional factors that explain (without really justifying) our kneejerk reluctance to debiasing. Brownstein suggested that it might have to do with the alienating perception that the training requires using myself (or my mind or body) as a mere means to an end. Vargas suggested that 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.

was also found to enhance implicit self-esteem. Perhaps research on approach training partly inspired a recent MSNBC commercial campaign, which featured ads that paired the progressive-sounding slogan "Lean Forward" with photos of its leading personalities:

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![60]

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'll just relearn the biases upon leaving the lab and that debiasing is creepy because it's like brainwashing. The anticipated relearning is presumably supposed to occur as a result of similarly brainwashing-esque procedures. 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 counter-brainwashing ourselves.[61]

I suspect that one of the most significant biases driving kneejerk pessimism about debiasing is the extent to which these studies *implicate us as individuals*. If individuals can really take their implicit biases into their own hands, that means *I* can do so, and if I can, then, other things being equal, I probably should. But if I can tell myself a plausible story about how it's a massive social-institutional problem that can't be solved at the individual level, then I don't have to feel bad for failing to take steps to improve myself. The primary oversight in this sort of self-deflecting bias 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*

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

[61] Another source of perceived creepiness (similar to Brownstein's suggestion 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.

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. There is almost a cottage industry devoted to understanding the cognitive biases that inhibit acting in response to these phenomena.[62] 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 social respect. This is just the sort of bias that approach training might help us to *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

[62] To see this, one need merely Google the phrase "cognitive bias" with any major social problem. For example, searching for "cognitive bias world poverty," turned up Thomas Pogge's (2008, 206) *World Poverty and Human Rights*: "This is due in part, no doubt, to powerful resistance against seeing oneself as connected to the unimaginable deprivations suffered by the global poor. This resistance biases us against data, arguments, and researchers liable to upset our preferred world view…This bias is reinforced by the cognitive tendency to overlook the causal significance of stable background factors (e.g. the role of atmospheric oxygen in the outbreak of a fire), as our attention is naturally drawn to geographically or temporally variable factors."

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.[63] I wonder whether this sort of desire for deep answers isn't responsible, in part, for the continued resistance to accepting that less sophisticated habits of thought, feeling, and action 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, or MacKinnon's account of discrimination before we get serious about combating discrimination. 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 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.)[64]

VII. Individualistic versus institutional approaches to discrimination

[63] 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."

[64] Nevertheless, I think there is another important intuition 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. My response is to invoke an analogy with linguistic fluency (see my 2012, chs.4-5 for more on the analogy). Memorizing vocabulary and grammar rules is not the same as becoming fluent in a second language. But you do need to memorize vocabulary and learn a bunch of rules before becoming truly fluent. These basic forms of training are the anti-prejudicial equivalent of memorizing flashcards. These are the *basics*, which will put you in a better position to actually *act* in unbiased ways in the real world, with all its inherent complexity.

Although activists generally agree that the pervasion of biased "microbehaviors" and judgments contributes to macro-level social injustices, many are skeptical of interventions that seek to change these microbehaviors by counterconditioning individuals' implicit biases. We should, they argue, instead focus on setting up institutional structures that preclude the operation of implicit biases in advance (such as blind reviewing) or counteract their operation after the fact (such as structures of affirmative action). I wholeheartedly support these structural interventions.[65] Far from being in *competition*, I believe debiasing will be integral to the successful implementation of broader systematic changes.

However, my first response to the claim that individualistic approaches are counterproductive is to ask just how productive institutional approaches have been. While the 20th century saw significant improvements in American legislation against *explicit* discrimination (although there is much more to do), there has been little progress in combatting less overt forms of discrimination.[66] A 5-4 Supreme Court 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. The overarching pattern has been to roll back existing structural interventions because they amount to "reverse" discrimination. So I fail to see how, in the contemporary political climate, institutional interventions have cornered the market on brass-tacks pragmatism.

Apart from asking how effective debiasing will be, then, we should ask *how much opposition will there be?* We can make counterstereotype or approach training widely available

[65] I am also extremely sympathetic with the criticism that philosophers have been especially prone to lose sight of the structural forest in the individualist trees. Legal theorists are way ahead of us on considering the political-institutional context of implicit bias. See the collection of papers in Levinson and Smith (eds.) (2012) *Implicit Racial Bias Across the Law* (none of which mentions Kawakami's research, although Dasgupta and Greenwald (2001) and Blair's (2002) review are cited).

[66] See, e.g., Lawrence (1987) for a seminal analysis of the failure to account for unconscious racism.

to individuals without overhauling institutional structures in potentially contentious ways. While we can (and 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. Giving individuals the free choice to take responsibility for debiasing themselves should appeal directly to the values of those who object to institutional interventions as paternalistic or reverse-discriminatory.

Debiasing is, in fact, often counted among the potential *benefits* of structural interventions such as affirmative action. Prominent psychologists and legal theorists have argued that promoting members of underrepresented groups to positions of prominence will produce "debiasing agents," counterstereotypical exemplars who debias their peers.[67] 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 (wrongly) perceive to be undue benefits, under-evaluate their performances in the future, and so on. For example, Kaiser and colleagues found that the mere presence of diversity-promoting structures can ironically lead some privileged individuals to become *more discriminatory*.[68] Given such findings, we cannot assume that institutional interventions will have debiasing effects. Implementing them without sufficient

[67] Jolls and Sunstein (2006) and Kang and Banaji (2006). (Anyone moved by Kantian objections to debiasing, i.e., that retraining itself involves using people as mere means (see notes #64 and 66), should be pretty alarmed by this proposal.) These theorists also argue that implicit bias constitutes a *new* justification for affirmative action, which is neither centered on redressing past injustice nor on paving the way for a better future. Research on implicit bias suggests that institutional redress may be necessary to counteract *ongoing discrimination*, which may be unwitting or unwilling. Of course, the reason that theorists need to come up with "new" arguments for affirmative action in the first place is that courts (at least in the US) have been pretty hard on such structural interventions in recent years.

[68] In a series of studies, Kaiser et al. (2012, 1) found that the mere presence of diversity-promoting structures produces "an illusory sense of fairness" in some privileged individuals, leading them to believe that the organization is procedurally fair, even when there is no evidence that the diversity structures are *effective*, and "even when it is clear that underrepresented groups have been unfairly disadvantaged." This perception of fairness, in turn, leads individuals "to legitimize the status quo by becoming less sensitive to discrimination targeted at underrepresented groups and reacting more harshly toward underrepresented group members who claim discrimination."

attention to the motivations 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, 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 also unfair and misguided. They are unfair because institutional sponsorship of debiasing need not take the (potentially) objectionable form of a universal debiasing mandate. 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 automatic dispositions of thinking, feeling, and acting into accord with the beliefs and values we already endorse, or at least claim to.

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 debiasing, 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), for example, 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 taking an approach mindset can "make interracial contact rewarding rather than depleting" (411).

Then again, maybe sometimes we should *approach egalitarianism* while at other times

we should *avoid prejudice*. Phills, Santelli, Kawakami, Struthers, and Higgins (2011) found that taking approach-equality strategies to positive images reduced implicit racial bias, as did taking avoid-prejudice strategies to negative images:

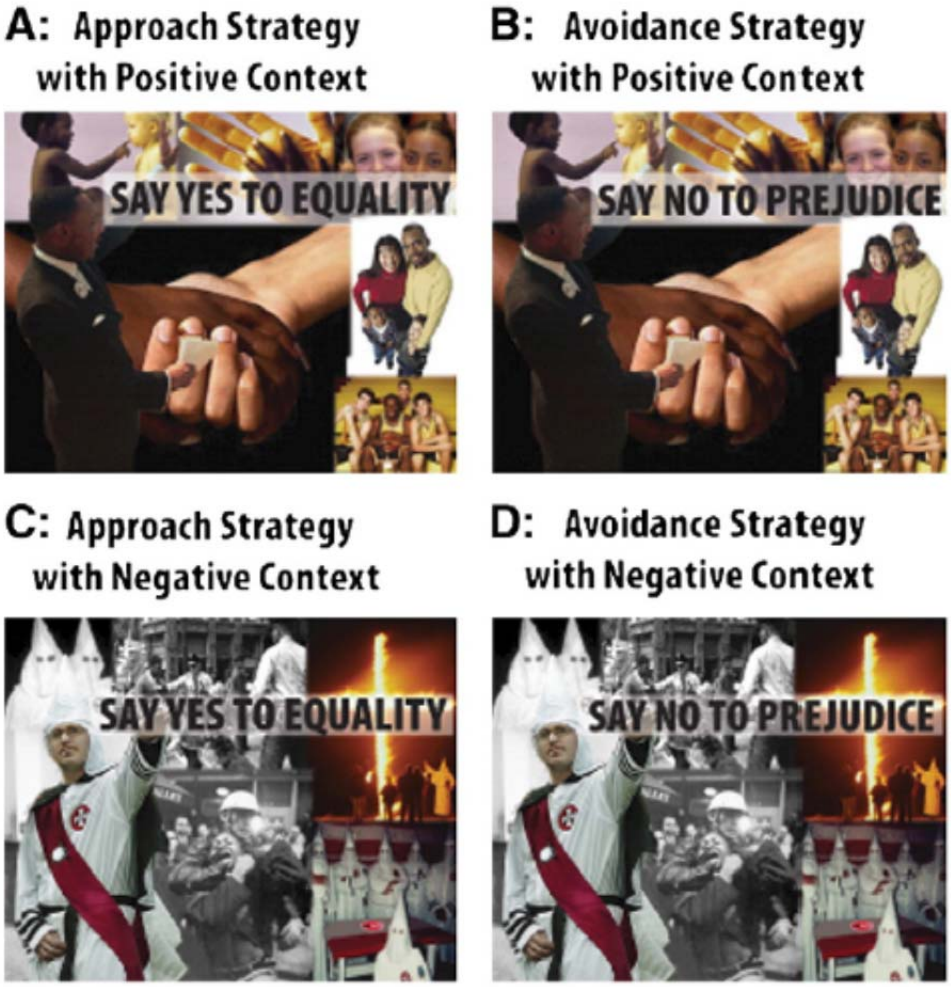

Fig. 1. Anti-racism advertisements presented to participants in Study 1. In particular, these advertisements combine an approach strategy with a positive (panel A) and negative (panel C) context as well as an avoidance strategy with a positive (panel B) and negative (panel D) context.

But it is far less effective to affirm equality while faced with images of the Ku Klux Klan (D in the above figure) and to negate prejudice while faced with images of Martin Luther King, Jr. 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*.

Bibliography

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

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

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.

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.

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.

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.

Garcia, S. M., Weaver, K., Moskowitz, G. B., & Darley, J. M. (2002). Crowded minds: The implicit bystander effect. *Journal of personality and social psychology, 83*(4), 843-853.

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.

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.

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.

Haslanger, S. 2012 lecture: Schemas and the Materiality of Social Practices. Presentation for *The Implicit Bias & Philosophy Workshop*. University of Sheffield, UK.

Houben, K., Nederkoorn, C., Wiers, R. W., & Jansen, A. (2011). Resisting temptation: Decreasing alcohol-related affect and drinking behavior by training response inhibition. *Drug and alcohol dependence*, *116*(1), 132-136.

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. Forthcoming: 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. (2012). *The Hidden Mechanisms of Prejudice: Implicit Bias & Interpersonal Fluency*. Doctoral dissertation, Columbia University, NY.

Madva, A. (Forthcoming). Virtue, Social Knowledge, and Implicit Bias. Volume on Implicit Bias and Philosophy, eds. Jennifer Saul and Michael Brownstein. Oxford University Press.

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. Manuscript: Attitude, Inference, Association: On the Propositional Structure of Implicit Bias.

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.

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.

Parker-Pope, T. October 31, 2008: Rewriting Your Nightmares. *The New York Times* URL = < http://well.blogs.nytimes.com/2008/10/31/rewriting-your-nightmares/ >

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.

Pogge, T. W. (2008). *World poverty and human rights*. Polity.

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.

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/ >

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.

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.