Alex Madva
Virtue, Social Knowledge, & Implicit Bias[1]

Forthcoming in *Implicit Bias and Philosophy*, edited by Jennifer Saul and Michael Brownstein, Oxford University Press.

Abstract
This paper is centered around an apparent tension that research on implicit bias raises between virtue and social knowledge. Research suggests that simply knowing what the prevalent stereotypes are leads individuals to act in prejudiced ways—biasing decisions about whom to trust and whom to ignore, whom to promote and whom to imprison—even if they reflectively reject those stereotypes. Because efforts to combat discrimination obviously depend on knowledge of stereotypes, a question arises about what to do next. I argue that the obstacle to virtue is not knowledge of stereotypes as such, but the "accessibility" of such knowledge to the agent who has it. "Accessibility" refers to how easily knowledge comes to mind. Social agents can acquire the requisite knowledge of stereotypes while resisting their pernicious influence, so long as that knowledge remains, in relevant contexts, relatively inaccessible.

I.   Introduction

Research suggests that most citizens of liberal democracies embrace egalitarian goals but

nevertheless exhibit predictable patterns of implicit social bias in, for example, hiring and legal

decisions. Recently, Tamar Szabó Gendler (2008b, 2011) and Andy Egan (2011) have argued

that implicit biases put us in a kind of tragic normative dilemma, in which we cannot jointly

satisfy all of our moral and epistemic requirements. Their arguments are, in part, responses to

---

alarming evidence that the strength of implicit biases correlates with the knowledge individuals have of prevalent stereotypes, regardless whether they reflectively reject or endorse the content of those stereotypes.[2] Simply knowing what the stereotypes are seems to make individuals more likely to act in biased ways.

These findings suggest an opposition between social knowledge and virtue. If mere knowledge of stereotypes hinders the possibility of acting ethically, should an individual with egalitarian goals aim to forget all she knows about discrimination and stereotypes? Returning to such a state of ignorance would incur serious costs. In such a state, she would not, for example, be in a position to recognize the injustices individuals suffer in virtue of being perceived in a stereotypical light. If, then, she ought not surrender her social knowledge, must she simply learn to live with the unfortunate effects that knowledge has?

I argue that this apparent opposition is misguided. Social knowledge as such poses no obstacle to virtue, but rather the relative *accessibility* of such knowledge does. In psychology, accessibility refers to "the ease with which a particular unit of information is activated or can be retrieved from memory" (Morewedge and Kahneman, 2010, 435). In some sense, certain bits of knowledge "come to mind" more readily than others. I intend to argue that the seeming tension between virtue and knowledge of stereotypes arises primarily insofar as that knowledge becomes too readily accessible. It is possible for social agents to acquire knowledge about the existence and effects of stereotypes, while working effectively toward being more virtuous, so long as that knowledge remains relatively *in*accessible.

---

[2] The term "stereotype" can be problematic and misleading, because some of the cognitive processes underlying objectionable stereotypes also underlie unobjectionable and even rational forms of social cognition. Valian (1998) adopts the less charged term "schema," while Beeghly (2013) argues that we should broaden our understanding of "stereotypes" to include both rational and problematic forms of social cognition. For my purposes, the negative connotation of stereotypes is useful for keeping in view that I am talking about a class of somehow *bad* or *undesirable* states. I agree with Antony (2002) that distinguishing the "good" stereotypes from the "bad" is fundamentally an empirical question. See also Brownstein and Madva (2012a,b).

This paper focuses on the problematic effects of social knowledge on the mental life and behavior of *individuals*, and, in turn, on what individuals can do about it. To consider what individuals can do is not to assume that the wrongs associated with implicit bias are primarily "individual" rather than "social." The harms and inequities suffered by individuals on the basis of race and gender depend to a great extent on social-institutional forces, and institutional change is necessary for redressing those harms. But institutions are composed both of a set of rules and laws as well as a set of individuals, and, if we want to bring about lasting change, we have to understand the roles that each plays in contributing to these large-scale harms. I focus on the role of individuals here. I intend to address what sorts of institutional policies might be warranted in light of these phenomena in future work.[3]

In what follows I recount the moral-epistemic dilemma posed by Gendler and Egan (§II). They argue that working toward the *ethical* ideal of being unprejudiced will inevitably incur certain "*epistemic* costs." I then critically examine the notion of an epistemic cost (§III), and I describe how an agent could, in principle, pursue her ethical aims without incurring any (§IV). Working toward this ideal ethical and epistemic state is no mere pipe dream, but depends on an agent's ability to regulate the accessibility of her social knowledge. I explain how the regulation of accessibility is key for resolving the dilemma (§V). I then describe a range of concrete strategies individuals can implement to better approximate the moral-epistemic ideal (§VI). Finally I consider in greater detail one of the contexts in which our ethical and epistemic aims are alleged to come into conflict (specifically, deciding whether to consider the racial composition of neighborhoods in setting home insurance premiums; §VII). This is very much

---

[3] See, e.g., Valian (1998, 2010) for many of the best institutional interventions for addressing gender bias in professional contexts; Kalev et al. (2006) for an informative meta-analysis of existing strategies to promote diversity; Levinson and Smith (2012) for a collection of essays on implicit racial bias and legal theory; and Anderson (2010, 2012) for pioneering work on prejudice and political philosophy.

the beginning of a larger treatment of these issues. A satisfactory account of how social

knowledge is accessed, and how to thwart the pernicious influence that such knowledge can have

on judgment and action, requires answering a series of interrelated questions in epistemology,

ethics, and psychology. My hope is to shed some light on which questions we should be asking

and make tentative proposals about how to answer them.

II.   A Moral-Epistemic Dilemma?

The point of departure for Gendler and Egan (the "dilemmists," as I will call them) is empirical

research on the undesirable effects that social knowledge can have on behavior. For example,

Correll and colleagues (2002) found that the magnitude of individuals' implicit racial bias did

not correlate with their self-reported racial beliefs but did correlate with their reports of what

most white Americans believe.[4] In other words, it seems that merely knowing what *many others*

believe about a group leads individuals to act in some respects as if they themselves believed it,

too.

Another problematic type of social knowledge might be knowledge of statistical

regularities of demographic variation, such as average differences between social groups in

crime rates and math SAT scores (Gendler, 2011, 56). Take, for example, evidence suggesting

that a woman sitting at the head of a table is often less likely to be identified as the group leader

than a man in the same position (Porter and Geis 1981). As Virginia Valian explains:

> failing to perceive a woman at the head of the table as the leader may have no
> discriminatory impetus behind it. On average, a woman is less likely to be a

---

[4] See Arkes and Tetlock (2004) for a review. See Nosek and Hansen (2008), Jost et al. (2009), and Uhlmann et al. (2012) for responses to Arkes and Tetlock's interpretation of implicit measures as reflecting *nothing but* innocuous cultural knowledge.

> leader of a group than a man is… Observers may be responding to the situation only on the basis of what is most likely, and men are more often leaders, wherever they sit. It is also important to notice, though, that regardless of the reason, a female leader sitting at the head of a table loses out compared to a male leader… She is less likely to obtain the automatic deference that marks of leadership confer upon men. Her position will be weakened—even if observers do not intend to undermine her authority. (1998, 127)

It seems that the knowledge that women are less likely to occupy leadership positions makes people (men and women alike) less likely to *treat* women as leaders. In other words, merely knowing *what is statistically likely* about a group leads individuals to act in some respects as if those statistical generalizations were *normative*, as if members of that group *ought* to be treated in a certain way (e.g., as if marks of leadership confer authority to men but not women). The general upshot seems to be that merely knowing certain social facts makes individuals more likely to act in biased ways. In what follows, I will, for ease of presentation, use the phrase "knowledge of stereotypes" as an umbrella term to refer jointly to all of these potentially problematic forms of social knowledge.[5]

The dilemmists claim that findings like these put us in an inescapable moral-epistemic bind, suggesting that cases will inevitably arise in which social categories are *epistemically* relevant but *ethically* objectionable. On one side of the dilemma, we have a range of epistemic goods, e.g., knowing about prevalent stereotypes, others' beliefs, demographic facts, and so on. On the other side, we have a range of ethical goods, e.g., treating people fairly, with respect, as individuals, and so on. Pursuing the former seems to compromise the latter. As Gendler (2011, 57) writes, "Living in a society structured by race appears to make it impossible to be both rational and equitable."

---

[5] I speak of "knowledge," rather than mere "awareness," of stereotypes, because epistemic justification is at issue. There is no moral-epistemic dilemma if the beliefs in question are false or unjustified. I ignore complications regarding whether true justified belief counts as knowledge.

Roughly, the dilemmists offer three ways we might respond to these findings, all of which come at a price:

> [1] We can use the categories unreflectively, and wind up with a bunch of bad stereotype-concordant inferences, judgments, attitudes, etc. [2] Alternatively, we can use the categories, but spend a bunch of cognitive resources suppressing or immediately excising the bad stereotype-concordant inferences, judgments, attitudes, etc. [3] Finally, we can avoid using the categories, and fail to code up the base rate information. (Egan 2011, 72)

In other words, the possible responses seem to be:

> (STATUS QUO) To continue unreflectively using stereotypes, or perhaps even come to reflectively endorse their content;

> (SUPPRESSION) To try to actively suppress the expression of our stereotypical thoughts and impulses;

> (IGNORANCE) To aim to ignore or forget the stereotypes altogether.[6]

I assume that STATUS QUO does not strike my readers as an attractive option. It amounts to waving the white flag in the fight against prejudice. The drawbacks of SUPPRESSION, according to the dilemmists, are that continually monitoring and suppressing our stereotypical thoughts threatens to make us so cognitively exhausted that we become less effective in reaching our other epistemic aims. It might even be self-defeating.[7] By trying to become ethically better we might become epistemically worse. But what about IGNORANCE? Suppose that we could erase knowledge of stereotypes from our minds, and by doing so, eliminate all the implicitly biased behaviors that harm ourselves and others. Our automatic, intuitive responses would be

---

[6] They describe the dilemma and the three possible responses in slightly different ways at different times (Gendler 2008b, 578; 2011, 37-38, 57; Egan 2011, 72-3). Egan notes that these responses need not be mutually exclusive.
[7] As when trying not to think about a white bear makes you think about a white bear. See Follenfant and Ric (2010), Huebner (2009) and Madva (2012, 103-104, 158-165) for discussion of several ways that efforts to resist stereotype-consistent thoughts and behaviors can backfire.

perfectly aligned with our reflective commitments. Egalitarian sunshine of the spotless mind. What would be the downside?

The dilemmists suggest that this option would also be epistemically suboptimal. We would lose out on information, namely, knowledge of what others believe, of demographic regularities, and so on (Gendler 2011, 56). By losing this information we suffer an epistemic cost. Ultimately, the dilemmists seem to think this cost is worth bearing, and that we should embrace some combination of SUPPRESSION and IGNORANCE.[8] But in what sense is IGNORANCE an epistemic cost? For that matter, what is it to *be* an epistemic cost? Answering this requires taking a stand on some broader issues in epistemology, which I do in the next section (§III). I then propose (§IV) that the dilemmists have either misdescribed IGNORANCE, insofar as we need not forget about stereotypes altogether in order to prevent their pernicious influence on behavior, or that they have failed to countenance a fourth option, namely, of *regulating* the cognitive accessibility of stereotypes, such that we can recall them when they are relevant and ignore them when not.

III.   The Aims of Knowledge

Our epistemic aim is not to know *everything*. Perhaps this is contentious, but it does not seem true that we have an omnipresent epistemic goal to achieve universal knowledge of all the facts

---

[8] Egan (2011, 78). Much of my concern here has to do with the claim that these responses are inherently epistemically costly. I doubt the dilemmists and I would disagree much over which practical strategies to pursue in order to make our behavior more aligned with our ethical aims, but rather over whether pursuing these strategies would force us to compromise our epistemic aims. I have benefited greatly from reading a longer, unpublished manuscript of Gendler (2011), in which she considers some of these ethically valuable but (putatively) epistemically costly strategies in greater depth.

there are. This is true neither of our everyday epistemic practices nor of our scientific inquiries

(nor of our theories of the theory of knowledge).

Given that we are not out to know it all, the sheer loss of information is not, just as such,

an epistemically bad turn of events. There are lots of facts we simply do not care about, such as

whether the number of oxygen atoms in the room is even or odd. Doubtless there are many facts

out there that we do not care about but *should*.[9] I should invest more effort than I do in

determining how big my carbon footprint is and what I can do to reduce it. At the same time,

however, there are lots of facts we *do* care about, but should not. I know plenty of facts that I

could do just as well without. Once upon a time it seemed important to memorize all the lyrics

to the opening theme song of *The Fresh Prince of Bel-Air*, but I would gladly trade in that

knowledge now to free up mental space for something else. How many hours did I spend

learning to write in cursive as a child only for it to become an almost perfectly useless skill as an

adult? Whether bits of knowledge like these are worth seeking or keeping is not a matter of their

intrinsic value, but of their value relative to some further aim. I think that memorizing that

theme song seemed important to me because I hoped it would impress my friends, but knowing

the lyrics is of little use to me now (except at retro or campy dance parties).

To consider examples nearer to the topic at hand, take our basic cognitive dispositions to

sort people into categories. From the beginnings of infancy, we start making distinctions

between ingroups and outgroups, and forming specific expectations about how respective

members of these groups will behave (Valian 1998; Leslie 2009). These cognitive dispositions

---

[9] Sometimes our ignorance reflects a motivated and systematic avoidance of information. See Mills' (1997, 18) discussion of the "epistemology of ignorance, a particular pattern of localized and global cognitive dysfunctions (which are psychologically and socially functional), producing the ironic outcome that whites will in general be unable to understand the world they themselves have made." I consider the moral culpability individuals might bear for subtle acts of selective attention and ignorance elsewhere (2012, Ch.3).

are surely indispensable, enabling us to deal with the overwhelming complexity of information in the world as well as the underwhelming poverty of information that confronts us at any given moment. We could not accomplish much of anything without them; they serve some pretty fundamental aims. Nevertheless, to grant that these cognitive dispositions are indispensable *on the whole* is not to say that their exercise on any particular occasion is useful or accurate. While there are plenty of useful categories and regularities that we pick up on, there are plenty of useless ones to which we devote undue attention and plenty of useful ones that we miss altogether.

Consider the relative difficulty people have in recognizing the faces of outgroup members, which Gendler (2011) discusses at length. One component of this outgroup recognition deficit seems to be that participants fail to notice individuating facial features and devote undue attention to others' outgroup status. This recognition deficit occurs for faces from different races as well as from different *classes*: middle-class white participants have a harder time recognizing faces seen in "impoverished contexts" than in "wealthy contexts" (Shriver et al. 2008). They hone in on the poverty of the social context (which is, given the task at hand, irrelevant) at the expense of noticing the idiosyncratic facial features that would enable recognition later on.

Recent evidence suggests that as soon as children begin to appreciate these group differences, they also begin to invest them with spurious significance, e.g., by implicitly *preferring* ingroups over outgroups, and high-status social groups over low (Dunham et al. 2013). Anti-outgroup biases begin to form during the first months of infancy (Ziv and Banaji 2012) and remain surprisingly stable through adulthood. Yet while it is relatively easy to imagine how a default disposition to, say, prefer the company of the rich and powerful might be

adaptive from an evolutionary perspective (e.g., because hangers-on could share in their abundant resources), such a disposition is pretty clearly bogus from a normative perspective (be it epistemic or ethical). Many ingroup preferences form during childhood, but decades of research also suggest that assigning adults into patently arbitrary groups can rapidly generate ingroup preferences as well (Ashburn-Nardo et al. 2001).

Perhaps the most notorious and egregious examples of investing social categories with undue epistemic significance are the so-called "fundamental" and "ultimate" attribution errors, both of which involve systematic asymmetries in the beliefs we tend to form about causes of behavior. We are more likely to seek out personality-based explanations—and ignore situational factors—when we try to understand the bad behavior and mistakes made by others (the fundamental error), especially when they are outgroup members (the ultimate error!), but exhibit the reverse tendency when it comes to explaining the mistakes made by us and our affiliates. For example, a white manager might assume that a black employee's lateness is due to laziness, while his white co-manager's lateness is due to unforeseeable traffic delays caused by a car accident.

Such cases of excessive attention to certain categories, and inattention to others (which are perhaps more predictive and relevant), comprise but a few examples of a more general feature of our epistemic state: that we know and attend to lots of things we should not, and don't know and ignore lots of things we should. There are lots of facts that we would do just as well to forget or ignore because our preoccupation with them prevents us from seeking out and remembering information more relevant to our ends. This is not to assume that it is always straightforward to determine which information is relevant for which aims, but the difficulty of figuring out what is relevant cuts both ways. It is not as if we should just think about as much as

we can, as the phrase "all things considered" suggests. Considering as much as we can because anything could be relevant is obviously going to be self-defeating. Considering patently irrelevant information, e.g., considering whether Pluto should be counted as a planet while deciding whether to turn left or right at a busy intersection, is, on its face, a bad epistemic practice, not just a gratuitous one. Sometimes it is worthwhile to err on the side of considering too much rather than too little, but doing so is nevertheless *to err*, to make (or increase the risk making) a certain sort of mistake. It is to consider something that is "beside the point" or "neither here nor there," i.e., irrelevant to the truth of the proposition in question.

I grant that, at the level of scientific and communal pursuits, it might very often be preferable to err on the side of taking in too much information rather than too little. (Bring on Big Data!) Something that we take to be irrelevant now might turn out to be important down the line. Perhaps, then, we should make an extra effort to be inclusive in how many hypotheses we entertain, how much data we gather, etc., but it obviously doesn't follow that we ought to take "every" hypothesis seriously, gather "all" possible data on a hypothesis, or treat everything that seems irrelevant as if it were relevant (if such norms even make sense). The history of science is replete with examples of our propensity to fixate on less predictive categories and overlook more relevant ones. We can safely say that the extensive efforts once devoted to finding a Philosopher's Stone that would turn lead into gold were wrongheaded. Such efforts were less guided by any sort of rational, inductive considerations than by fantasies of wealth and power.[10] More recently, it is absurd how much empirical attention is devoted to uncovering potential evolutionary, biological, or neuroscientific explanations for the underrepresentation of women in certain fields, e.g., the current hypothesis that prenatal hormone exposure predisposes girls to be

---

[10] This is, of course, not to indict alchemy *in general*, the history of which is intricately tied to chemistry and medicine (Principe 2013), but only to find fault with one of its notoriously misguided pursuits.

innately less *interested* in quantitative subjects than boys (Jordan-Young 2010). Scientists are earnestly investigating the possibility that prenatal testosterone makes boys *like math more* than girls (a possibility that is ostensibly more "PC" than the hotly debated proposal that testosterone makes boys *better* at math). While the effects of prenatal hormone exposure on interest in math is an empirical question like any other, the extensive efforts devoted to uncovering such gender differences seem misplaced, especially given the overwhelmingly credible, empirically well-supported alternative explanations in terms of social and institutional factors. It is epistemically unwarranted to devote *so much* attention to these hypotheses. Indeed, this is plausibly a case of attribution errors writ large, in that researchers are searching for explanations in terms of enduring features of our gendered brains instead of our situations.[11]

The upshot is that the loss (and the acquisition) of information is not epistemically good or bad in itself, but only so relative to some more particular aims or values. Is it an epistemic deficiency not to know the names of foreign countries and their leaders? If you want to be a public official, definitely yes. If you just want to run a chain of pizza restaurants, perhaps no. Maybe there are *some* things worth knowing for their own sake, without qualification, such as the Form of the Good envisioned by Plato. But stereotypes do not fall into that category.

IV.   **The Right Thought at the Wrong Time?**

---

[11] Note also that Gendler and Egan frame the moral-epistemic dilemma in terms of a single individual's cognitive limitations, not in terms of the communal advancement of scientific knowledge. The dilemma arises on the assumption that an individual can notice and track only a sharply circumscribed number and range of properties, and must therefore glom onto those properties that give her the most inductive bang for her cognitive buck. By contrast, communal pursuits of knowledge are not so sharply constrained in this way. This asymmetry between individual and communal epistemic constraints explains why scientific endeavors can afford, and even benefit from, taking in ostensibly irrelevant information, while it is typically epistemically perilous for individuals to do the same.

Stereotypes might seem to be just the sorts of items we would be eager to forget. Insofar as they are often false or misleading, who needs them? Unfortunately, matters are not so simple. For one thing, it is not obvious that stereotypes are generally false (although they may still often be misleading), as in Valian's discussion of how individuals tend not to assume that a woman at the head of the table is a leader.[12] We can, however, bracket questions regarding their accuracy, because it is clearly important for individuals with egalitarian aims to know about stereotypes *at least insofar* as such knowledge enables them to recognize cases when someone suffers by virtue of being perceived in a stereotypical light.[13] For example, suppose a job candidate is not hired because the employer judged, on the basis of an objectionable stereotype, that people from the job candidate's race or gender are ill-qualified for the job. We lose the ability to identify the wrong that was done if we cannot make reference to the stereotype.

Stereotypes are worth knowing, then, not for their own sake, but for specific purposes. Does the importance of retaining our knowledge of stereotypes rule out an option like IGNORANCE? It does insofar as we ought not forget about stereotypes *altogether*, but maybe we don't really need to. We simply need to be able to think about the stereotypes for certain purposes and in certain contexts, and not in others. We go epistemically astray insofar as our knowledge of stereotypes is *accessed* or *activated* in the wrong contexts and for the wrong ends.

The dilemma with which we began, however, was ethical as well as epistemic. Knowledge of stereotypes seems to make us act out of step with our considered ethical commitments. However, a second look at the data suggests that that *mere possession* of

---

[12] See Beeghly (2013) for further discussion. Another common error is to radically over- or underestimate the relevant probabilities (e.g., sharks are more likely to attack people than are other creatures of the sea, but radically less likely to do so than most people believe). I believe the sort of hyper-accessibility I discuss here plays a major role in distorting estimations of probability, but I will not take up that issue here.

[13] There will be many cases when it is important to know about stereotypes. This case is among the least objectionable.

knowledge of stereotypes does not just as such tend to lead to implicitly biased behavior—

anymore than mere possession of knowledge of the *falsity* of stereotypes tends to lead to

implicitly *unbiased* behavior. Typically, studies that find relationships between implicit bias and

social knowledge are about *highly accessible* knowledge of *culturally prevalent* stereotypes

(Arkes and Tetlock 2004). The problem is not mere social knowledge, but rather *hyperactive*

social knowledge, agitating our minds in moments when it ought to keep silent. The normative

costs of social knowledge arise primarily insofar as that knowledge becomes too accessible**.** We

ought to access the stereotypes when they are relevant, and ignore them when not. I take the

operative "ought" here to be both epistemic and ethical. We should, if possible, embrace the

response of giving up some measure of cognitive access to our social knowledge. Clearly, this

does not actually amount to returning to a state of total ignorance about social stereotypes. The

aim is not to unlearn what we know, and thereby surrender (potentially valuable) knowledge.

We want to reduce the accessibility of knowledge of stereotypes, such that it is not the first thing

that comes to mind, but not lose access to it altogether. It seems, then, that the dilemmists have

mischaracterized IGNORANCE (and perhaps SUPPRESSION as well, as some of the evidence I

introduce below suggests). Alternatively, it might be more accurate to conclude that the list of

three possible responses offered by the dilemmists is incomplete, and I am recommending a

fourth option.[14] While the dilemmists refer to accessibility at several points (Gendler 2011, 37;

Egan 2011, 74), they seem not to appreciate the possibility of *regulating* the accessibility of our

social knowledge in order to have that information available when and only when we need it.

So how might this be possible? By way of suggesting how it is, I say something, first,

about how this sort of cognitive access is understood in psychology (§V), and second, about

---

[14] Thanks to Erin Beeghly and Alice Crary, who separately urged this reformulation.

what we as individuals can do to *change* the accessibility of our knowledge—in particular, our knowledge of stereotypes and other problematic social information (§VI).

V.   Primer on Accessibility

A substantial body of research shows that our decisions and actions are often swayed by the bits of knowledge that are most "accessible." In some sense, some ideas come to mind more "readily" or "easily" than others. For example, I know that in the United States, individuals of Asian descent are stereotyped to excel in math. I also know that this stereotype is less prevalent in Canada.[15] Although I have not taken a test measuring my implicit associations between Asians and mathematical ability, I should assume that I am like most Americans (including Asian Americans) and that the American stereotypes of Asians are more accessible to me than the Canadian ones; I have been repeatedly exposed to the former and not the latter. (Nor do I predict that my biases would shift were I to cross the border from Vermont to Quebec.) Knowledge of Asian-math stereotypes has been so pounded into our heads as to become *chronically accessible*. It comes too often and too easily to mind; the mere perception of a cue related to Asians or math might activate it (although we might not be fully aware that such a stereotype has been activated, or of how it influences our thoughts and behaviors).

Over the years we learn many stereotypes that do not stick in the same way. I was once told that black people think that "white people smell like wet dogs when they come out of the

---

[15] I learned this from a study on the accessibility of math-aptitude stereotypes for women of Asian descent. Shih et al. (1999) found that Asian-American female undergraduates whose "Asian identity" had surreptitiously been made salient performed better on a math test than those whose "Female identity" had been made salient and those who received no priming. Researchers ran the same study with high school students in Vancouver, Canada, and found that in this case activating an Asian identity *degraded* mathematical performance. The math-aptitude stereotype may be less prevalent in Vancouver because "the Asian community is largely recently immigrated" (82).

rain." Hearing this certainly made an impression, but I can safely say that I have never tested it

or modified my automatic dispositions (such as they are) to sniff or avert my nose around

members of different races who have just come out of the rain, or to act insecurely around non-

white people after I myself have just come out of the rain. Aside from being the first example

that comes to mind when I try to think of a silly stereotype, this item of knowledge is *not*

chronically accessible.

Knowledge can also be *temporarily* or *transiently accessible*. Suppose that upon arrival

in Montreal, a number of locals, including a guide at the tourism information desk, offer me this

inside tip: "You know, in Montreal, the elderly are exceptionally good drivers. You should

always try to hail taxis driven by very old people while you're here." I might briefly, for the

next fifteen minutes or so, show some increased disposition to look favorably upon cab drivers of

advanced years, but it is unlikely that my newfound knowledge (of what the Montréalais believe

of the elderly) would have any enduring effect on my behavior. I might recall the information

from time to time, but it would not remain close to the mental surface. Given my extensive

socialization into a world where the elderly are routinely depicted as bad drivers, this newfound

knowledge would exert little influence on my day-to-day dealings. Similarly, if you imagine

yourself in a post-apocalyptic world in which all the flowers are poisoned with radiation and the

insects are the only healthy things to eat, your typical preference for roses over roaches will

briefly be reversed.[16] You will, on some measures, temporarily show an implicit preference for

insects over flowers. But this sort of imaginative exercise will not transform you into a bug lover

---

[16] See Foroni and Mayr (2005), although Han et al. (2010) found that, while these transient effects occur for the standard IAT that uses words such as "pleasant/unpleasant," they do not occur for the "personalized" IAT that uses the words "I like/I don't like." Han, Fazio, and colleagues argue the standard IAT often measures mere "extrapersonal associations" (roughly akin to mere knowledge of prevalent stereotypes believed by others), whereas others (Nosek and Hansen 2008) take the grammatical sentences of the personalized IAT to induce greater self-regulation.

if you were not one already; the thoughts and feelings that come most readily to mind when you think of insects will, in a few days at the most, be much as they were before. In cases like these, a bit of knowledge is just transiently accessible, briefly rendered salient by virtue of some anecdote or exemplar.

It is fair to wonder, however, what this talk of accessibility really *means*. The natural ways of explaining it are metaphorical. Something is accessible, e.g., if it is hovering "in the back of your mind." Psychological definitions of accessibility are not entirely perspicuous, either. Morewedge and Kahneman (2010, 435) define accessibility as "the ease with which a particular unit of information is activated or can be retrieved from memory." I have some intuitive sense of what they are talking about, but the notion of "ease of access" is rather obscure. One thing this definition brings out is that accessibility is intimately tied to the notion of *knowledge activation*. Eitam and Higgins define accessibility and activation reciprocally:

> When initially conceived, accessibility referred to the ease with which a mental representation *could* be activated by external stimulation, and activation meant that a representation *has* been accessed for use… In other words, a mental representation's accessibility referred to the amount of external stimulation needed for it to shift from a latent state (available in the mind but currently inactive) to an active one (involved in current thought and action). (2010, 951)

Again, I more or less know what they mean, but these definitions are just cycling through synonyms: it is accessible in the sense that it is available; available in the sense that it is easily retrieved; easily retrieved in the sense that it is easily activated.

Reflecting on how accessibility is measured goes some way toward illuminating what it is. In effect, a bit of knowledge is said to be more accessible to the extent that an individual is more likely to recall it upon request, or to recall it faster. (Thus, the moral-epistemic dilemma arises in part because stereotypes are too likely and too quick to reach the mental surface.)

Accessibility is then often defined computationally as "the probability of retrieval."[17] "Ease" of retrieval is thereby replaced with "probability." This definition, however, still invokes "retrieval," which is effectively synonymous with "access" (as if to say, "retrieve-ability refers to the probability of retrieval"). Terms like "retrieval" do little more than re-label commonsense notions of *remembering* in terms of a storage-space metaphor.[18] The underlying psychological constructs and operations remain murky. For now, perhaps the best way to understand accessibility is to understand how to *change* it.

VI.   The Malleability of Accessibility: Concrete Strategies

Reducing the accessibility of our knowledge only constitutes a normative cost, whether epistemic or ethical, if we cannot access that knowledge when we need it. If we can intervene to influence the accessibility of our knowledge in the right way, we can mutually satisfy (or at least come significantly closer to mutually satisfying) our epistemic as well as ethical aims. Can we do this?

It is not just wishful thinking to envision lining up the relative accessibility of our knowledge with our ethical aims. There is a significant tradition of research suggesting that accessibility is often highly goal-dependent.[19] What most readily comes to mind is often a

---

[17] See Bahrick (1971). Thanks to Edouard Machery for emphasizing this.

[18] Fine-grained distinctions between these terms (accessibility, availability, activation, retrieval, recall) are sometimes made, but I ignore them here. Another concern is that it is completely non-obvious why *probability* and *speed* of recall should be lumped together as two measures of a single construct, accessibility.

[19] See Kunda and Spencer's (2003) goal-dependent theory of stereotype activation (which they define as "the extent to which a stereotype is accessible in one's mind" (522)). For reviews, see Moskowitz (2010) and Uhlmann et al. (2010). Eitam and Higgins (2011) seek to explain accessibility *entirely* in terms of the "motivational relevance" of a bit of knowledge to an agent, and even recommend replacing the term "accessibility" with "relevance" (e.g., distinguishing "chronic" from "transient" relevance (960)). Given its obscurity, abandoning the term "accessibility" might be wise, but is "relevance" an obvious improvement? That equally fraught term/concept threatens to muddle

function of what is most relevant to our aims. Some of the most important aims in this regard are those of being fair and egalitarian, as I explain below. But there are surprising ways in which other aims can help us reduce the accessibility of stereotypes as well.

For example, one way to block the activation of stereotypical thoughts seems to be to adopt the aim of being *creative*. Stereotypical thinking is *typical* thinking; it is unoriginal. Agents who are motivated to think creatively will automatically ignore stereotypical associations. In Sassenberg and Moskowitz (2005), some participants were put in a "creative mindset" by being asked to recall a few occasions in which they had been creative. Participants next performed a lexical decision task, which required them to identify as fast as possible whether letters on a computer screen made up a real word or not. They saw either a black or white male face immediately followed by a nonsense word, a stereotype-consistent word, or a stereotype-irrelevant word. For example, an image of a black face followed by the word "rhythmic" would be a stereotype-consistent pairing. Those who had been primed to be creative were significantly slower to identify stereotype-consistent words than stereotype-irrelevant words, whereas participants in other conditions exhibited the reverse tendency. The stereotypical associations were irrelevant to the task of distinguishing words from non-words, and the goal of being creative enabled participants to ignore that irrelevant but otherwise chronically accessible knowledge. Being in a creative mindset prevented the stereotypes from being the first thoughts to come to mind.[20]

description with prescription, i.e., the distinction between what agents happen to take to be relevant and what actually (as a normative matter) *is* relevant. I use the term "relevance" here exclusively in the normative sense.
[20] This study may point to certain *aesthetic* implications of research on implicit cognition. Recommending that we should resist relying on stereotypes in art might ring in some ears as advocating the oppression of artistic creativity. But studies like this suggest that stereotypes are precisely *not* creative. Of course, creativity is a complex phenomenon. For example, an artist—or a comedian (Anderson, April 2012)—can appeal to a stereotype in order to subvert it. My point is simply that research on social cognition does not bespeak any fundamental opposition between our aesthetic and ethical aims (any more than it does between our epistemic and ethical aims). Reducing

In that case, the goal to be creative was more or less unwittingly activated. Are there conscious, intentional strategies for influencing accessibility? This may be exactly the effect of implementation intentions, which are if-then plans that link a specific cue to a specific response, such as, "If I feel a craving for cigarettes, then I will chew gum!" Concrete plans specifying when, where, and how an action will be performed are far more effective than unspecific plans, such as, "I should cut down on smoking!" Webb and Sheeran (2008) argue that these if-then plans work in part by making the specified cue more accessible. Participants who formed an if-then plan to retrieve a coupon after the experiment exhibited heightened access to the "if" components of the plan on a lexical decision task like the one described above. They were also almost twice as likely to follow through on retrieving the coupon than participants who did not form an if-then plan. In contexts where we know we are wont to attend to the wrong categories, we can form if-then plans to guide our attention to the right ones.

The influence of implementation intentions on "shooter bias" is a case in point. Participants must press a button labeled "shoot" when they see an image of a person holding a gun, and "don't shoot" when they see a person holding a cell phone. Many participants, including African Americans, are faster and more likely to "shoot" unarmed blacks than unarmed whites. Mendoza and colleagues' (2010) found that participants' shooter bias was significantly reduced after they formed such intentions as, "If I see a gun, then I will shoot!" This intention plausibly makes the relevant cue (the gun) more accessible, and makes the irrelevant cues (such as the race of the person holding it) less accessible. In fact, there is no *more* relevant cue than the gun in this context. Participants are directing their attention to the precise property required by the task, so, given the dilemmists' assumption that our cognitive finitude requires us to focus

---

the accessibility of stereotypes might even *improve* creativity by preventing us from falling back on hackneyed depictions of members of social groups.

on just a few features, this strategy incurs pure epistemic benefit with no cost. Stewart and Payne (2008) found that participants who formed the intention to think the word "safe" when they saw a black face also showed significantly less implicit racial bias. Thinking counterstereotypical thoughts seems to reduce the accessibility of stereotypical ones.

Insofar as we already know the content of many stereotypes, and we are learning more and more about the contexts in which stereotypes do harm, then we can formulate if-then plans that pick out those contexts and specify responses. In the case of shooter bias, researchers identified a context in which racial stereotypes are highly accessible, and specified responses that counteract the stereotype. Generally speaking, in cases where stereotypes are apt to do harm, the right if-then plans might have a structure roughly along the lines of, "If I perceive a member of group G in context C (acting in way W, or being treated in way Y), then I will perform action A!!" During Q&A after a lecture in December 2011, Louise Antony proposed that a person concerned to avoid over-interrupting women could form the plan, "If she's talking, then I won't!"

Many of these plans will be *other-directed*, in that they regard how we treat others, but we can also form *self-directed* plans to arm ourselves against potentially harmful environmental cues and stereotype threat. For example, a person stereotyped to underperform in math can block the negative effects of stereotype threat by forming the plan: "And if I start a new problem, then I will tell myself: I can solve it!" Bayer and Gollwitzer (2007) found that participants who rehearsed this plan solved significantly more problems on a test of logical reasoning than those who rehearsed other plans, such as, "And I will tell myself: I can solve these problems!"

I mentioned that the accessibility of knowledge is highly goal-dependent, but I have not discussed the goals most relevant to reducing the accessibility of stereotypes, namely, the aims to

be fair and unbiased. There is substantial evidence that these aims are effective as well, whether they are held chronically or transiently (Moskowitz 2010). The goal to be unprejudiced can be induced, for example, by having participants contemplate a time when they *failed* to live up to this ideal. Some agents seem to have chronic egalitarian goals that are automatically activated when they find themselves in a situation in which they or others might be inclined to act in a biased way. Perhaps they have internalized the plan, "If I see injustice, then I will fight it!"

Figuring out precisely which if-then plans to adopt is a substantive project, but I see no reason to be pessimistic about taking that project pretty far. For example, perhaps one appropriate plan might be, "If I see a woman at the head of the table, then I will treat her like a leader!" Then again, perhaps the plan ought to refer generically to a *person* at the head of the table, which could help reduce the influence of other sorts of biases, e.g., regarding race, class, and disability.[21] However, if men are generally more likely to be seated at the heads of tables, then referring generically to a person might simply reinforce our initial disposition to assume men are leaders. I am skeptical that the "person" formulation would incur such unforeseen, undesirable consequences, but it is an open empirical matter. My broader point is that, like all heuristics and rules of thumb, no simple if-then plan (or collection of such plans) will provide a universally accurate guide to true belief or right action. Implementation intentions are not a magical cure-all for discrimination. Nor will they empower us to transcend human finitude and fallibility. Whichever plans we identify as most effective for advancing our aims, there will always be tradeoffs, with certain plans working better in certain contexts and for certain purposes but not others.

---

[21] Thanks to Jenny Saul for helping me think through this example.

As far as I can tell, however, these tradeoffs need not be between our ethical and epistemic aims. The initial dilemma was that knowledge of stereotypes led to unethical behavior, and that ethical improvement would incur epistemic costs. Supposing that implementation intentions lead to ethical improvement, do they bring the costs of SUPPRESSION and IGNORANCE in tow?

Employing these if-then plans constitutes a *kind* of self-control, but they do not sap our limited cognitive resources. Implementing them is not a matter of engaging in SUPPRESSION—or, to the extent that they do involve SUPPRESSION, then that option need not be as epistemically costly as the dilemmists suggest. Rehearsing them is easy and, once formed, putting them into action often requires little or no additional effort. Gollwitzer and Sheeran (2006) go so far as to call them "instant habits." Of course, there will be limitations to how radically an if-then plan can reconfigure our ingrained dispositions, but there is no question that these plans provide a powerful way to influence the accessibility of our knowledge. Another way to reduce the accessibility of stereotypes is to *practice*, e.g., repeatedly affirming counterstereotypes. For example, repeatedly pressing a button labeled "YES" in response to counterstereotypical stimuli, such as a black face paired with the word "friendly," significantly reduces stereotype accessibility (Kawakami et al. 2000).

Moreover, these interventions do not cause participants to completely forget about stereotypes. It is implausible that a person with the right retraining, mindset, or if-then plan would somehow cease to know the contents of prevalent stereotypes. She would not suddenly profess ignorance about stereotypes regarding, say, Asians and mathematical aptitude. She

might, however, have a harder time coming up with a list of such stereotypes out of the blue.[22]

This is, I suggest, essentially what we are after in trying to line up the accessibility of our

knowledge with our ethical aim of treating others fairly.

I by no means intend to suggest that implementing these and other strategies will swiftly

transport us to an ideal state of moral-epistemic virtue where we can access our knowledge at all

and only the right times and where all our impulses and habits will be effortlessly unprejudiced.

Rehearsing these if-then plans and training procedures are, I submit, the anti-prejudicial

equivalent of using flashcards to memorize a new vocabulary. The expectation is not that

adopting these cue-behavior rules of thumb will transform us into ethically and epistemically

ideal agents, any more than memorizing a set of rules and words will transform an individual

into a fluent speaker of a second language.[23] Nevertheless, these if-then plans enable us to make

meaningful progress in that direction, towards lining up our accessible social knowledge with our

considered aim of being egalitarian—without incurring epistemic costs. We can significantly

close the distance between our current sorry state and our normative ideals, and, in closing this

distance, we need not make a forced choice between pursuing ethical and epistemic aims.

VII.   Tragic Cases?

In (§VI), I emphasized cases, like the influence of a creativity mindset, where stereotypes are

irrelevant. What happens when stereotypes *are* relevant to the task at hand? Not to trivialize the

---

[22] I doubt this has been tested, but I predict that such debiased participants would be able to check off a list of stereotypes put in front of them, but be less able to generate an extensive list of such stereotypes without prompting. Some evidence suggests, however, that open-ended list-generating is an unreliable indicator of stereotype knowledge and accessibility (Nosek and Hansen 2008).

[23] For more on the analogy between egalitarian agency and linguistic fluency, see my (2012, ch.4).

issue, but that's when we should access them. The dilemmists, however, seem to have cases in mind where stereotypes are, specifically, epistemically relevant but ethically objectionable.

A prominent case that Gendler and Egan discuss is Tetlock and colleagues' (2000) study on "forbidden base rates." Participants were asked to imagine an executive setting the premiums on home insurance in various neighborhoods. All participants were told that some of the neighborhoods were higher risk than the others. Some participants were also told that the high-risk neighborhoods were predominantly black, while others received no information about race. Here is the entire scenario participants read in the race-relevant condition:

> Dave Johnson is an insurance executive who must make a decision about whether his company will start writing home insurance policies in six different towns in his state. He classifies three of the towns as high risk: 10% of the houses suffer damage from fire or break-ins each year. It turns out that 85% of the population of these towns is Black. He classifies the other three towns as relatively low risk: less than 1% of the houses suffer fire or break-in damage each year. It turns out that 85% of the population of these towns is White. (Tetlock, 2000, 860-1)

Participants who were not given any race-related information tended to say that the executive should charge a higher insurance premium for the houses in the high-risk neighborhoods, but those given the information about race insisted that he should charge the same premium for all (this was especially true for politically liberal participants). Gendler (2011, 55) refers to this as "a kind of epistemic self-censorship on non-epistemic grounds." Tetlock and colleagues suggest that these "people are striving to achieve neither epistemic nor utilitarian goals" (2000, 853), and specifically compare this case to the classic base-rate neglect literature (854).

However, it is not clear why we should think of this as a dilemma between epistemic and non-epistemic (moral) requirements. In fact, we don't know which requirements are at issue because we don't know which problem the participants took themselves to be solving. We don't know what their *aims* were. The participants were not, for example, told in advance that they

should make the decision in the best economic interests of the insurance executive. All they

were told was that "the research goal was to explore how people make judgments" (860).

Plausibly, the participants who just read about high-risk neighborhoods and insurance premiums,

without any reference to race, thought the task *was* just how to maximize profits, or something

similar. But once race was introduced, they may very well have thought the task was how best to

avoid being prejudiced, or even how to compensate for systemic racial injustice! This is quite

plausible in light of how *overt* the reference to race was. Even Tetlock and colleagues explain

that:

> liberals did not indiscriminately embrace any justification for not using the base
> rates. Liberals viewed the pragmatic or empirical grounds offered for dismissing
> the base rates as implausible. They were not more inclined to challenge the
> statistics or to argue that the best long-term profit-maximizing strategy is to
> charge the same price. Instead, liberals invoked a straightforward moral defense
> against policies that harmed the already disadvantaged. (863-4)

Relative to the aim of preventing racial injustice, there is nothing epistemically deficient

about discounting the background information about high-risk neighborhoods. Participants may

have been explicitly *counting* that information as part of the justification for charging the same

premiums, because this policy would prevent the exacerbation of injustice. The real test would

be if participants who had categorically adopted the goal of maximizing profits still overlooked

this information when race was introduced.

If I am right about the Tetlock study, then its failure to constitute a genuine moral-

epistemic dilemma is telling. Such cases might be harder to find than one might think. What

seemed like a moral-epistemic dilemma might be no dilemma at all, because a *morally tinged*

aim sets the epistemic agenda. There is nothing inherently wrong with the epistemic agenda

being set by a moral aim; there has got to be some aim or other, and it might as well be a moral

one. Consider, for example, the long tradition of research seeking to identify factors that predict

death sentencing. Recently, Eberhardt and colleagues (2006) found that death sentencing "is influenced by the degree to which a Black defendant is perceived to have a stereotypically Black appearance" (383), and Williams and Holcomb (2004) found that a death sentence is significantly more likely in cases when the victim is white and female. Of course, knowledge of these predictive factors is not being pursued "for its own sake," but because we want to know whether the practice of capital punishment is *fair*. If the race and gender of defendants and victims are strong predictive factors, that gives us reason to believe that the practice is not fair. The purely epistemic project of acquiring true beliefs about the factors that predict death sentences is, in other words, set in motion by ethical aims.[24]

I believe that there can be tragic normative conflicts with no ideal solution, such as Sartre's classic example of the individual torn between fighting in the revolution and staying home to take care of his ailing mother, and I will not offer a principled argument against the possibility of similar unresolvable conflicts arising between ethics and epistemology. I submit, however, that we have not been given definitive examples of these which pit the ethical fight against prejudice against the epistemic project of identifying the properties that give us the most inductive bang for our buck. Insofar as genuine conflicts between epistemic and ethical aims do arise in lived experience, solving them may often not be a matter of choosing which to pursue and which to sacrifice, but reconsidering the merits of the aims themselves. I have assumed that we are holding our aims fixed, but they themselves can be called into question. If our ethical and epistemic aims conflict, that may signal that we are operating with the *wrong* aims. Participants who ignored the information about high-risk neighborhoods may not have been compromising their epistemic aims; they may have been (rationally) revising them.

---

[24] See Anderson (2004) for more on the role of "value judgments" in science.

I hope these considerations suggest that the dilemma is less of a theoretical nature, about the principled impossibility of jointly satisfying competing normative requirements, and more of a practical nature, about what we can do concretely to thwart the pernicious influence that knowledge of stereotypes has on our judgment and behavior. There is reason to be optimistic that we can improve along this ethical front, and that we can do so without compromising our epistemic aims, if only we try.

Bibliography

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia*, *19*(1), 1-24.

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

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

Anderson, L. 2012 April: Why So Serious? An Inquiry into Racist Jokes. Presentation for the *Implicit Bias & Philosophy Workshop*. University of Sheffield, UK.

Antony, L. (2002). Quine as feminist: The radical import of naturalized epistemology. In *A mind of one's own*, *Feminist essays on reason and objectivity* 2nd edition, Antony, L. M., Witt, C., & Atherton, M. (Eds.). Boulder, CO: Westview Press.

Arkes, H.R., and Tetlock, P.E. 2004: Attributions of implicit prejudice, or "would Jesse Jackson 'fail' the Implicit Association Test?". *Psychological Inquiry* 15, 257-278.

Ashburn-Nardo, L., Voils, C.I., and Monteith, M.J. 2001: Implicit associations as seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology*, 81, 789-799.

Bahrick, H.P. 1971: Accessibility and Availability of Retrieval Cues in the Retention of a Categorized List. *Journal of Experimental Psychology* 89 (1), 117-125.

Bayer, C., and Gollwitzer, P.M. 2007: Boosting scholastic test scores by willpower: The role of implementation intentions. *Self and Identity* 6, 1-19.

Beeghly, E. 2013: *Seeing Difference: The Epistemology and Ethics of Stereotyping*. PhD diss., University of California, Berkeley, California.

Brownstein, M.S, & Madva, A.M (2012a). Ethical automaticity. *Philosophy of the Social Sciences*, *42*(1), 68-98.

Brownstein, M.S., and Madva, A.M. 2012b: The Normativity of Automaticity. *Mind & Language* 27 (4), 410-434.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. 2002: The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology* 83 (6), 1314-1329.

Dunham, Y., Chen, E., & Banaji, M.R. (2013). Two signatures of implicit intergroup attitudes: Developmental invariance and early enculturation. *Psychological Science*.

Eberhardt, J.L., Davies, P.G., Purdie-Vaughns, V.J., and Johnson, S.L. 2006: Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes. *Psychological Science* 17 (5), 383-386.

Egan, A. 2011: Comments on Gendler's "The epistemic costs of implicit bias." *Philosophical Studies* 156, 65-79.

Eitam, B. and Higgins, E.T. 2010: Motivation in Mental Accessibility: Relevance of a Representation (ROAR) as a New Framework. *Social and Personality Psychology Compass* 4 (10), 951-967.

Follenfant, A., & Ric, F. (2010). Behavioral rebound following stereotype suppression. *European Journal of Social Psychology*, *40*(5), 774-782.

Foroni, F. and Mayr, U. 2005: The power of a story: New, automatic associations from a single reading of a short scenario. *Psychonomic Bulletin & Review* 12 (1), 139-144.

Gendler, T.S. 2008a: Alief and belief. *The Journal of Philosophy* 105 (10), 634-663.

Gendler, T.S. 2008b: Alief in action (and reaction). *Mind and Language* 23 (5), 552-585.

Gendler, T.S. 2011: On the epistemic costs of implicit bias. *Philosophical Studies* 156, 33-63.

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

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

Jordan-Young, R. 2010: *Brain Storm: The Flaws in the Science of Sex Differences*. Cambridge, MA: Harvard University Press.

Jost, J.T., Rudman, L.A., Blair, I.V., Carney, D.R., Dasgupta, N., Glaser, J., and Hardin, C.D.. 2009: The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior* 29, 39-69.

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.

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review*, *71*(4), 589-617.

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 .

Kim, B. 2012: *The Context-Sensitivity of Rationality and Knowledge*. PhD diss., Columbia University, New York.

Kunda, Z., & 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.

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

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

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

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

Mills, C.W. 1997: *The Racial Contract*. Ithaca: Cornell University Press.

Morewedge, C.K. and Kahneman, D. 2010: Associative Processes in Intuitive Judgment," *Trends in Cognitive Sciences* 14 (10), 435-440.

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

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

Porter, N., and Geis, F.L. 1981: Women and Nonverbal Leadership Cues: When Seeing Is Not Believing. In *Gender, Androgyny, and Nonverbal Behavior*, edited by C. Mayo and N. Henley. New York: Springer Verlag, Inc.

Principe, L. M. (2012). *The secrets of alchemy*. University of Chicago Press.

Sassenberg, K. and Moskowitz, G.B. 2005: Don't stereotype, think different! Overcoming automatic stereotype activation by mindset priming. *Journal of Experimental Social Psychology* 41, 506-514.

Shriver, E. R., Young, S. G., Hugenberg, K., Bernstein, M. J., & Lanter, J. R. (2008). Class, race, and the face: Social context modulates the cross-race effect in face recognition. *Personality and Social Psychology Bulletin*, *34*(2), 260-274.

Shih, M., Pittinsky, T.L., and Ambady, N. 1999: Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, 10, 80-83.

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.

Tetlock, P.F., Kristel, O., Elson, B., Green, M., and Lerner, J. 2000: The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78(5), 853-870.

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.

Uhlmann, E. L., Poehlman, T. A., & Nosek, B. A. (2012). Automatic Associations: Personal Attitudes or Cultural Knowledge? *Ideology, Psychology, and Law*.

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

Valian, V. (2010). What works and what doesn't: How to increase the representation of women in academia and business. In *Gender Change in Academia* (pp. 317-328). VS Verlag für Sozialwissenschaften.

Webb, T.L. and Sheeran, P. 2008: Mechanisms of implementation intention effects: The role of goal intentions, self-efficacy, and accessibility of plan components. *British Journal of Social Psychology*, 47, 373–395.

Williams, M.R. and Holcomb, J.E. 2004: The Interactive Effects of Victim Race and Gender on Death Sentence Disparity Findings. *Homicide Studies* 8 (4), 350-376.

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