

# On the epistemic costs of implicit bias

Tamar Szabó Gendler

Published online: 1 September 2011  
© Springer Science+Business Media B.V. 2011

## 1 Prologue: two apparently unrelated stories (don't let the details distract you...)

### 1.1 Base rates and the CIA

The Central Intelligence Agency's *Psychology of Intelligence Analysis* (Heuer 1999) provides concise and practical summaries of recent work in cognitive psychology using real-world examples intended to help CIA agents perform their daily activities more effectively.<sup>1</sup> Its 12th chapter—"Biases in Estimating Probabilities"—describes a number of widely-discussed findings from the heuristics and biases tradition. The chapter points out that among the errors that a well-trained CIA agent should aim to avoid are mistakes arising from what is commonly referred to as the "base-rate fallacy." Here is an example<sup>2</sup>:

During the Vietnam War, a fighter plane made a non-fatal strafing attack on a US aerial reconnaissance mission at twilight. Both Cambodian and Vietnamese jets operate in the area. You know the following facts:

---

<sup>1</sup> In his Foreword, CIA Deputy Director of Intelligence Douglas MacEachin notes: "I know from first-hand encounters that many CIA officers tend to react skeptically to treatises on analytic epistemology. This is understandable." (Heuer 1999) I know likewise from first-hand encounters that many analytic epistemologists tend to react skeptically to CIA treatises on intelligence analysis (cf. Gendler and Hawthorne 2005; as well as responses to von Fintel and Gillies 2008). While this too is understandable, it would be misplaced in this particular case.

<sup>2</sup> As Heuer notes, the case is a terminological variant (due to Frank J. Stech) of Kahneman and Tversky's (1972) classic green taxicab (85%)/blue taxicab (15%) case.; cf. also (Kahneman and Tversky 1973).

---

T. S. Gendler (✉)  
Yale University, New Haven, CT, USA  
e-mail: tamar.gendler@yale.edu

- (a) Specific case information: The US pilot identified the fighter as Cambodian. The pilot's aircraft recognition capabilities were tested under appropriate visibility and flight conditions. When presented with a sample of fighters (half with Vietnamese markings and half with Cambodian) the pilot made correct identifications 80 percent of the time and erred 20 percent of the time.
- (b) Base rate data: 85 percent of the jet fighters in that area are Vietnamese; 15 percent are Cambodian.

Question: What is the probability that the fighter was Cambodian rather than Vietnamese?

When the example is presented in this sort of probabilistic format, participants typically form their estimates by focusing on the case-specific information in (a) rather than the base-rate information in (b), with most giving estimates well over 50%, and many giving estimates in the range of 80%.

In fact, the probability is actually only 41%, which can be seen most easily if we consider the case in terms of frequencies.<sup>3</sup> (Nothing in the remainder of the article turns on the details here, so unless you are in a particularly conscientious mood, you may skip to the last sentence of this paragraph.) The information in (a) tells us that the pilot makes correct identifications 8 out of 10 times. The information in (b) tells us that for every 100 planes he will encounter, 85 will be Vietnamese and 15 Cambodian. Together, (a) and (b) imply that of those 85 + 15 planes, he will correctly identify 68 of the 85 Vietnamese planes as Vietnamese, and incorrectly identify 17 of them as Cambodian; likewise, he will correctly identify 12 of the 15 Cambodian planes as Cambodian, and incorrectly identify 3 of them as Vietnamese. This means for every 100 planes, he will identify 71 as Vietnamese (68 correctly, 3 incorrectly) and 29 as Cambodian (12 correctly, 17 incorrectly). So of planes that he identifies as Cambodian, 12/29—that is, 41%—will actually be Cambodian (whereas the remaining 17—that is, 59%—will actually be Vietnamese). So *even though our pilot is correct in his identifications 80% of the time, the relative rarity of Cambodian planes means that even a plane he identifies as Cambodian is still more likely to be Vietnamese.*

Failure to take into consideration background information about the relative distribution of properties is a classic failure of reasoning. When assessing your evidence, if you want to be rational, *you need to take base rates into account.* If you want to minimize the likelihood of making mistakes in situations where you are operating under something less than certainty—and this, of course, is most of the time—you would do well to pay careful attention to the distribution of

<sup>3</sup> Though human beings are notoriously bad at making use of explicitly-presented probabilistic base rate information (like the 85:15 information in the airplanes case), we are actually very good at making use of base rate information that is acquired through implicit learning or presented to us in terms of frequencies rather than probabilities. (Indeed, this fact will play a role in the next example.) For detailed discussion of and evidence for this claim, see Koehler (1996). (Thanks to David Hilbert for emphasizing the need to include reference to this body of work, and for pointing out the internal tensions in earlier drafts of the paper that resulted from my not having done so explicitly.)

properties across individuals of various categories. Keep this in mind as we continue.<sup>4</sup>

## 1.2 John Hope Franklin and the Cosmos Club

In the summer of 1995, historian John Hope Franklin—author of *From Slavery to Freedom*—received a call from the White House informing him that President Clinton planned to present him with the Presidential Medal of Freedom, the nation’s highest civilian honor. On the night before the award ceremony, Franklin hosted a dinner for a small group of friends at the Cosmos Club, a Washington DC social organization of which he was a member. He writes: “It was during our stroll through the club that a ...woman called me out, presented me with her coat check, and ordered me to bring her coat. I patiently told her that if she would present her coat to a uniformed attendant, ‘and all of the club attendants were in uniform,’ perhaps she could get her coat” (Franklin 2005, p. 340).

Clearly, the woman who presented Professor Franklin with her coat check was guilty of some sort of error. But one error of which she seems clearly not to have been guilty is that of base-rate neglect. (Actually, things are a bit more complicated—but let’s follow this line of thought for a bit.) Franklin had been the Cosmos Club’s first black member, and was still one of its very few. By contrast, nearly all of the club’s numerous attendants were men of African descent. So when the woman was presented with the visual experience of a black man in the club’s lobby, she endorsed an empirically well-supported hypothesis—one that took full account of prior probabilities. The likelihood that a black man present in the Cosmos Club was a member of the staff rather than a member of the club was very high—high enough, perhaps, to make it rational to assume that even though he was wearing a suit rather than a uniform, he was nonetheless an employee rather than a host.

There’s a lot more to talk about here, but before doing that, I want to talk about you. You, dear reader, now know that John Hope Franklin was a black man. Perhaps you knew that before you started reading this paper. But if you didn’t, I’m willing to wager that you formed the belief that John Hope Franklin was black before you read the sentence telling you that “Franklin had been the Cosmos Club’s first black member.” You probably formed it when you read the phrase “author of *From Slavery to Freedom*,” which I included at that point in the essay precisely so that I could make this point at this one. But if you didn’t, you almost certainly formed it when you read Franklin’s story about being presented with a “woman[’s]...coat check.”<sup>5</sup>

<sup>4</sup> Readers who enjoyed this brief excursion into CIA publications and who are looking to while (not “wile”; Zwicky 2008) away a few spare hours might enjoy playing some of the Agency’s games for children at <https://www.cia.gov/kids-page/games/index.htm>. (Central Intelligence Agency 2010) (Readers wondering how I while away mine are invited to follow this paper’s footnotes for a guided tour.)

<sup>5</sup> Actually, in the original quotation, Franklin writes: “It was during our stroll through the club that a *white* woman called me out, presented me with her coat check, and ordered me to bring her coat” (italics added). I have omitted the crucial term because including it renders race so salient that it would provide an alternative explanation for the response, diluting the force of my *tu quoque*. (Thanks to Keith DeRose for pointing this out.)

Now, let me ask you a question. On what basis did you form your belief? Presumably on grounds similar to the ones that the woman herself used. You took into consideration facts about base rates and cultural associations. “The author of a book entitled *From Slavery to Freedom*” you tacitly thought to yourself, “is likely to be of African descent. A man to whom a coat check would be presented at a DC social club is presumably a person of dark complexion. Examples that would be presented in a paper such as this one are presumably included to illustrate points about race.” And, of course, unlike the woman who used base rate information to infer that Franklin was a club attendant, you were correct in your conclusion that Franklin was black.

Perhaps part of the reason you were correct is that the prior probabilities were different in the two cases. Perhaps there’s a greater chance that a black man in the lobby of the Cosmos Club is a club member than that the author of *From Slavery to Freedom* is white. (Though, in fact, there isn’t.<sup>6</sup>) But this is hardly the issue. As the old joke goes—attributed alternatively to Churchill, Shaw and Twain—“we’ve already established what kind of person you are: we’re just quibbling about the price.”<sup>7</sup>

### 1.3 You (and me) (and our minds) (and our society)

What kind of person you are, of course, is one with capacities and categories that enable you to encode certain sorts of regularities in your environment. If you weren’t that kind of person, you certainly wouldn’t be reading this article. You wouldn’t be reading anything at all—or speaking a human language, or thinking complex thoughts, or perceiving objects in the way that you do. People are categorizers—and a good thing too—else the world in all its specificity would be unnavigable: it would be, to borrow an overused phrase, a “blooming, buzzing confusion” (James 1890, p. 462).

This is hardly news. If you’ve read pretty much any philosopher in the last, say, 2000 years—or taken Introduction to Psychology, or paid careful attention to your child during her first year, or reflected systematically on the nature of your own experience—you’ve probably thought about some of the features and implications of our tendency to make sense of the world using categories and extrapolations.

But to me at least, it was surprising to discover that the existence of race as a category that gives rise to certain sorts of automatic associations is, for those who disavow the normative content of these associations, hazardous in a particular way:

<sup>6</sup> Indeed, as far as I can tell, of the seven authors listed in the first substantive paragraph of the Digital History site’s “succinct essay on the best books on the history of slavery,” *none* is of African descent. (Digital History n.d.) So if this was your (tacit) reasoning process, you should be credited not with accurate base-rate encoding, but—more likely—with employing what psychologists call the “representativeness heuristic.” (Thanks to Steve Darwall for reminding me that most books on African-American slavery are in fact by non-black authors.)

<sup>7</sup> In case you are fortunate enough not to know the joke, here’s a version from Wikiquote, with George Bernard Shaw (“GBS”) in the leading role. GBS: “Madam, would you sleep with me for a million pounds?” Actress: “My goodness, Well, I’d certainly think about it.” GBS: “Would you sleep with me for a pound?” Actress: “Certainly not! What kind of woman do you think I am?!” GBS: “Madam, we’ve already established that. Now we are hagglng about the price.” (“George Bernard Shaw” 2011).

racially-based inequities—and the psychological processes by which we inevitably encode them—appear to carry not merely moral, but also epistemic, costs.<sup>8</sup> The costs take various forms, and in the remainder of the paper I'll consider several examples in some detail. But if you're looking for a (not completely accurate) slogan to hold in mind throughout the paper, here's one: if you live in a society structured by racial categories that you disavow, either you must pay the epistemic cost of failing to encode certain sorts of base-rate or background information about cultural categories, or you must expend epistemic energy regulating the inevitable associations to which that information—encoded in ways to guarantee availability—gives rise.<sup>9</sup> This dilemma is particularly profound for white Americans who occupy positions of class privilege. I certainly fall into that demographic, and I suspect many of my readers do as well. But in my discussion below, I will also discuss ways in which the implicit coding of racial (and other) stereotypes can be cognitively costly for all members of a society.<sup>10</sup>

In the remainder of the paper, I will proceed as follows. I will begin by presenting a bit of background concerning the development and structure of racial categories. I'll then describe three phenomena—difficulties in cross-race face identification, stereotype threat, and cognitive depletion following interracial interaction—each of which brings out a particular type of epistemic cost incurred by participants who are sensitive to race in ways that give rise to certain sorts of automatic associations whose contents they disavow.<sup>11</sup> (These are three very different sorts of cases, most

---

<sup>8</sup> I am thus trying to cash out in more detail some remarks I made in the closing pages of Gendler (2008b), where I remarked: “[L]iving in a society that violates one's normative ideals has unavoidable cognitive consequences. For either you will need to deliberately restrict your attention or experiences so as not to encode certain sorts of genuine regularities (for example, by deliberately preventing yourself from acquiring and attending to the fact that, in contemporary American society, certain racial categories are associated with certain sorts of highly-valenced affective content.) Or you will need to engage in alief-driven rationalization, changing your normative ideals to accord with the relevant sorts of experienced regularity (for example, by coming to endorse the legitimacy of these stereotypical associations.) Or you will experience the cognitive costs of disharmony, redeploing cognitive energy to suppress the pull of your belief-discordant aliefs (for example, by expending executive control in cases of interracial interaction to suppress your aliefs, thereby temporarily depleting your cognitive resources.) This is the trichotomy of norm-discordant alief.” (Gendler 2008b, p. 578).

<sup>9</sup> Matters here are actually quite complicated, and those familiar with the relevant psychological literature will be aware that I am skating over a large number of important distinctions, both in the base rate literature itself (again, see Kohler 1996 for a review) as well as more specific work on the relation between stereotypes and base-rate neglect (e.g., Locksley et al. 1980; Kunda and Spencer 2003.) Since the goal of this (already overlong) paper is just to explore one aspect of this incredibly complex and multifaceted phenomenon, I beg forgiveness for neglecting these important subtleties.

<sup>10</sup> An important bibliographic caveat. In my discussion below, I will make reference to a wide range of empirical literature in psychology, but almost none to the tremendously important work in philosophy that has addressed these questions. It should be obvious to those familiar with those works that I have been influenced profoundly in my thinking by books such as (Alcoff 2006; Anderson 2010; Appiah and Gutman 1996; Blum 2002; Mills 1997; Shelby 2005; Sullivan 2006; as well as numerous articles by these and other authors.)

<sup>11</sup> In this paper, I focus largely on racial categories in late 20th and early 21st century America. But the basic psychological mechanisms that underlie the general points I am making appear to be universal. See, for example, Hodson et al. (2005) (a study of aversive racism outside the US); Sangrigoli et al. (2005) (a study of cross-race face identification in Europe); Tanaka et al. 2004 (a cross-cultural study of cross-race face identification); and Dunham et al. (2006) (a study exploring implicit bias in America and Japan).

likely operating through three different sorts of mechanisms, but I introduce them to give the reader a sense of the range of ways in which implicit racial bias can give rise to epistemic costs.) One way of avoiding at least some of these costs would seem to be through base-rate neglect or a failure to encode information about cultural associations, since the associations in question would then not be automatically activated. But this, of course, carries its own epistemic problems, ones to which I adverted in the opening pages. I discuss these issues in the final substantive section. I conclude by suggesting that what the argument reveals are the epistemic costs of situations where there are systematic discrepancies between the way things are and the way you wish things to be.<sup>12</sup>

Before beginning, it is important to emphasize the problem I am addressing here generalizes<sup>13</sup>: it arises for any prevalent yet disavowed categorization schema. This makes the issues discussed below more interesting in some ways, and less interesting in others. They are more interesting because the phenomenon in question appears to tell us something quite general about the human mind, less interesting because they are, in many ways, just another example of the consequences of our finitude. That said, it may nonetheless be interesting to recognize explicitly that whenever there are discrepancies between our automatic or habitual or default responses on the one hand, and our reflective commitments on the other, we can expect to pay an epistemic price.

## 2 Background

In the discussion that follows, I take as established a number of framework-setting results from contemporary social psychology. The purpose of this section is to make these assumptions explicit, and to offer some background on the empirical work that justifies them. This section has three parts: the first is a rapid overview of psychological work on categorization and stereotyping; the second is a short discussion of my own work on alief; the third is a quick presentation of recent work on racial categories, including an introduction to the distinction between *dominant* and *aversive racism*, followed by some brief remarks on some consequences of the material presented throughout the section.

### 2.1 Categorization and stereotyping

Categorization and stereotyping are tools used by finite minds to operate effectively in an overwhelmingly complex environment. Cognizing entities in ways that involve classifying them as members of larger collections solves two problems simultaneously. On the one hand, it allows us to navigate a world whose complexity

---

<sup>12</sup> I had originally hoped to include a final section describing promising strategies for overcoming some of the phenomena discussed in this paper, but the essay is already far over its allotted length. I hope to explore these issues in a subsequent paper (especially since eliminating that section required me to leave out some of the paper's best jokes.) For a philosophically informed overview of some of this literature, see Kelly and Roedder (2008) and Kelly et al. (2010).

<sup>13</sup> Thanks to Matthew Noah Smith for encouraging me to be explicit about this point from the outset.

exceeds our cognitive capacities: the world is complicated, our cognitive resources are limited, and classifying objects into groups allows us to proceed effectively in an environment teeming with overwhelming detail. On the other hand, it allows us to navigate a world where in any given encounter, an object presents us with only a few of its potentially relevant features. Classifying objects into groups allows us to readily extrapolate properties that that object is likely to share with other members of its kind.<sup>14</sup> So categorization is not an optional way of making sense of the world: it is our means of dealing with the problems of global complexity and experience poverty.

Given this, it's perhaps unsurprising that categorization is something that comes on line early and forcefully in the course of human development. Here's a particularly nice characterization of the phenomenon from a recent paper by Sarah-Jane Leslie (Leslie forthcoming<sup>15</sup>):

Long before we learn to talk, our expectations concerning novel members of a category are shaped by our experience with already encountered members. We expect, for example, that objects that share obvious perceptible qualities will also share dispositional properties. If a given item rattles when shaken, nine-month olds expect other items which share the same perceptible profile will rattle when shaken (Baldwin et al. 1993). By our first birthday, these inductive inferences are guided by language; we expect that even superficially dissimilar objects will share their hidden properties if they are identified by the same common noun; if, for example, each is introduced as 'a blickett' (Graham et al. 2001). From the very beginning, we are inclined to generalize from experience with a given item to other items that we perceive as belonging to a common category.

In short, rapid, automatic generalization on the basis of categories is fundamental to how we make sense of the world.

This is true not only of object categories, but of social categories as well.<sup>16</sup> Following a long tradition that can be traced to Walter Lippmann (1922) and Gordon Allport (1954), it is a central tenet of contemporary social psychology that (as two recent encyclopedic discussions explain), stereotypes<sup>17</sup> which encode social categories "are fundamental to the ability to perceive, remember, plan, and act"

<sup>14</sup> For related psychological discussion see (Hamilton and Troiler 1986; Taylor 1981; Medin 1988; Oakes and Turner 1990; Schneider 2004).

<sup>15</sup> For interesting related work see (Huebner 2009).

<sup>16</sup> Cf. also the overview in Hamilton and Sherman (1994) and Devine (2009).

<sup>17</sup> The contemporary term *stereotype* was coined by Lippmann in (1922). He writes: "For the most part we do not first see, and then define; we define first and then see. In the great blooming, buzzing confusion of the outer world we pick out what our culture has already defined for us, and we tend to perceive that which we have picked out in the form stereotyped for us by our culture." (Lippmann 1922, p. 81). In keeping with the dual themes of the opening section of this essay, it's worth noting that Lippmann is also the person who introduced the term *Cold War* into popular discourse (Lippmann 1947), following its use in the context of US-Soviet relations by Bernard Baruch, an advisor to Harry Truman, in speech before Congress on April 14, 1947. (Baruch 1947).

(Banaji 2002, p. 15102)<sup>18</sup> in that they “serve to simplify perception, judgment and action... [by] spar[ing] perceivers the ordeal of responding to an almost incomprehensibly complex social world” (Macrae et al. 1994, p. 37). They are, like categories in non-social domains, ineliminable, fundamental, experience-shaping features of our perceptual and cognitive repertoire

In conjunction with more general facts about our cognition, this has three related consequences which will be central to the remainder of our discussion. I will discuss the first two briefly and the third in somewhat more detail. The first is that because categorization—including social categorization—is driven by a goal of simplicity, there is a tendency to view individuals within a category in ways that emphasize their similarities, and view individuals between categories in ways that emphasize their differences; that is, there is a tendency towards what psychologists call *intracategory assimilation*, and *intercategory contrast*.<sup>19</sup>

The second is that because information processing at all levels is influenced in part by top-down considerations, stereotypes play a role “from the earliest moments in the information processing sequence.” They do so by “giving preference to stereotype-consistent options” in ways that “influence the manner in which information is sought, perceived, remembered, and judged.” Stereotype-congruent information is attended to and encoded; stereotype-incongruent information is ignored and unassimilated (Banaji 2002, p. 15102). This is an instance of the more general phenomenon of confirmation bias, whereby both the search for and interpretation of information tends to be done in ways that favor hypotheses that the subject already holds. (For a nice overview, see Kunda 1999.) In the discussion below, we will look at a number of examples of this phenomenon.

The third is that as we become familiar with these ways of classifying objects in the world, our expectations and associations become automatized, so that the mere presence of a category-linked cue may be sufficient to activate a train of valenced

<sup>18</sup> The entry continues: “Functionally, they may be regarded as mental helpers that operate in the form of heuristics or short-cuts. Historically, this view of stereotypes as devices that allow a sensible reading of a complex world marked a breakthrough in research on stereotypes. Advanced by Allport (1954) and Tajfel (1969), the idea that stereotypes were inherent to the act of social categorization now forms the basis of the modern view. As such, stereotypes are regarded to be ordinary in nature, in the sense that they are the byproducts of basic processes of perception and categorization, learning, and memory. This cognitive view of stereotypes has dominated the field since the early 1980s” (Banaji 2002).

<sup>19</sup> In the seminal study exploring this effect, Tajfel and Wilkes (1963) asked participants to estimate category exemplars (lines) that varied continuously along a physical dimension (length). Participants provided their estimates in one of three conditions: Longer lines were systematically given a different label from shorter lines (with an arbitrary point along the continuum selected for the division), each line was randomly given one of the two labels, or no labels were presented. As predicted, the systematic association of categorical labels with exemplars that varied along a physical continuum increased the perceived differences between exemplars from different categories: Participants perceived a greater difference between the shortest of the long lines and the longest of the short lines in the systematic-categorization condition than they did in the other two conditions (thereby providing support for Tajfel and Wilkes’ hypothesis that some aspects of stereotyping have their origins in cognitive-perceptual processes that lead to an exaggeration of perceived differences between members of different social groups). In subsequent research in this tradition, Tajfel and others have demonstrated this effect in a wide range of domains (cf. Tajfel 1959; see also Ford and Stangor 1992; Krueger and Clement 1994; Macrae et al. 1993; Macrae et al. 1994; Stangor 2009; Tajfel 1981; Tajfel and Wilkes 1963).

and non-valenced associations.<sup>20</sup> Ultimately, encountering or thinking about a member of a well-learned category activates what I have called, in other work, an *alief*: an innate or habitual propensity to respond to an apparent stimulus, often with an automatized representational-affective-behavioral triad (more on this in the next subsection). Such aliefs are triggered whether or not they accord with our explicit beliefs—indeed, even when they run explicitly counter to them. And because they operate at a level that is relatively (though not completely) impenetrable by controlled rational processes, their regulation is best achieved by strategies that exploit capacities other than rational argument and persuasion.

## 2.2 Interlude: Alief

Since alief may be an unfamiliar notion to many of my readers, and since I see this paper as part of a larger project continuous with the work I began in Gendler (2008a, b), it may be useful to say a few words about the concept.<sup>21</sup> As I noted above, to have an *alief* is to have an innate or habitual propensity to respond to an apparent stimulus in a particular way. In paradigm cases and on strict usage, this response involves an automatized representational-affective-behavioral triad. I'll run through a specific case with this strict characterization, and then say something about how the notion generalizes.<sup>22</sup>

When I stand on a transparent glass walkway projecting over the Grand Canyon, this renders occurrent in me an alief whose content includes the visual appearance as of a cliff (representation), an alertness to danger (affect), and the readying (though not necessarily full activation) of a motor routine associated with retreat (behavior) (cf. Gendler 2008a). Depending on my other occurrent aliefs, beliefs and desires, this may have various consequences. The activation of a mindset associated with danger, for example, may direct my attention in certain ways: without deliberately intending to, I may repeatedly turn my eyes downward, encoding visual information about the distance between my feet and the raging Colorado River. Well-worn connections between my visual and motor systems may be activated, leading me to tremble. My experience of trembling may reinforce my feelings of discomfort, increasing my danger-alertness in cyclical fashion. Eventually, despite my uncompromised belief in the safety and stability of the structure on which I am standing, and my unwavering reflective desire to view the Canyon from the

---

<sup>20</sup> Cf. Patricia Devine: “[D]uring socialization, a culture’s beliefs about various social groups are frequently activated and become well-learned. As a result, these deep-rooted stereotypes and evaluative biases are automatically activated, without conscious awareness or intention, in the presence of members of stereotyped groups (or their symbolic equivalent)” (Devine and Sharp 2009, p. 62; cf. Devine 1989).

<sup>21</sup> Thanks to George Bealer for encouraging me to include more detail on this issue.

<sup>22</sup> I present the skywalk case making use of a strict characterization of alief, according to which the alief’s content is a representational-affective-behavioral triad. But the term fits more easily into ordinary talk if it is given a propositional complement. When the associative chain is widely shared, little information is lost thereby. So I might say, speaking loosely: I believe that the walkway is solid but alieve that it is precarious.

vantage-point that the walkway offers, I may find myself with a powerful—though reflectively disavowed—contrary inclination to remove myself from it.<sup>23</sup>

If I am strongly motivated by my reflective desire to remain on the walkway and my endorsed belief that the surface is safe, and if I have some (conscious or tacit) insight into the triggers of my alternate responses, I may (deliberately or automatically) engage in certain actions. I may direct my visual attention away from the glass surface beneath my feet (by closing my eyes or looking elsewhere). I may imagine the surface to be opaque. I may focus my mental energies on thoughts like “it’s only an illusion” or “lots of other people are standing here safely beside me” or “I paid a lot of money to do this.” I may ask you to block my way if I try to leave the walkway. Or I may give up, go inside, and buy a t-shirt at the gift shop.

What is important to the case is not what I end up doing, but how it illustrates an immensely common sort of situation. What happens in the skywalk case—and, I suggest, in numerous other cases at many other moments in our lives—is that our reflective desires and beliefs suggest one type of response or course of action, while our implicit associations and habitual patterns of response render occurrent another sort of response routine.<sup>24</sup>

Now, since a key source of my visually-triggered aliefs in the skywalk case is an evolutionarily-instilled (and highly advantageous) reluctance to place myself on an unsupported surface, I may be happy not to meddle (perhaps futilely) with the associative chain that underlies my responses. But in other cases—stereotypical racial associations being one of them—I may wish to replace some of my current associative chains with others. Likewise, since part of my motivation in stepping out on the skywalk might have been to induce precisely this sort of internal tension, I may be content to revel in the alief-belief discordance.<sup>25</sup> But in other cases—stereotypical racial associations being one of them—I may find this sort of discordance to be problematic.

<sup>23</sup> Literally on the day that this paper was going to press, *New York Times* reporter Edward Rothstein (author of the Grand Canyon Skywalk article cited at the beginning of “Alief and Belief”) published a piece about Chicago’s Skydeck Ledge, a transparent glass box suspended on the side of the 103-story Willis Tower, some 1,353 feet above the city’s busy streets. In perfect keeping with the alief-belief theme, Rothstein reports: “Despite the reassuring rivets in the 1,500-pound glass panels, the calm stillness of the air at the Windy City’s pinnacle and the security of a 10,000-pound weight capacity for each of the four 4.3-foot-deep glass boxes that protrude past the sheer edge of the Western Hemisphere’s tallest building—despite all that, you still feel twinges of queasiness... It is comforting to know that they were designed by the building’s original architects, Skidmore, Owings and Merrill, and mounted on tracks that allow them to be pulled inside for cleaning and maintenance. But any reassurance is undercut by the elimination of nearly all visible support; you walk within them, 1,353 feet above South Wacker Drive, surrounded by open air and unbounded space.” (The article continues with some interesting observations about why the effect of the Ledge is more powerful than that of the Canyon Skywalk; see Rothstein (2011)).

<sup>24</sup> For discussion of a wide range of such cases, see Gendler (2008a, b).

<sup>25</sup> Another domain where alief-belief discordance may be valuable is in the horror that most humans confront in undertaking necessary violence. Even if a person believes it is necessary to harm or kill for the greater good, we think they have lost something fundamental to their humanity if they lose or suppress the ‘killing-is-bad’ alief. (Thanks to Brendan Dill for discussion here.)

### 2.3 Dominant and aversive racism

In spelling this point out more explicitly, it will be useful to know something about the set of stereotypical associations connected with American blacks—a topic which has been researched intensively for nearly a century. Among the most recent large-scale studies was one conducted in 1995 at the University of Wisconsin, where participants were asked to specify the traits they thought were stereotypically associated with (American) blacks. The traits most frequently cited were: lazy, ignorant, musical, stupid, unreliable, loud, aggressive, athletic, rhythmic, low in intelligence, sexually perverse, uneducated, poor, hostile, and criminal (Devine and Elliott 1995, pp. 1144–1146).<sup>26</sup>

Importantly, these characteristics were identified as constituting the cultural stereotype even by those who rejected their accuracy, and knowledge of the stereotype—as measured by the subject’s ability and willingness to produce these terms when prompted to specify traits stereotypically associated with (American) Blacks—did not correlate in any way with measures of explicit prejudice. Study after study has shown that that “high- and low-prejudiced individuals d[o] not differ in their knowledge of the stereotype of Blacks but diverge sharply in their endorsement of the stereotype” (Devine and Elliott 1995, p. 1139).

Moreover, *awareness* of these stereotypes is sufficient to give rise to the relevant associative chains. The costs with which I will be concerned below are ones that arise simply from *having encoded* the stereotypes, whether or not the subject *endorses* them. As in the skywalk case, my reflective beliefs and desires may be unqualified: I desire to treat all races uniformly, and believe without qualification in their fundamental equality. But if I know that blacks are considered (by others, not by me) to be “lazy, ignorant, musical, stupid, unreliable, loud, aggressive, athletic, rhythmic, low in intelligence, sexually perverse, uneducated, poor, hostile, and criminal,” then—at least some of the time—those associations will be triggered by my thoughts about or interactions with members of those groups.<sup>27</sup> As in the skywalk case, my reflective desires and beliefs may suggest one type of response or course of action, while my implicit associations may render occurrent another sort of response routine.

This structure is well-known in the psychological literature: it is captured by the notion of *aversive racism*. The term can be traced to 1970, when psychiatrist Joel Kovel distinguished between what he called “dominant racism”—straightforward endorsement of racially-based discrimination and explicit commitment to race-based denigrations—and “aversive racism”—characteristic of those who “sympathize with the victims of past injustice, support the principle of racial equality, and regard themselves as non-prejudiced” but who, because of their explicit or implicit awareness of the negative traits stereotypically associated with members of the dominated racial group, exhibit behaviors indicative of “negative feelings and beliefs about blacks, which may be unconscious” (Kovel 1970, p. 54; Dovidio and Gaertner 2004).

<sup>26</sup> Devine and Elliott (1995); for 21st century discussion, see Amodio and Devine (2006).

<sup>27</sup> As Patricia Devine writes: “during socialization, a culture’s beliefs about various social groups are frequently activated and become well-learned. As a result, these deep-rooted stereotypes and evaluative biases are automatically activated, without conscious awareness or intention, in the presence of members of stereotyped groups (or their symbolic equivalent)” (Devine 2009, p. 62; cf. Devine 1989).

Aversive racism “is presumed to characterize the racial attitudes of a substantial portion of well-educated and liberal whites in the United States” (Dovidio and Gaertner 2004, p. 315), since it is the natural psychological outcome of a culture that embodies, on the one hand, an explicit ideal of racial equality, and, on the other, bears the legacy of centuries of personal and institutionalized racial discrimination.<sup>28</sup> It predicts that “for most individuals, overtly biased behavior will be rare, consistent with those individuals’ non-prejudiced self-concepts. More subtle expressions of bias, however, will be relatively common during intergroup interactions, as will discriminatory behavior that is attributionally ambiguous—that is, behavior that could be caused either by prejudice or by some other factor.”<sup>29</sup> (Dovidio et al. in press).

Even among those who are explicitly and sincerely committed to anti-racism, the legacy of having lived in a society structured by hierarchical and hostile racial divisions retains its imprint. So, for example, White participants primed with images of Black faces tend to be faster to identify an ambiguous image as a gun, and more likely to misidentify a (non-gun) tool as a gun (Payne 2001; Payne et al. 2002; cf. also Payne 2006). Likewise, participants playing a video game are quicker at deciding to shoot an unarmed black target than an unarmed white target, even when both targets are armed at equal rates in the context of the game (Correll et al. 2002). Otherwise identical resumés bearing stereotypical black names (e.g. Jamal, Lakisha) are less likely to result in interviews than resumés bearing stereotypical White names (Emily, Greg) (Bertrand and Mullainathan 2004.) State legislators are less likely to respond to requests for help with voter registration when the requests come from individuals with stereotypically black names (Butler and Broockman 2011). Black cab drivers receive lower tips than white cab drivers (Ayres et al. 2004–2005). And so on.

Indeed, even those who devote their lives to counteracting such stereotypical associations are not immune from them. Consider the following quotation:

There is nothing more painful to me at this stage in my life than to walk down the street and hear footsteps and start thinking about robbery. Then look around and see somebody White and feel relieved.

Its utterer is none other than Jesse Jackson (nd).

With this background in place, I now turn to the examples of the epistemic costs that this fact brings in its wake.

<sup>28</sup> Nor is this tension a new one. In 1944, Gunnar Myrdal wrote of “the ever-raging conflict between, on the one hand, the valuations preserved on the general plane which we call the ‘American creed,’ where the American thinks, talks, and acts under the influence of high national and Christian precepts and, on the other hand, the valuations on the specific planes of individual and group living, where personal and local interests; economic, social, and sexual jealousies; consideration of community prestige and conformity; group prejudice against particular persons or types of people; and all sorts of miscellaneous wants, impulses, and habits dominate his outlook. (Myrdal 1944, p. xliii, as cited in Gaertner and Dovidio 2000).

<sup>29</sup> Cf. also Devine and Elliot: “The coexistence of a rejected yet enduring negative stereotype of Blacks and a positive set of beliefs about Blacks places the low-prejudiced social perceiver in a precarious position. A considerable amount of research indicates that stereotypes can be automatically activated by the perception of social stimuli (e.g. exemplars of the group, group labels), resulting in prejudice-like feelings, thoughts and behaviors for high- and low-prejudiced individuals alike (Devine 1989; Klinger and Beall 1992).” (Devine and Elliott 1995, p. 1147).

### 3 Difficulties in cross-race face identification

Let's start with the first epistemic cost of living in a society structured by race, a phenomenon that is generally called the *cross-race recognition deficit* (CR deficit). Across hundreds of studies in dozens of cultures, psychologists have shown a tendency for participants to exhibit superior recognitional capacities for own-race as compared to other-race faces (cf. Meissner and Brigham 2001).

Typically, such studies take something like the following form. Participants sit in front of a computer screen where they are presented with a series of images of faces of various races—say, 12 black faces and 12 white faces, interspersed in random order—that they observe sequentially for a certain number of seconds each. A short time later, they are shown an additional series of faces on a screen, some of which were and some of which weren't faces that appeared in the series they initially observed. Their task is to indicate, for each of the faces in the second series, whether or not it is a face that they had previously been shown. There are thus two kinds of correct identifications—correctly identifying a previously-seen face as previously-seen (a “hit”), and correctly identifying a novel face as novel—and two kinds of mistakes—mistakenly identifying previously-seen as novel (a “miss”) or mistakenly identifying a novel face as previously-seen (a “false alarm”).

Cross-race recognition deficit is the phenomenon whereby faces of the out-group. (in this case, faces of the race other than the subject's own) are less consistently recognized (misses) or remembered (false alarms) than faces of the in-group (in this case, faces of the subject's own race.) In a comprehensive 2001 review article, data from more than 5000 participants across nearly 40 studies consistently revealed a “mirror effect” pattern whereby in-group faces yielded both “a higher percentage of hits and a lower percentage of false alarms compared with other-race faces” (Meissner and Brigham 2001, p. 1).<sup>30</sup>

Nor is the effect limited to laboratory settings. It occurs in informal and formal individual and group interpersonal interactions, and in low-stakes and high-stakes identification situations, including eyewitness testimony (Wells and Olson 2003). Both in structured experimental settings, and in unstructured field settings, participants are worse at recognizing outgroup than ingroup faces: they are worse at telling that they've seen a particular face before when they have, and worse at telling that they haven't seen a particular face before when they haven't.

Two general sorts of mechanisms seem to underlie this phenomenon. Some of the variance appears to result from differences in *perceptual expertise*. Since the “ability to extract information from an environment improves with experience,” it is perhaps unsurprising that “[i]ndividuals are more accurate at recognizing types of faces with which they have had more exposure” (Hehman et al. 2010, p. 445, citing Sporer 2001). In a culture in which interpersonal encounters are at least

<sup>30</sup> Similar patterns were observed in a 1986 meta-analysis of 128 eyewitness identification and facial recognition studies, involving 960 experimental conditions and 16,950 participants (Shapiro and Penrod 1986).

partially structured by racial-group membership, and where racial categorization partially tracks independently specifiable visual features,<sup>31</sup> participants gain exposure to and consequent expertise with faces of their own racial group, but not with faces of other groups. As a result, their ability to extract information about faces from their own racial group on the basis of short exposure increases, whereas their ability to extract information about faces from other racial groups does not develop. Evidence from cross-cultural studies of infancy and early childhood suggest that perceptual expertise likely plays some role in explaining cross-race facial effect.<sup>32</sup>

A second source of variance appears to result from *social-cognitive* factors, that is, from “the different ways people process information as a function of categorizing others as ingroup or outgroup members,<sup>33</sup>” over and above (or independently of) issues of exposure and expertise. It appears, for example, that even among same-race faces, participants are generally more effective at encoding information about faces that are marked as belonging to their ingroup (for example, as attending their university, or as liking the same sports teams they do) and less effective at encoding information about faces marked as belonging to their outgroup (for example, as attending a different university, or as liking a different team). Perhaps even more strikingly, there is evidence that providing participants with a superordinate ingroup/outgroup classification that cuts across racial lines results in a facial recognition deficit that tracks the constructed in-group/out-group boundary, rather than the long-learned racial categories (Kurzban et al. 2001; cf. Ackerman et al. 2006; Cosmides et al. 2003; cf. Gaertner and Dovidio 2000).

The sources of CR deficit are undoubtedly complicated and manifold. Remembering that our concern in this article is with the epistemic costs of automatically-activated reflectively-disavowed racial aliefs, I’ll focus discussion on one social-cognitive explanation which is of particular interest in this context. According to the *asymmetric feature selection* hypothesis, at least part of the explanation is the following: when participants encounter other-race faces, one of the visual features they typically encode is information about race, whereas no such information is typically encoded for same-race faces. So, for example, when a white subject sees a novel black face, in addition to coding information like “eyes here” “nose there” “ears there,” she also uses some of her limited cognitive resources (cf. Miller 1956) to encode “black;” whereas, typically, when she sees a novel white face, the category of race is not encoded as such: this leaves her with more cognitive space to encode an additional fact about the same-race face—say, “eyebrows there.” Since other-race faces are processed as racially marked, cognitive resources that would otherwise be available for encoding specific information about the face are deployed to encode coarse-grained information about category-membership; since same-race faces are processed as racially unmarked, the subject’s limited

<sup>31</sup> This is an important caveat, and one that brings out the ways in which I am ignoring in my discussions a large and important literature on racial categorization itself.

<sup>32</sup> Cf. (Kelly et al. 2007; also Bar-Haim et al. 2006; for an overview of this literature, see Hehman et al. 2010).

<sup>33</sup> Cf. (Bernstein et al. 2007; Sporer 2001; Hehman et al. 2010, p. 445).

cognitive resources can be deployed to encode more fine-grained information.<sup>34</sup> (As an analogy, think about how we encode local versus long-distance phone numbers: in the case of the former, we need only the seven digits, since the area code is encoded as “like me;” in the case of the latter, we need not only the seven specific digits, but the area code as well.<sup>35</sup>)

One source of evidence in favor of this hypothesis comes from visual search tasks. When typical white participants are presented with an array of luminance-corrected faces arranged in a large row-and-column matrix, they tend to be much faster at finding a single black face in a sea of white faces than they are at finding a single white face in a sea of black faces. This appears to be an instance of a more general phenomenon in visual search. It is comparatively easy to find an R hidden in a matrix of Ps, or a Q hidden in a matrix of Os; it is comparatively difficult to find a P hidden in a matrix of Rs or an O hidden in a matrix of Qs. More generally, in situations that require locating a “feature-present” entity in an array of “feature-absent” ones, it is an easy task for the visual system to pick out the discrepant object; in situations that require locating a “feature-absent” entity in an array of “feature-present” ones, the visual system finds the task difficult. The asymmetry observed in white participants’ facial-search speed in the black/white matrix task suggests such a feature-present/feature-absent explanation.

So what is happening in cases of CR deficit may be something like this: when white participants are briefly exposed to a face of another race, they typically use up one of their information slots encoding coarse-grained information about its race; as a result, the amount of individuating information available to them about other facial details is smaller, and there is a consequent reduction in their recognitional accuracy. A related explanation runs through directed attention: people tend to direct their attention to individuating features of ingroup members, and to classificatory features of outgroup members.<sup>36</sup> Either makes the cross-race recognition deficit an epistemically interesting phenomenon. It occurs because, typically, the information that people select in cross-race faces is optimal for classification but not for recognition, whereas the information that people encode for same-race faces is optimal for individual recognition. Subsequent research in this tradition has confirmed this basic hypothesis (e.g. Hugenberg et al. 2007; Lebrecht et al. 2009; Tanaka and Peirce 2009).<sup>37</sup>

There are strategies that can be used to overcome this effect (See, e.g., Lebrecht et al. 2009; Hills and Lewis 2006; McGugin et al. 2011; Tanaka and Pierce 2009). But because racial categorization is one of the most commonly and easily activated patterns of classification available in a society structured by race, in ordinary day-to-day interactions where deliberate strategies for alternative attentional patterns are

<sup>34</sup> To my knowledge, there have been no empirical studies in this paradigm looking at black responses to white faces. One might expect parallel results (*mutatis mutandis*). Or one might expect that because white identity is encoded as the unmarked case in the dominant culture, the effect would be less pronounced.

<sup>35</sup> Thanks to Christopher Peacocke for the suggestion that prompted this analogy.

<sup>36</sup> Thanks to Brendan Dill for pointing out the importance of making this distinction, and for suggesting the second formulation.

<sup>37</sup> For an interesting discussion of the interaction of race and class in CR deficit, see Shriver et al. (2008).

not in play, even people whose normative commitments are anti-racist may find themselves differentially encoding information about same-race and other-race faces. Because this is a real-time process, and because race is a salient category in many interpersonal interactions, race-associated schemata tend to be activated automatically. As a result, it will typically be the case that coarse-grained information useful for classification will use up some of the cognitive resources that would otherwise be available for fine-grained information useful for recognition.

This is our first example of a case where the existence of race as a category that gives rise to certain sorts of automatic associations is epistemically hazardous, even for those who disavow the normative content of some of those associations. On the assumption that at least in many cases, the facial information that would be useful to encode is individuating information, the automatically-triggered tendency to encode other-race faces in ways optimal for classification rather than individuation directs attentional resources in ways that leave participants epistemically worse off.

As we will see in the next section, stereotype threat provides a second example with this sort of structure: another case where attentional resources are directed in epistemically costly ways.

#### 4 Stereotype threat

Stereotype threat is a well-documented phenomenon whereby activating an individual's thoughts about her membership in a group that is associated with impaired performance in a particular domain increases her tendency to perform in a stereotype-confirming manner.<sup>38</sup> So, for example, as Claude Steele and Joshua Aronson hypothesized in their original 1995 paper, “whenever African American students perform an explicitly scholastic or intellectual task, they face the threat of confirming or being judged by a negative social stereotype—a suspicion—about their group’s intellectual ability and competence...The self-threat [this] ...may interfere with the intellectual functioning of these students, particularly during standardized tests” (Steele and Aronson 1995, p. 797).

In Steele and Aronson's original study, black and white college students were given a half-hour test consisting of a selection of challenging items from the verbal section of the Graduate Record Exam (GRE) under one of two conditions. In the *stereotype-threat condition*, the test was described as “diagnostic of intellectual ability;” in the *non-stereotype-threat condition*, “the same test was described simply as a laboratory problem-solving task that was nondiagnostic of ability.” (Steele and Aronson 1995, p. 799.) The differences in results between the two conditions were striking: When differences in SAT scores were controlled for, “Black participants performed worse than White participants when the test was presented as a measure of their ability, but improved dramatically, matching the performance of Whites, when the test was presented as” nondiagnostic (Steele and Aronson 1995, p. 801).

<sup>38</sup> For an accessible overview of recent work in this area, see Stroessner et al. (n.d.). A highly readable introduction to the phenomenon by the person who coined the term is Steele (2010).

Since then, the effect has been shown in hundreds of studies in dozens of domains. So, for example.

When a task is described as diagnostic of intelligence, Latinos and particularly Latinas perform more poorly than do Whites (Gonzalez et al. 2002), children with low socioeconomic status perform more poorly than do those with high socioeconomic status (Croizet and Claire 1998), and psychology students perform more poorly than do science students (Croizet et al. 2004). Even groups who typically enjoy advantaged social status can be made to experience stereotype threat... White men perform more poorly on a math test when they are told that their performance will be compared with that of Asian men (Aronson et al. 1999), and Whites perform more poorly than Blacks on a motor task when it is described to them as measuring their natural athletic ability (Stone 2002; Stone et al. 1999). (Schmader et al. 2008, pp. 336–337).

The phenomenon has been demonstrated with verbal tasks (Steele and Aronson 1995), complex mathematical tasks (Quinn and Spencer 2001), tests of memory (Hess et al. 2003), and mental rotation tasks (Wraga et al. 2006.) It has been demonstrated with social tasks “such as maintaining a fluid interaction with someone in the face of negative stereotypes suggesting malicious intentions in that interaction” (Bosson et al. 2004; Richeson and Shelton 2003), and in tasks involving “sensorimotor skills or other tasks that entail fluid movement or automated behavioral processes” (e.g., Beilock et al. 2006; Stone et al. 1999; all citations in this paragraph drawn from Schmader et al. 2008, p. 340).<sup>39</sup>

In one of the most striking demonstrations of the phenomenon, young girls of Asian-American descent who ranged in age from kindergarten to 8th grade were given tasks that rendered salient either their female identity, their Asian identity, or neither identity (control). Subsequently, they were given a series of items from a grade-appropriate standardized math test. Girls from lower-elementary and middle school grades showed a striking pattern of results: those whose Asian identity had been emphasized showed an improvement in scores when compared with controls, whereas those whose female identity had been emphasized showed a decrement<sup>40,41</sup> (Ambady et al. 2001; cf. Shih et al. 2006).

What implications does all of this have philosophically? Stereotype threat appears to interfere with knowledge in at least two ways.<sup>42</sup> Participants may

<sup>39</sup> The magnitude of these effects can be quite powerful: in some cases, participants in stereotype threat conditions answer only half as many questions correctly as those in non-threat conditions (Steele and Aronson 1995).

<sup>40</sup> Interestingly, the effect seems to be most profound for challenging rather than simple problems. Neville and Croizet (2007) found, for example, that activating gender identity *increased* girls' performance on easy math problems, but that this advantage was outweighed by a large *decrease* in their performance on more difficult problems.

<sup>41</sup> For a stereotype-threat-avoidance strategy that exploits this phenomenon, see Rydell et al. (2009).

<sup>42</sup> As with any complex social phenomenon, there are undoubtedly a range of factors that contribute to the effect. Theories to explain the phenomenon (or, more likely, phenomena) are manifold, and a full survey would require a full paper. For a representative example of such a theory, cf. Schmader et al. (2008), who hypothesize that “stereotype threat disrupts performance via 3 distinct, yet interrelated, mechanisms: (a) a physiological stress response that directly impairs prefrontal processing, (b) a tendency

temporarily lose *access* to the contents of certain of their true beliefs: a subject suffering from stereotype threat may find herself unable to recall whether Chester Arthur was the 20th or the 21st President, or which one is xylem and which one is phloem—despite finding such questions easy in contexts outside of the threat-inducing situation. And participants may temporarily lose *confidence* in their true beliefs: our subject may find herself double- and triple-checking that  $11 \times 11$  is indeed 121, or running through an entire conjugation to make sure that ‘sunt’ is the third person plural form of ‘esse’—despite finding such questions trivial in other contexts.<sup>43</sup>

This means that stereotype threat should be at least *prima facie* interesting to epistemologists: it’s an easily-specifiable type of circumstance that appears to have predictable effects on participants’ knowledge. But is it deeply interesting? I think it may be, though the ideas in the next paragraphs are admittedly somewhat fuzzy.

Let’s start by thinking about cases where you lose access to the contents of or confidence in your true beliefs due to simple physical *causes*. After several gin-and-tonics, you may find yourself unable to recall whether Chester Arthur was the 20th or the 21st President<sup>44</sup>; if you’re under anesthesia, you probably won’t be in a position to say which one is xylem and which one is phloem. If you bump your head, you may find yourself triple-checking your arithmetical calculations; and if I give you a Latin quiz when we are 15-miles into a grueling marathon, you may find yourself running through the full conjugation to check whether ‘esse’s’ third-person plural is ‘sunt’. Such cases are important, of course: they remind us of a number of facts that are boring, but significant. For example, that crediting people with knowledge requires various sorts of idealizations, and that facts about our bodies cause predictable consequences for our minds. But so far this is old news.

It’s also old (but important) news that there are cases where you may lose access to or confidence in your true beliefs as a result of *reasons*. If I read and am convinced by a revisionist history book or botany text, and come to believe that everything I had learned about the American presidency and the biology of trees is demonstrably false, I may replace my beliefs about Chester Arthur and xylem with new beliefs that leave my previous ones relevantly inaccessible. Likewise, if I think through the Cartesian dreaming argument, I may lose confidence in my (presumably) true belief that I have two hands; when I consider various lottery cases, I may lose confidence in my (presumably) true belief that my car is parked outside in my driveway. Again, this is all old news.

Now what *may be* interesting, epistemically and more generally about the stereotype threat cases is that they appear to fall somewhere in between the reason cases and the cause cases. I repeat: this is not well-worked out and I am

---

Footnote 42 continued

to actively monitor performance, and (c) efforts to suppress negative thoughts and emotions in the service of self-regulation. These mechanisms combine to consume executive resources needed to perform well on cognitive and social tasks. The active monitoring mechanism disrupts performance on sensorimotor tasks directly.” (Schmader et al. 2008, p. 336).

<sup>43</sup> I don’t know of empirical work that tries to tease apart these two possibilities.

<sup>44</sup> Cf. YouTube’s “Drunk History” videos, e.g. <http://www.youtube.com/watch?v=ipV2u-MxIFc> (see Drunk History n.d.) (Thanks to Brendan Dill for this pointer).

counting on my readers to help me in thinking this through in later papers. In the stereotype threat cases, the loss of knowledge is due at least in part to a feeling of anxiety that is induced through activations of self-referential cultural stereotype aliefs whose content you disavow. This loss of knowledge isn't just the result of something straightforwardly causal like bumping your head and getting a concussion—or even just something like being hungry, or tired, or preoccupied; nor is it the result of something straightforwardly reason-based like reading a revisionist textbook or thinking through a Brain-in-a-Vat scenario. That is, it's not due to pushes and pulls and bumps and bounces—nor is it due to beliefs and arguments and reflections and persuasions. Rather, it's the result of something that—for want of a better term—I'll provisionally call an *eason*: something that is not sufficiently well-conceptualized to call a reason, but that (in a way in between a reasony and a causy fashion) *eases* us towards a certain outlook on the world.<sup>45</sup>

Roughly speaking, on this picture, beliefs are to reasons as aliefs are to easons: we justify beliefs by appeal to reasons; we explain aliefs by appeal to easons. Our imagined victim of stereotype threat has an *alief* with the content “‘Female’ applies to me and ‘female’ is associated with poor math performance; (anxiously) better make sure that I’m doing these math problems correctly; double-check double-check double-check.” The *eason* that she has this alief is similarly complex. It consists in some sort of complex interplay among stress, anxiety-induced self-monitoring and self-regulative emotion suppression. This case is distinct from one where our subject *believes* “perhaps the answer to the third question is not in fact 17” or “I, like most females, am bad at math.” If asked explicitly, she would deny both, and offer reasons for each that she reflectively endorses: “Of course the sum of 1, 2, 3, 4 and 7 is 17:  $1 + 2 + 3 + 4 = 10$ , and  $10 + 7 = 17$ ” or “Of course I’m good at math: when I’m at home after school, I love to play math games on my computer.” Nor is it simply a case where she is merely *caused* out of holding that  $1 + 2 + 3 + 4 + 7 = 17$ : she hasn’t been hit in the head by an errant baseball, or had arithmetic-scrambling electrodes implanted in her brain, or forgotten to eat breakfast, the most important meal of the day (WebMD 2010). Rather, her alief is the eason for her knowledge-loss.

To repeat, this is sketchy and in need of sharpening. But it’s as far as I can get without my readers’ help, which I hereby solicit.

To return to the paper’s main thread: even without the issues just raised, stereotype threat is still important and interesting. It provides us with our second example of an epistemic cost resulting from the existence of race (and gender) as categories that give rise to certain sorts of automatic associations, even for those who disavow their normative content. Raising to salience a negative stereotype about a group with which one self-identifies—even if one explicitly believes those negative associations to be false—impairs cognitive performance.

<sup>45</sup> Relatedly, extant hypotheses that give rise to confirmation biases can be seen as operating via easons rather than reasons. (Thanks to Brendan Dill for discussion on this point.)

## 5 Cognitive depletion following interracial interaction

Interesting though they are, in at least one sense the issues raised in the last section are not so surprising: it's not particularly newsworthy that racism and sexism are harmful to the targets of discrimination (though it may be somewhat surprising exactly what form that harm takes.) What may be less expected is the epistemic costs that racism carries even for some of those who are not its target. It is to this issue that I turn in the section below.<sup>46</sup>

For roughly the last two decades, a large body of psychological research has made use of a research paradigm known as the Implicit Association Test (IAT)—a test which measures what I would call occurrent aliefs. In its canonical form, the test involves asking a subject to engage in a categorization task intended to measure—for two categories A and B—whether the subject finds it more natural to associate As with Bs, or with  $\sim$ Bs. More specifically, participants are presented with a series of words and images on a computer screen in a series of trials that embody one of two different conditions. In trials governed by the first condition, participants are presented with words and images which they are asked to classify as belonging either to the category A-or-B or to the category  $\sim$ A-or- $\sim$ B. In trials governed by the second condition, participants are asked to classify the words and images as belonging either to the category A-or- $\sim$ B or  $\sim$ A-or-B. If they are faster in trials governed by the first condition, then it is concluded that they find it more natural to associate A with B (as opposed to with  $\sim$ B); if they are faster in the second, it is concluded that they find it more natural to associate A with  $\sim$ B (as opposed to with B).

So, for example, participants might be presented with sequence of White (A) and Black ( $\sim$ A) faces, interspersed with a sequence of positive (B) and negative ( $\sim$ B) words (e.g. “happy” and “harmful”), and instructed to classify the words and images as quickly as possible either into the categories White-or-positive (A-or-B) and Black-or-negative ( $\sim$ A-or- $\sim$ B), or—alternatively—into the categories White-or-negative (A-or- $\sim$ B) or Black-or-positive ( $\sim$ A-or-B).<sup>47</sup> Hundreds of studies conducted in dozens of laboratories over nearly two decades have shown that—for most twentieth and twenty-first-century American participants in most circumstances—the first of these pairings is more natural than the second.<sup>48</sup> That is, these participants on the whole are faster to make classifications into the categories White-or-positive and Black-or-negative than to their converses. This suggests that the former categories are, for most participants, more readily constructed or more easily accessed, and hence experienced as more “natural” than the latter (or, in my terminology, that most participants have

<sup>46</sup> Some of the discussion in this section follows that in the closing pages of Gendler (2008b).

<sup>47</sup> To try a version of the test, see IAT Corp (1998) or Pious (2002). For discussion, see Nosek et al. (2006).

<sup>48</sup> Analogous studies have been conducted in other societies, with parallel results. Cf. Dunham et al. (2006, 2007).

highly-accessible aliefs that encode these contents).<sup>49</sup> That said, there is a good deal of individual variation in IAT results, and this will turn out to be important in the discussion that follows.

In a series of studies conducted by the psychologist Jennifer Richeson (cf. Richeson and Shelton 2007; Richeson et al. 2005; Richeson and Shelton 2003; Trawalter and Richeson 2006; Trawalter and Richeson 2008; Trawalter et al. 2009), white college-aged participants who had previously taken a Black/White IAT were asked to interact with either a same-race (White) or different-race (Black) peer confederate, who was presented to them as being the student manager of the laboratory where the study was being conducted. After the interaction, participants were asked to undertake an ostensibly unrelated task—the Stroop color-naming task<sup>50</sup>—which is standardly used as a measure of executive control and cognitive depletion: the faster participants are at calling out the colors in which the words are printed (which requires suppressing the initial tendency to instead utter the name of the color-word which is printed on the page), the higher their level of current executive control; the slower they are, the more cognitively depleted they are assumed to be.

Richeson reports her findings as follows:

Consistent with the prediction that interracial contact stress will undermine subsequent executive control, White individuals, on average, performed more poorly on the Stroop task after contact with a Black experimenter than they did after contact with a White experimenter.

That is, White participants on the whole performed worse on the Stroop task following interaction with a Black peer than with a White peer—even though the interactions themselves were, in both the White–White and the White–Black case—unerringly polite. But, in addition to the main effect of cross-race vs. same-race interaction, there were also individual differences that correlated with IAT scores. Richeson continues:

Furthermore, the greater the...relative ease with which [these participants] associate[d]...negative words with...Black American racial categories [e.g: IAT scores]... the poorer their Stroop performance after interracial interactions. (Richeson and Shelton 2007, pp. 316–317)

She concludes:

[T]his...suggests that, like other stressors, interracial interactions can be cognitively costly (Richeson and Shelton 2007, p. 317)

Subsequent neuroimaging work (Richeson et al. 2003) suggests that the difference between the groups can be traced to differential activation of areas in

<sup>49</sup> Given the argument I am making, it does not matter whether the implicit associations revealed by the IAT measure the subject's ease of access to associations that she implicitly (though perhaps not explicitly) endorses or whether it measures her ease of access to culturally-encoded associations that she both implicitly and explicitly rejects. For discussion of this controversy, see (Karpinski and Hilton 2001; Olson and Fazio 2004; Arkes and Tetlock 2004).

<sup>50</sup> On the unlikely chance that you are not familiar with this effect, see Chudler (n.d.). A nice summary of the task can be found in the opening paragraphs of Richeson and Shelton (2007).

Is it just that greater level of IB → greater cognitive depletion, or is it greater IIB -EBI → greater cognitive depletion?

prefrontal cortex associated with executive function and self-regulation. Together, the neuroimaging and behavioral evidence suggest that participants whose occurrent aliefs—as manifest through their IAT results—were out of line with their conscious goal of acting in a non-discriminatory fashion expended significant cognitive effort to suppress the response-tendencies activated through these associations. So, as with the stereotype threat case, people whose aliefs and beliefs are out of line pay cognitive penalties.<sup>51</sup>

This is our third example of an epistemic cost resulting from the existence of race as a category that gives rise to certain sorts of automatic associations, even for those who disavow their normative content. Raising to salience (though interaction) a negative stereotype about a group with which one does *not* self-identify—even if one explicitly believes those negative associations to be false—impairs cognitive performance if one is led to expend cognitive effort keeping those associations at bay.

\*even\* for those or especially/ only for those?

Perhaps this is more surprising and newsworthy: it turns out that living in a culture structured by a legacy of racism may be epistemically costly exactly for those who understand but wish to resist that legacy.

## 6 Base rate neglect

One obvious way to resist the legacy—and thereby to avoid the three types of epistemic costs detailed above—would be to fail to encode the base rate information and cultural associations that give rise to these problematic aliefs. And, indeed, recent psychological and legal literature has looked at two related domains where those committed to a certain sort of anti-racism exhibit a tendency to ignore base rates. The first is a phenomenon that Philip Tetlock has dubbed “the psychology of the unthinkable” (Tetlock et al. 2000); the second is the phenomenon of resistance to racial profiling. I had initially hoped to include a discussion of racial profiling in this paper,<sup>52</sup> but the essay is already too long as it stands. So I will focus my attention here only on the first of these two cases.

In an influential paper entitled “The Psychology of the Unthinkable,” Philip Tetlock and his colleagues discuss three sorts of cases in which participants appear to engage in cognitive self-censorship, stifling processes of reasoning that would otherwise allow them to achieve certain epistemic goals.<sup>53</sup> The first are *taboo trade-offs*: scenarios in which participants resist considering how they would resolve certain dilemmas because they involve presenting “‘sacred values’ like honor, love, justice and life” as “fungible” in the sense that they can be traded for particular quantities of secular goods like money (Tetlock et al. 2000, p. 854). (Think of Kant on price vs. dignity.) The second are *heretical counterfactuals*: “assertions about

<sup>51</sup> For connection of this to historical discussions of the “harmonious soul” see Gendler (2008b).

<sup>52</sup> For representative philosophical discussion of these issues, see (Risse and Zeckhauser 2004; Lever 2005).

<sup>53</sup> For Tetlock’s own diagnosis of these cases, see (Tetlock et al. 2000, p. 853; also Arkes and Tetlock 2004).

historical causality (framed as subjunctive conditionals with false antecedents) that pass conventional cognitive tests of plausibility but that many people greet with indignation because the assertions subvert a core tenet of their religious belief systems” (Tetlock et al. 2000, pp. 854–855<sup>54</sup>). The third are *forbidden base rates*: statistical generalizations “that devoted Bayesians would not hesitate to enter into their probability calculations but that deeply offend a religious or political community” (Tetlock et al. 2000, p. 854). In all three of these cases, Tetlock found that a majority of participants exhibit a tendency to self-impose “normative proscriptions...on cognitive processes that are fundamental to rationality in the intuitive scientist and economist traditions,” that is, on cognitive processes that would typically be described as rational (Tetlock et al. 2000, p. 854). Most relevant to our purposes here is the third—*forbidden base rates*.

In a series of studies, Tetlock and his colleagues presented participants with cases in which decision-makers were confronted with base rate information that either “did or did not turn out to be correlated with the racial composition of neighborhoods” in order to test what he called “the symbolic antiracism hypothesis,” namely, “that people would regard actuarial risk as a legitimate rationale for price discrimination in setting insurance premiums only when the correlation between actuarial risk and racial mix of neighborhoods is not mentioned. When the correlation is highlighted, people—especially liberals—will vehemently reject race-tainted base rates and invoke multiple grounds for rejecting them” (Tetlock et al. 2000, p. 860). The hypothesis was borne out by the data: for most participants (and especially for self-identified liberals) base-rates that they considered “permissible” for consideration in a race-neutral context became “off-limits” when “the linkage with race was revealed” (Tetlock et al. 2000, p. 863). That is: participants engaged in a kind of epistemic self-censorship on non-epistemic grounds.

The phenomenon of *Forbidden Base Rates* highlights some of the ways in which it is costly to adopt a particular sort of anti-racism in a racially stratified society. It is costly in a narrow economic sense because it causes participants to discount information that, if taken into consideration, would increase their narrowly-construed financial well-being; and it is costly in an epistemic sense because it causes participants to discount information that might be relevant to their full consideration of both background and foreground conditions.

Of course, the narrow financial costs to sellers of insurance may be considered non-problematic, even desirable, if one’s concern is equitable spread of risk across parties in a society structured by racial inequality: after all, it is presumably the benefits of white privilege that make it the case that there is a correlation between

---

<sup>54</sup> Heretical counterfactuals are essentially cases that invoke what I call “imaginative resistance” (Gendler 2000, 2006). So, for example, Tetlock et al. presented participants with cases such as the following: “Consider the argument If Joseph had not believed the message that Mary had conceived a child through the Holy Ghost and that there was no reason to fear taking Mary as his wife, then Jesus would have grown up without the influence of a father and would have formed a very different personality. Is it easy or difficult to accept the premise that Joseph could have decided not to believe the angel’s message?” (Tetlock et al. 2000, p. 864.) I offer an alief-involving explanation of imaginative resistance in Gendler (2000, 2006, 2010).

actuarial risk and neighborhood racial composition, and in light of that, it may well be appropriate to take this information into consideration in setting insurance costs. People who endorse this analysis might be in a position to adopt an epistemic strategy in which the enlarged perspective allows them to take into consideration—though subsequently discount—the base-rate data that Tetlock’s participants reported as “off-limits.” But within the larger context of the set of cases that Tetlock describes, it is clear that the “apposite functionalist metaphor” is one that “depicts people engaged in a continual struggle to protect their private selves and public identities from moral contamination by impure thoughts and deeds” (Tetlock et al. 2000, p. 853).

Now, a person who has accurate statistical knowledge of demographic variation will, by definition, know about racial differences in crime rates. Whether or not one is aware of the precise statistics, to know nothing of these data would require one to cultivate ignorance about a striking feature of contemporary American society. “In the mid-1990s,” for example, “23 of the 80 largest cities in the United States...had black homicide arrest rates that were more than 10 times higher than white rates” (LaFree et al. 2010, p. 94).<sup>55</sup> Even if, as is surely the case, arrest rates are proportionally higher for crimes committed by historically underrepresented racial groups, there is general consensus even among those most critical of contemporary policing practices that the actual rates of commission differ across races. (It goes without saying that the explanation for these differences lies in the nation’s legacy of racial injustice (which is part of what makes racial categories unavoidable.) It is likewise important to note that blacks are also disproportionately represented among the *victims* of such crimes.)

This base-rate information has consequences for both reflective and non-reflective behavior, regardless of the degree to which one dislikes the correlation. (Think about the Jesse Jackson story above.) Explicitly, it should lead a person to update her prior probability; rationally, she should come to believe with respect to certain racial groups that the likelihood of a member of that racial group committing a certain sort of crime is higher than for a member of some other racial group. And if one’s goal is the reduction of crime, this distributional information has—as Arkes and Tetlock (2004), following Farmer and Terrell (2001) contend—rather striking practical implications:

As long as differential crime rates exist across groups in society, minimizing the overall crime rate will result in far more convictions of innocent members of the minority group even if racism is not at work. It follows mathematically (within a wide range of plausible assumptions) that by requiring less evidence to convict members of a smaller but higher crime group, one will simultaneously lower the overall crime rate and increase the overall probability of convicting an innocent person. This troubling state of affairs must be considered in light of the opposite option: to rectify racial inequality in the probability of erroneous convictions, society must tolerate a higher crime rate, whose victims will predominantly come from the minority group.

---

<sup>55</sup> LaFree et al. (2010) also includes a discussion of effectiveness of various ameliorative strategies.

Indeed, Farmer and Terrell (2001) have estimated that approximately 1,900 more murders per year will occur if racial inequality is removed from the erroneous conviction rates.” (Arkes and Tetlock 2004, p. 272)

This extremely disturbing mathematical fact brings out a perverse feature of the logic of discrimination, namely, its quasi-rational self-perpetuation. Though this is an important topic, it is peripheral to the argument of this particular paper. What is relevant for our purposes are the psychological implications of this correlation.

As I have argued above, given the salience of race in American culture (and hence the likelihood that racial aliefs will be easily rendered occurrent in daily interactions), the encoding of certain sorts of associations with certain racial groups carries the sorts of costs I illustrated in the previous three sections. Because of how human minds work, these associative chains—whether endorsed or denied—will be triggered in a wide range of circumstances. This means that, given facts about race and crime, if a person encodes this information and then takes the IAT, she will be more likely to associate crime-related behavior with African American men than with white men. And if so, she will be subject to the epistemic costs enumerated above.

In short, as long as there’s a differential crime rate between racial groups, a perfectly rational decision maker will manifest different behaviors, explicit and implicit, towards members of different races. This is a profound cost: living in a society structured by race appears to make it impossible to be both rational and equitable.

## 7 The sad conclusion

The existence of race as a category that gives rise to certain sorts of automatic associations is hazardous, even for those who disavow the normative content of those associations. It affects your ability to encode individuating information about faces of persons that you apprehend as belonging to a different racial group—even if you explicitly avow racial equality. It impairs cognitive performance when a negative stereotype about a group with which you self-identify is brought to salience—even if you explicitly believe those negative associations to be false. It leads to cognitive exhaustion following interracial interaction when there are discrepancies between your explicit commitments and your implicit associations—even if your explicit commitments are wholly sincere. And it makes encoding information about racial inequity itself problematic—you are faced with a choice between explicit irrationality through base-rate neglect or implicit irrationality through encoding associations that you reflectively reject.

Racially-based inequities—and the psychological processes by which we inevitably encode them—carry not merely moral, but also epistemic, costs. And they carry them regardless of what we believe.

**Acknowledgements** For comments and conversation regarding early drafts, I am grateful to Richard Brooks, Jack Dovidio, Andy Egan, Jeffrey Fagan, David Hilbert, Richard Holton, Joshua Knobe, Tony Laden, Alex Madva, Jennifer Nagel, Aaron Norby, Jonathan Phillips, Richard Samuels, Eric

Schwitzgebel, Susanna Siegel, Matthew Noah Smith and J.D. Trout. Thanks also to the members of John Bargh's and Jack Dovidio's labs (Fall 2009–Spring 2010), who introduced me to the bodies of psychological literature to which this paper makes appeal. I am immensely grateful to the audiences to whom I had a chance to present this material, including those at the Oberlin Colloquium in Epistemology (April 2010) and the Harvard/MIT "Belief and its Cousins" Mini-conference (January 2011), and at the Departments of Philosophy at Brown University (October 2010), Columbia University (February 2011), Northwestern University (March 2011), University of Illinois at Chicago Circle (March 2011) and the University of Pittsburgh (April 2011); I regret that I did not take notes in ways that allow me to acknowledge specific contributions in those settings, but they are manifold. Special thanks to my Yale Philosophy colleagues—including Facundo Alonso, George Bealer, Steve Darwall, Jay Elliot, Verity Harte, Shelly Kagan, David Possen, Barbara Sattler and Zoltán Gendler Szabó—for an incredibly helpful Faculty Lunch discussion in April 2011, and to Brendan Dill for a careful reading of the paper's penultimate draft. My greatest thanks are reserved for my extraordinary research assistant, Takuya Sawaoka, for his tireless bibliographic work and numerous outstanding suggestions, both stylistic and substantive.

## References

- Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., et al. (2006). They all look the same to me (unless they're angry): From out-group homogeneity to out-group heterogeneity. *Psychological Science*, *17*(10), 836–840.
- Alcoff, L. M. (2006). *Visible identities: Race, gender, and the self*. Oxford: Oxford University Press.
- Allport, G. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.
- Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, *12*(5), 385–390.
- 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), 652–661.
- Anderson, E. (2010). *The imperative of integration*. Princeton, NJ: Princeton University Press.
- Appiah, K. A., & Gutman, A. (1996). *Color conscious: The political morality of race*. Princeton, NJ: Princeton University Press.
- Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?". *Psychological Inquiry*, *15*(4), 257–278.
- Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Browin, J. (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, *35*(1), 29–46.
- Ayres, I., Vars, F. E., Zakariya, N. (2004–2005). To insure prejudice: Racial disparities in taxicab tipping. *Yale Law Journal*, *114*, 1613.
- Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development*, *64*(3), 711–728.
- Banaji, M. R. (2002). Stereotypes, social psychology of. In N. Smelser & P. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 15100–15104). New York, NY: Pergamon.
- Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and nurture in own-race face processing. *Psychological Science*, *17*(2), 159–163.
- Baruch, B. M. (1947). More production for peace. *Vital Speeches of the Day*, *13*(14), 425.
- Beilock, S. L., Jellison, W. A., Rydell, R. J., McConnell, A. R., & Carr, T. H. (2006). On the causal mechanisms of stereotype threat: Can skills that don't rely heavily on working memory still be threatened? *Personality and Social Psychology Bulletin*, *32*(8), 1059–1071.
- Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The cross-race category effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological Science*, *18*(8), 706–712.
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review*, *94*(4), 991–1013.
- Blum, L. (2002). *"I'm not a racist, but": The moral quandary of race*. Ithaca, NY: Cornell University Press.

- Bosson, J. K., Haymovitz, E. L., & Pinel, E. C. (2004). When saying and doing diverge: The effects of stereotype threat on self-reported versus non-verbal anxiety. *Journal of Experimental Social Psychology*, *40*(2), 247–255.
- Butler, D. M., & Broockman, D. E. (forthcoming). Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*.
- Central Intelligence Agency. (2010). *Games*. <https://www.cia.gov/kids-page/games/index.html>. Accessed 28 April.
- Chudler, E. H. (n.d.). *Colors, colors*. <http://faculty.washington.edu/chudler/words.html>. Accessed 31 May 2011.
- 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.
- Cosmides, L., Tooby, J., & Kurzban, B. (2003). Perceptions of race. *Trends in Cognitive Sciences*, *7*(4), 173–179.
- Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, *24*(6), 588–594.
- Croizet, J.-C., Desprès, G., Gauzins, M.-E., Huguet, P., Leyens, J.-P., & Méot, A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin*, *30*(6), 721–731.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18.
- Devine, P. G., & Elliott, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin*, *21*(11), 1139–1150.
- Devine, P. G., & Sharp, L. B. (2009). Automatic and controlled processes in stereotyping and prejudice. In T. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 61–82). New York, NY: Psychology Press.
- Digital History. (n.d.). [http://www.digitalhistory.uh.edu/modules/slavery/bibliographical\\_essay.html](http://www.digitalhistory.uh.edu/modules/slavery/bibliographical_essay.html). Accessed 21 May 2011.
- Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology*: (Vol. 36, pp. 1–52). San Diego, CA: Academic Press.
- Dovidio, J. F., Gaertner, S. L., & Kawakami, K. (in press). Racism. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *Handbook of prejudice, stereotyping, and discrimination*. Thousand Oaks, CA: Sage.
- Drunk History. (n.d.). <http://www.hbo.com/funny-or-die-presents/about/video/drunk-history.html>. Accessed 31 May 2011.
- Dunham, Y., Baron, A. S., & Banaji, M. R. (2006). From American city to Japanese village: A cross-cultural investigation of implicit race attitudes. *Child Development*, *77*(5), 1268–1281.
- Dunham, Y., Baron, A. S., & Banaji, M. R. (2007). Children and social groups: A developmental analysis of implicit consistency in Hispanic Americans. *Self and Identity*, *6*(2), 238–255.
- Farmer, A., & Terrell, D. (2001). Crime versus justice: Is there a trade-off? *Journal of Law and Economics*, *44*(2), 345–366.
- Ford, T. E., & Stangor, C. (1992). The role of diagnosticity in stereotype formation: Perceiving group means and variances. *Journal of Personality and Social Psychology*, *63*(3), 356–367.
- Franklin, J. H. (2005). *Mirror to America: The autobiography of John Hope Franklin*. New York, NY: Farrar, Straus and Giroux.
- Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing intergroup bias: The common ingroup identity model*. Philadelphia, PA: Psychology Press.
- Gendler, T. S. (2000). The puzzle of imaginative resistance. *The Journal of Philosophy*, *97*(2), 55–81.
- Gendler, T. S. (2006). Imaginative resistance revisited. In S. Nichols (Ed.), *The architecture of the imagination* (pp. 149–174). Oxford: Oxford University Press.
- Gendler, T. S. (2008a). Alief and belief. *Journal of Philosophy*, *105*(10), 634–663.
- Gendler, T. S. (2008b). Alief in action (and reaction). *Mind & Language*, *23*(5), 552–585.
- Gendler, T. S. (2010). *Intuition, imagination and philosophical methodology: Selected papers*. Oxford: Oxford University Press.
- Gendler, T. S., & Hawthorne, J. (2005). The real guide to fake barns: A catalogue of gifts for your epistemic enemies. *Philosophical Studies*, *124*(3), 331–352.

- George Bernard Shaw. (n.d.). In Wikiquote. [http://en.wikiquote.org/wiki/George\\_Bernard\\_Shaw#Disputed](http://en.wikiquote.org/wiki/George_Bernard_Shaw#Disputed). Accessed 31 May 2011.
- Gonzalez, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin*, 28(5), 659–670.
- Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2001). Words and shape similarity guide 13-month-olds' inferences about nonobvious object properties. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the twenty-third annual conference of the cognitive science society* (pp. 352–357). Hillsdale, NJ: Erlbaum.
- Hamilton, D. L., & Sherman, J. W. (1994). Stereotypes. In R. S. Wyer Jr. & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., Vol. 2, pp. 1–68). Hillsdale, NJ: Erlbaum.
- Hamilton, D. L., & Trolier, T. K. (1986). Stereotypes and stereotyping: An overview of the cognitive approach. In J. Dovidio & S. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 127–163). Orlando, FL: Academic Press.
- Helman, E., Mania, E. W., & Gaertner, S. L. (2010). Where the division lies: Common ingroup identity moderates the cross-race effect. *Journal of Experimental Social Psychology*, 46(2), 445–448.
- Hess, T. M., Auman, C., Colcombe, S. J., & Rahhal, T. A. (2003). The impact of stereotype threat on age differences in memory performance. *Journal of Gerontology*, 58(1), 3–11.
- Heuer, R. J. (1999). *Psychology of intelligence analysis*. Washington, DC: Center for Study of Intelligence, Central Intelligence Agency.
- Hills, P. J., & Lewis, M. B. (2006). Reducing the own-race bias in face recognition by shifting attention. *Quarterly Journal of Experimental Psychology*, 59(6), 996–1002.
- Hodson, G., Hooper, H., Dovidio, J. F., & Gaertner, S. L. (2005). Aversive racism in Britain: Legal decisions and the use of inadmissible evidence. *European Journal of Social Psychology*, 35(4), 437–448.
- Huebner, B. (2009). Troubles with stereotypes for spinozan minds. *Philosophy of the Social Sciences*, 39(1), 63–92.
- Hugenberg, K., Miller, J., & Claypool, H. M. (2007). Categorization and individuation in the cross-race recognition deficit: Toward a solution to an insidious problem. *Journal of Experimental Social Psychology*, 43(2), 334–340.
- IAT Corp. (1998). *Project implicit*. <https://implicit.harvard.edu/implicit/>. Accessed 31 May 2011.
- James, W. (1890). *The principles of psychology*. New York, NY: Henry Holt and Co.
- Jesse Jackson. (n.d.). In Wikiquote. [http://en.wikiquote.org/wiki/Jesse\\_Jackson](http://en.wikiquote.org/wiki/Jesse_Jackson). Accessed 31 May 2011.
- Kahneman, D., & Tversky, A. (1972). On prediction and judgment. *Oregon Research Institute Research Bulletin*, 12(4).
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.
- Karpinski, A., & Hilton, J. L. (2001). Attitudes and the implicit association test. *Journal of Personality and Social Psychology*, 81(5), 774–788.
- Kelly, D. R., Machery, E., & Mallon, R. (2010). Race and racial cognition. In J. M. Doris, et al. (Eds.), *The Oxford handbook of moral psychology* (pp. 433–472). Oxford University Press: Oxford.
- Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy. *Psychological Science*, 18(12), 1084–1089.
- Kelly, D. R., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass*, 3(3), 522–540.
- Klinger, M., & Beall, P. (1992, May). *Conscious and unconscious effects of stereotype activation*. Paper Presented at the 64th Annual Meeting of the Midwestern Psychological Association, Chicago.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Normative, descriptive, and methodological challenges. *Behavioral & Brain Science*, 19(1), 1–17.
- Kovel, J. (1970). *White racism: A psychohistory*. New York, NY: Pantheon Books, Random House.
- Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, 67(4), 596–610.
- Kunda, Z. (1999). *Social cognition: Making sense of people*. Cambridge, MA: MIT Press.
- 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.
- Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, 98(26), 15387–15392.

- LaFree, G., Baumer, E. P., & O'Brien, R. (2010). Still separate and unequal? A city-level analysis of the black-white gap in homicide arrests since 1960. *American Sociological Review*, *75*(1), 75–100.
- Lebrecht, S., Pierce, L. J., Tarr, M. J., & Tanaka, J. W. (2009). Perceptual other-race training reduces implicit racial bias. *PLoS ONE*, *4*(1), e4215.
- Leslie, S. J. (forthcoming). The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*.
- Lever, A. (2005). Why racial profiling is hard to justify: A response to Risse and Zeckhauser. *Philosophy & Public Affairs*, *33*(1), 94–110.
- Lippman, W. (1922). *Public opinion*. New York, NY: Harcourt, Brace.
- Lippman, W. (1947). *The cold war: a study in US foreign policy*. New York, NY: Harper.
- Locksley, A., Borgida, E., Brekke, N., & Hepburn, C. (1980). Sex stereotypes and social judgments. *Journal of Personality and Social Psychology*, *39*(5), 821–831.
- Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology*, *23*(1), 77–87.
- Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, *66*(1), 37–47.
- McGugin, R. W., Tanaka, J. W., Lebrecht, S., Tarr, M. J., & Gauthier, I. (2011). Race-specific perceptual discrimination improvement following short individuation training with faces. *Cognitive Science*, *35*(2), 330–347.
- Medin, D. L. (1988). Social categorization: Structures, processes, and purposes. In R. Wyer & T. Srull (Eds.), *Advances in social cognition*: (Vol. 1, pp. 119–125). Hillsdale, NJ: Erlbaum.
- Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 343–355.
- Mills, C. W. (1997). *The racial contract*. Ithaca, NY: Cornell University Press.
- Myrdal, G. (1944). *An American dilemma: The Negro problem and the modern democracy*. New York, NY: Harper & Bros.
- Neuville, E., & Croizet, J. C. (2007). Can salience of gender identity impair math performance among 7–8 years old girls? The moderating role of task difficulty. *European Journal of Psychology of Education*, *22*(3), 307–316.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2006). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Philadelphia: Psychology Press.
- Oakes, P. J., & Turner, J. C. (1990). Is limited information processing capacity the cause of social stereotyping? In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 1, pp. 111–135). Chichester, UK: Wiley.
- Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the implicit association test: Personalizing the IAT. *Journal of Personality and Social Psychology*, *86*(5), 653–667.
- Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192.
- Payne, B. K. (2006). Weapon bias: Split-second decisions and unintended stereotyping. *Current directions in Psychological Science*, *15*(6), 287–291.
- Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in racebased misperceptions of weapons. *Journal of Experimental Social Psychology*, *38*(4), 384–396.
- Pious, S. (2002). *Implicit association test*. <http://www.understandingprejudice.org/iat/>. Accessed 31 May 2011.
- Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat on women's generation of mathematical problem solving strategies. *Journal of Social Issues*, *57*(1), 55–71.
- Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., et al. (2003). An fMRI examination of the impact of interracial contact on executive function. *Nature Neuroscience*, *6*(12), 1323–1328.
- Richeson, J. A., & Shelton, J. N. (2003). When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science*, *14*(3), 287–290.
- Richeson, J. A., & Shelton, J. N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. *Current Directions in Psychological Science*, *16*(6), 316–320.

- Richeson, J. A., Trawalter, S., & Shelton, J. N. (2005). African-Americans' racial attitudes and the depletion of executive function after interracial interactions. *Social Cognition, 23*(4), 336–352.
- Risse, M., & Zeckhauser, R. (2004). Racial profiling. *Philosophy & Public Affairs, 32*(2), 131–170.
- Rothstein, E. (2011). *Suspended in space, 103 stories over Chicago*. [http://www.nytimes.com/2011/05/31/arts/design/willis-tower-suspends-visitors-above-chicago.html?\\_r=2&scp=2&sq=chicago&st=cse](http://www.nytimes.com/2011/05/31/arts/design/willis-tower-suspends-visitors-above-chicago.html?_r=2&scp=2&sq=chicago&st=cse). Accessed 31 May 2011.
- Rydell, R. J., McConnell, A. R., & Beilock, S. L. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology, 96*(5), 949–966.
- Sangrigoli, S., Pallier, C., Argenti, A.-M., Ventura, V. A. G., & de Schonen, S. (2005). Reversibility of the other-race effect in face recognition during childhood. *Psychological Science, 16*(6), 440–444.
- Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat on performance. *Psychological Review, 115*(2), 336–356.
- Schneider, D. J. (2004). *The psychology of stereotyping*. New York, NY: Guilford Press.
- Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100*(2), 139–156.
- Shelby, T. (2005). *We who are dark: The philosophical foundations of black solidarity*. Cambridge, MA: Harvard University Press.
- Shih, M., Pittinsky, T. L., & Trahan, A. (2006). Domain-specific effects of stereotypes on performance. *Self and Identity, 5*(1), 1–14.
- Shriver, E. R., Young, S. G., Hugenberg, K., Bernstein, M. J., & Lantner, 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.
- Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7*(1), 36–97.
- Stangor, C. (2009). The study of stereotyping, prejudice, and discrimination within social psychology: A quick history of theory and research. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 1–22). New York, NY: Psychology Press.
- Steele, C. M. (2010). *Whistling Vivaldi: And other clues to how stereotypes affect us (issues of our time)*. New York, NY: W. W. Norton and Co.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811.
- Stone, J. (2002). Battling doubt by avoiding practice: The effects of stereotype threat on self-handicapping in white athletes. *Personality and Social Psychology Bulletin, 28*(12), 1667–1678.
- Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on black and white athletic performance. *Journal of Personality and Social Psychology, 77*(6), 1213–1227.
- Stroessner, S., Good, C., & Webster, L. (n.d.). <http://reducingstereotypethreat.org>. Accessed 21 May 2011.
- Sullivan, S. (2006). *Revealing whiteness: The unconscious habits of racial privilege*. Bloomington and Indianapolis, IN: University of Indiana Press.
- Tajfel, H. (1959). Quantitative judgment in social perception. *British Journal of Psychology, 50*(1), 16–29.
- Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Social Issues, 25*(4), 79–97.
- Tajfel, H. (1981). *Human groups and social categories*. Cambridge, MA: Cambridge University Press.
- Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgment. *British Journal of Psychology, 54*(2), 101–114.
- Tanaka, J. W., Kiefer, M., & Bukach, C. M. (2004). A holistic account of the own-race effect in face recognition: Evidence from a cross cultural study. *Cognition, 93*(1), B1–B9.
- Tanaka, J. W., & Piece, L. J. (2009). The neural plasticity of other-race face recognition. *Cognitive Affective Behavioral Neuroscience, 9*(1), 122–131.
- Taylor, S. E. (1981). A categorization approach to stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 145–182). Hillsdale, NJ: Erlbaum.
- Tetlock, P. F., Kristel, O., Elson, B., Green, M., & 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.
- Trawalter, S., & Richeson, J. A. (2006). Regulatory focus and executive function after interracial interactions. *Journal of Experimental Social Psychology, 42*(3), 406–412.

- Trawalter, S., & Richeson, J. A. (2008). Let's talk about race, baby! When whites' and blacks' interracial contact experiences diverge. *Journal of Experimental Social Psychology*, *44*(4), 1214–1217.
- Trawalter, S., Richeson, J. A., & Shelton, J. N. (2009). Predicting behavior during interracial interactions: A stress and coping approach. *Personality and Social Psychology Review*, *13*(4), 243–268.
- von Fintel, K., & Gillies, A. (2008). CIA leaks. *Philosophical Review*, *117*(1), 77–98.
- WebMD. (2010). *Why breakfast is the most important meal of the day*. <http://www.webmd.com/diet/guide/most-important-meal>. Accessed 27 May 2011.
- Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, *54*(1), 277–295.
- Wraga, M., Duncan, L., Jacobs, E. C., Helt, M., & Church, J. (2006). Stereotype susceptibility narrows the gender gap in imagined self-rotation performance. *Psychonomic Bulletin & Review*, *13*(5), 813–819.
- Zwicky, A. (2008). *Wile away*. <http://languageblog ldc.upenn.edu/nll/?p=466>. Accessed 31 May 2011.