1.1

Who's Responsible for This?

Moral Responsibility, Externalism, and Knowledge about Implicit Bias

Natalia Washington and Daniel Kelly

1 The Cognitive Monster

Recently, philosophers have become increasingly concerned about a cluster of issues that arise at the intersection of ethics and psychology. The general worry was expressed by John Bargh in his influential paper "The Cognitive Monster" (1999):

> If it were indeed the case, as research appeared to indicate, that stereotyping occurs without an individual's awareness or intention, then the implications for society—specifically, the hope that prejudice and discrimination could eventually be eradicated—were tremendous, as well as tremendously depressing. Most ominously, how could anyone be held responsible, legally or otherwise, for discriminatory or prejudicial behavior when psychological science had shown such effects to occur unintentionally? (363)

We agree with Bargh that the picture emerging from many areas of empirical psychology is both theoretically intriguing and morally troubling. Taken as a whole, he sees this picture as suggesting that a significant amount of human behavior is at the behest of a cognitive monster that operates outside of our conscious awareness, producing behaviors unguided by explicit intention. When they are produced in such a way, it is difficult to see how we could justifiably be held responsible for those behaviors, given the principles that typically govern practices of holding people responsible. On the one hand, facts about the kind of implicit and automatic mental processes that Bargh alludes to in his description of the "cognitive monster" do not fit easily with a folk psychological picture of the mind, and as we will see, the way that they deviate from that picture can make

them seem not just counterintuitive but somewhat unsettling. On the other hand, theories of moral responsibility are often grounded in intuitions, social norms, and practices that rely on commonsense conceptions of the sources of behavior. This raises a difficulty concerning how to square these facts with these theories, and how the latter can best be brought to bear on the former. In this paper we aim to think systematically about, formulate, and begin addressing some of the challenges to applying theories of moral responsibility to behaviors shaped by a particular subset of unsettling psychological complexities: namely, implicit biases.[1]

One might initially be skeptical that implicit biases raise any special challenge to moral responsibility. We disagree. Echoing some of Bargh's themes, Jennifer Saul (2013) sketches a position about implicit bias and blameworthiness in terms of awareness and control:

> I think it is also important to abandon the view that all biases against stigmatized groups are blameworthy. My first reason for abandoning this view is its falsehood. A person should not be blamed for an implicit bias that they are completely unaware of, which results solely from the fact that they live in a sexist culture. Even once they become aware that they are likely to have implicit biases, they do not instantly become able to control their biases, and so they should not be blamed for them. (55)

Saul's discussion shows that there are intricacies here that deserve careful philosophical attention. Indeed, we follow her in framing our questions about responsibility and blame in terms of control and awareness or knowledge.[2] Moreover, we agree with her more specific claim that individuals do not become able to control their biases simply by or immediately upon learning about them. The case we will make below takes this as given, and argues that there are nevertheless cases in which an individual should be held responsible for actions that are influenced by her implicit biases—even if she cannot control them at the time of the behavior, and even if she does not know she has those implicit biases, and would disavow those biases were she their influence.[3] For reasons that will

[1] Implicit bias hardly exhausts the domain of interesting, counterintuitive, and philosophically relevant "unsettling facts" being discovered about human psychology. In order to focus our discussion in this paper we will be bracketing some other relevant areas of research, including those on the unreliability of self-report (Nisbett and Wilson, 1977; Wilson, 2002), confabulation and post hoc rationalization (Hirstein, 2005; Haidt, 2006; Tavris and Aronson, 2007), and stereotypes and stereotype threat (Carr and Steele, 2010).

[2] See Kelly and Roedder (2008: 532) for similar worries expressed in terms of Frankfurtian identification and Fischer and Ravizza's notion of reasons-responsiveness.

[3] In a recent paper, Jules Holroyd (2012) contends that arguments in support of the conclusion that individuals cannot be held responsible for manifesting implicit biases are untenable. We want to make the stronger claim, that in many situations, individuals *are* responsible for manifesting implicit biases. We owe much to earlier discussions with her on this subject.

become evident as we go, our discussion will focus not on control but rather on knowledge about implicit biases. We should also be clear that the view we articulate and defend applies first and foremost to actions, rather than the agent who engages in an action. For the purposes of this essay we will remain neutral on the issue of whether or not someone is a bad person merely in virtue of harboring implicit biases, or if simply having implicit biases should be reflected in assessments of that person's character. Rather, we are concerned with whether and how a person should be held responsible when her biases are allowed to manifest in a specific piece of behavior. For us, the primary target of evaluation is action itself, even though responsibility and blame are eventually ascribed to the agent who engages in it.

Also central to our framing of questions about responsibility will be the notion of an exculpating condition. Broadly speaking, exculpating conditions specify factors that can excuse a person for an action, absolving her of responsibility and blame. In Section 2 we explain how such conditions operate in the common norms and practices that govern how we typically hold people responsible for some behaviors, but let them off the hook for others. Next we lay out several core features of implicit bias, emphasizing those features that correspond to common exculpating conditions, and spelling out how such correspondences motivate the kinds of worries about responsibility expressed by Bargh and explored by Saul. We then present a thought experiment designed to show that in certain circumstances, an individual who does not know she has implicit biases can nevertheless be held responsible for behaviors that were crucially influenced by those implicit biases. In reflecting on the intuition our thought experiment is designed to pump[4], we note that it implies that not all of the knowledge relevant to moral responsibility and exculpation need be "in the head" of the individual whose actions are being evaluated. Hence, one of our main claims is that an individual can be open to blame for manifesting implicit biases when knowledge about such mental states is available in her epistemic environment, and that individual occupies a social role to which implicit biases and knowledge about them are clearly relevant. Since we see ourselves as articulating the best way to extend current practices concerning responsibility to actions that involve the psychological complexities of the cognitive monster, along the way we illustrate and defend this claim by comparing our thought experiment to a number of parallel but more familiar cases. In Section 6 we respond to some common objections and comment on the context and pragmatic implications of our position.

[4] See Dennett (2013) for a discussion of thought experiments as "intuition pumps."

2 Holding Ourselves Morally Responsible: Common Practices, Typical Excuses

In broad strokes, it is common practice to hold people morally responsible for many of their actions, where being held responsible is tied up with notions of praise and blame, reward, and punishment. When a person performs an action that is morally good, particularly when it is exemplary or supererogatory, the person may be praised, and can be justifiably rewarded. More importantly, when a person performs an action that is morally bad or wrong, the person is open to blame, and may be justifiably punished.[5]

Of course, the picture is not so simple. For instance, people are not held morally responsible for all of their behaviors. It is also common practice to excuse people for behaviors in special kinds of circumstances. Broadly speaking, such circumstances include those in which the behavior in question is forced or coerced, in which the behavior is accidental, or behaviors in which the agent is ignorant or unaware of some key element of her situation.[6] In these cases, we typically do not praise or reward the agent, even if the behavior has a morally good or desirable outcome, nor do we blame or punish the agent if the behavior has a morally bad or undesirable outcome. Behaviors that do not occur in such circumstances are sometimes said to be free. They can be described by phrases like "the behavior was freely chosen" or "it was the result of a genuine decision;" in some relevant sense, the agent could have reasonably done otherwise. In short, we hold each other morally responsible for those actions that are freely chosen—expressions of free will.[7]

[5] It is broad common practice in the cultural environment that the authors and probably most readers of this chapter inhabit. Whether even this very general characterization of the norms and practices concerning ascription of responsibility and blame applies to all cultures, or merely to those cultures that are WEIRD or heavily influenced by WEIRD cultures, is a fascinating and still largely underexplored question (Henrich et al., 2010; though see Sommers, 2011).

[6] Although we mention here only control- and knowledge-based exculpation conditions, we do not assume these are the only two types; for instance, see Machery et al. (2010) for discussion of a rationality-based condition.

[7] We realize we are passing over some rich philosophical ground quickly in these paragraphs, but we will complicate the picture as needed to make our key points as we go. Here we simply wish to give a coarse-grained overview of the types of connections between the key concepts of free will, responsibility, and praise and blame, as they are construed by folk psychology and common practice (or at least as how many "Introduction to Philosophy" courses depict folk psychology and common practices as construing those connections.) For discussion of some more fine-grained notions related to responsibility that philosophers have advanced, see Watson (1996), Shoemaker (2011), and Smith (2012). Most of what talk we about in terms of responsibility seems to us to most comfortably fall under the category of "accountability." However, little appears to be settled in this area (cf. Sripada, forthcoming). Even if the distinct notions prove to be defensible and significant, questions about if

Philosophers interested in these issues have long been concerned about the possibility of global threats to free will and moral responsibility. For instance, one such threat seems to emerge from contemporary physics, and an appreciation of the fact that there seems to be no room for genuine choice or moral responsibility if we are living in a deterministic universe. Others worry that another sweeping threat may be looming in the results of recent cognitive neuroscience which suggests that our actions are the result of mental processes that completely bypass our conscious, reflective deliberation and decision making (see Nahmias 2010, Roskies 2006). Although fascinating, our concern here will not be with either of these "global challenges" to moral responsibility. Rather, we will adopt the assumption built into our ordinary, everyday practices of holding people responsible—typical of compatibilist approaches to free will—that we *are* responsible for *some* of our behaviors.

To illustrate, consider this piece of behavior: Cate eats a batch of the cookies that her roommate made especially for tomorrow's bake sale. In the first variation of this scenario, Cate knows full well that her roommate made the cookies especially for the bake sale, but eats them anyway, for no other reason than because she is hungry. Described as such, Cate's behavior is not just callous, but it is of the sort for which she is responsible; she is a straightforward target for blame. However, in other circumstances we would not hold her responsible for the same piece of behavior. In a second variation of this scenario, imagine that Cate eats a batch of the cookies that her roommate made especially for tomorrow's bake sale, but she does so while in a somnambulant daze. Since in this case there is an important sense in which she did not know what she was doing, it is not her fault for ruining her roommate's contribution to the bake sale, and she should not be blamed but excused, because people are not commonly held responsible for their behaviors while they are sleep-walking.

Generalizing from this second variation, we will say that behaviors that occur in these kinds of scenario satisfy an *exculpating condition*. We will focus on two of the most important kinds of exculpating condition: those that center on knowledge and control, respectively. To a first approximation, in cases that satisfy the knowledge condition, the agent is excused for the behavior because she did not know or was unaware of relevant features of the situation, and in cases that satisfy the control condition, the agent is excused for the behavior because she did not have the right kind of control over it. This makes sleep-eating Cate an especially apt case for exculpation, since at first blush it seems that she is both unaware that

and how each one applies to the kinds of cases we consider here, while potentially interesting and important, would require more space than we have here to address properly.

she is sleep-eating, and unable to stop herself. In any event, we can begin articulating this line of thought as follows:

Knowledge Condition: An agent is exculpated for having done X if
(1) she did not know that she did X, or
(2) she did not know why she did X.

Control Condition: An agent is exculpated for having done X if
(1) she was constrained or coerced to do X, or
(2) she lacks proper control over doing X.

In addition to letting people off the hook for what they do in certain circumstances, ordinary practices of ascribing responsibility also allow for another wrinkle; in our terminology, exculpating conditions also have exception clauses. When a case of behavior occurs in circumstances that satisfy an exculpating condition, but the circumstances *also* meet one of the exculpating condition's exception clauses, then the agent is not excused, but instead is held responsible for the action. Consider a third variation on our cookie thief: Cate, while in a somnambulant daze, eats a batch of the cookies that her roommate made especially for tomorrow's bake sale, but she was in that somnambulant daze because she had taken a hefty dose of Ambien. Cate has a long history, of which she and her roommate are both well aware, of sleep-walking and binge eating whenever she takes Ambien. In this case, Cate's behavior indeed satisfies an exculpating condition (indeed, a case could be made that she satisfies both), but it also satisfies an exception clause (or two): somnambulant Cate is unaware of and unable to consciously control what she is doing, but, like a drunk driver, she is responsible for having put herself in that compromised condition, and is blameworthy for what she does once she inhabits it—in this case, especially since she has a well-known history of such Ambien-induced destructive behavior.[8] Excuses, as they say, wear thin. More generally, exception clauses can be represented like this:

Knowledge Condition: An agent is exculpated for having done X if
(1) she did not know that she did X (except when she is responsible for having been unaware of doing X), or
(2) she did not know why she did X (except when she is responsible for having been ignorant)

[8] For more on these kinds of so-called "tracing" cases, see Vargas (2005) and Fischer and Tognazzini (2009), and for a discussion focused on culpable ignorance, see Smith (2011).

Control Condition: An agent is exculpated for having done X if
(1) she was constrained or coerced to do X (except when she is responsible for having been constrained), or
(2) she did not have volition or control over doing X (except when she is responsible for lacking that control)

We will return to this framework in Section 4, where we will show how it applies to and illuminates a few more fairly mundane cases, before bringing it to bear on a case that involves implicit bias. First, however, we will briefly describe how we are construing this range of unsettling psychological facts.

3 "Textbook" Facts about Implicit Bias

Much is still being discovered about the character of implicit biases, and we do not wish our argument to depend on any of the more controversial or uncertain aspects of the ongoing research. To that end, we will work with a fairly broad conception of implicit biases as unconscious and automatic negative evaluative tendencies directed towards people based on their membership in a stigmatized social group—for example, on gender, sexual orientation, race, age, or weight. Such biases appear to be widespread in many populations, cultures, and countries. For ease of exposition and to focus the discussion to come, we will confine most of our attention to implicit racial biases. We hope and suspect that what we have to say generalizes straightforwardly to other types of implicit biases as well.[9]

Here are four features of implicit bias that will be important for the discussion to come:

1) *Dissociation* In a single individual, implicit racial biases can coexist with explicit racial attitudes that are diametrically opposed to them. For example, a person can explicitly hold genuine anti-racist or egalitarian views that they sincerely endorse upon reflection, and yet at the same time harbor implicit biases again members of certain races.
2) *Introspective opacity* Typically, a person who is explicitly biased knowingly and intentionally evaluates others negatively based on their race. In contrast, a person who is only implicitly biased has tendencies whose presence and influence on thought and behavior is not easily detectable via introspection.

[9] For citations and details, see Brownstein and Saul's Introduction in Volume 1. The view of implicit biases articulated there is fairly standard, and bears much in common with the minimal view on which we rely. For more detailed discussion see the chapters by Frankish, Huebner, Holroyd and Sweetman, Machery, and Mallon in the first part of that volume, as well as Staats and Patton (2013) and Banaji and Greenwald (2013).

Moreover, her sincere self-reports about her own attitudes are not likely to reflect her implicit tendencies. Indeed, much of what is known about implicit biases comes not from self-report, but is rather inferred from indirect experimental techniques like the Implicit Association Test, startle eye blink tests, and semantic priming tests. These experimental techniques are indirect in that they do not directly rely on participants' powers of introspection or the accuracy of descriptions of their own psychological makeup.

3) *Recalcitrance* Implicit biases operate not just implicitly but automatically. It is difficult to completely suppress the manifestation of an implicit bias in either judgment or behavior. They are also much easier to acquire than they are to eradicate or completely remove from one's psychological makeup. Moreover, while they are amenable to some methods of control, directly suppressing their expression once activated requires vigilance and effort, is mentally fatiguing, and can backfire in a number of ironic ways.[10]
4) *Widespread effects on behavior* Implicit racial biases can influence judgments and behaviors in subtle but important ways, even in real world situations. Many studies suggest implicit biases influence snap decisions, such as determining whether or not a person is holding a weapon (Payne, 2005, 2006), or if a basketball player has committed a foul (Price and Wolfers, 2010). There is reason to believe that implicit biases can also influence more deliberate, temporally extended decision making, despite confidence that in such cases behavior is more likely to reflect explicit attitudes and considered views. Examples include what diagnosis or type of health care a medical patient should get (Blair et al,. 2011), who should or should not to serve on a jury (Haney-López, 2000), and whom to hire, or which resumé gets an interview (Bertrand and Mullainathan, 2004; see also Kawakami et al., 2007).

Given these four features, we can further refine the worry behind Bargh's invocation of a cognitive monster. We have formulated two core exculpating conditions: one for knowledge and one for control. These conditions align neatly with two prominent properties of implicit biases: namely, that their existence and influence is not easily detectable via introspection—an individual can have and be

[10] Happily, it seems that implicit biases are not completely intractable or uncontrollable. Work on what we will call the malleability of implicit biases has shown a number of techniques to be quite promising, including both implementation intentions and exposure to counterstereotypical exemplars. For further discussion on the implications of this work, see Madva (ms.) and the chapters by Rees and Brownstein in these volumes .

influenced by implicit biases without *knowing* it—and that they are recalcitrant, liable to run automatically in the face of contradictory, explicitly held beliefs—an individual's implicit biases operate without and sometimes beyond her *control*. Given this, the potential trouble can be put rather starkly: behaviors driven by implicit biases will, in virtue of that neat alignment, satisfy one or both of the two key exculpating conditions, so the agent who engages in implicit bias-driven behavior should be absolved of responsibility or blame for it. Expressed another way:

Epistemic Worry Since implicit biases are opaque to introspection and can operate outside of conscious awareness, a person should be exculpated, and not blamed or held responsible for behaviors that manifest them.

Bypassing Worry Since the operation of implicit biases is recalcitrant and automatic, a person should be exculpated, and not blamed or held responsible for behaviors that manifest them.

In Sections 4 and 5 we spell out one way to dispel these two worries generated by the neat alignment between these exculpating conditions and this set of unsettling psychological facts. We argue that the growing knowledge about implicit biases, which includes a body of empirical research showing how implicit biases can best be brought under control, has important implications for responsibility and blame. Indeed, we suggest how this knowledge should be incorporated into what Manuel Vargas (2013) calls our *moral ecology*. We do this by showing how that research and its dissemination are relevant to the kinds of exception clauses and exculpating conditions discussed above.

4 The Hiring Committee

Consider three different people, each with one of the following psychological profiles:[11]

The Earnest Explicit Racist (circa whenever) The earnest explicit racist has implicit racial biases, but these are accompanied by explicitly racist attitudes as well. She is fully aware that she holds these explicit views and is able and willing to articulate them, though perhaps only among trusted friends. She reflectively endorses her racist attitudes, and acts on them without compunction when given the chance. Though she does not know about her implicit biases, if made aware she would take pride in the fact that these instinctive evaluative tendencies run in tandem with her more reflective judgments, and that both express her considered values.

[11] We are not under the mistaken impression that these characters exhaust the range of possible or interesting cases when it comes to evaluating implicitly biased behavior. We have chosen these three to highlight a particular situational feature, and to make the case that people can be blamed for behaviors driven by implicit biases they do not know they have. In Section 6 we will briefly consider some questions about how to extend our approach to other cases.

The Old-School Egalitarian (circa 1980) The old-school egalitarian is explicitly anti-racist. She genuinely holds egalitarian views, which she honestly reports when asked about her views on race, and which she sincerely endorses upon reflection. However, the old-school egalitarian also harbors implicit racial biases. Like almost everyone in 1980, though, she does not know this fact about herself. Not only is she unaware that she herself is implicitly biased, she has never heard of implicit biases at all. Unlike the explicit racist, she suffers from what we called dissociation, and so if it were somehow revealed to her that she was implicitly biased, she would be surprised and taken aback. In fact, she would disavow those evaluative tendencies, and would acknowledge that in cases where her implicit racial biases influenced her decisions or behaviors, something had gone wrong. Such decisions and behaviors would not express her considered values, and she would be falling short of her own avowed ideals.

The New Egalitarian (circa 2014) Like the old-school egalitarian, the new egalitarian is genuinely explicitly anti-racist and egalitarian. He too harbors implicit racial biases but does not know this fact about himself. Like many others in 2014, however, the new egalitarian is vaguely aware of the phenomenon of implicit bias, but has not much looked into the matter, and so does not know any details. Nor has he checked to see if he has any implicit biases himself. As a result, he takes no precautions and makes no adjustments to his own behavior to suppress or counteract them. However, as with the old-school egalitarian, if the new egalitarian came to know that he was implicitly biased he would not endorse those evaluative tendencies, but would sincerely disavow them. He too would acknowledge that in cases where his implicit racial biases influenced his decisions or behaviors, something had gone wrong. He would be failing to express his values and to live up to his own stated ideals.

Now indulge us in a little bit of science fiction fancifulness (given how the members of our cast of characters are indexed to different times), and imagine a scenario where these three people comprise a hiring committee. They have all the usual duties that come with membership on such committees, but most important is their task of sorting through the set of resumés submitted for consideration for the job, reading them over with an eye toward deciding which candidates to interview and, ultimately, to hire. As such, each committee member puts in the considerable time it takes to sort and evaluate those resumés, and the individual and collective decision-making processes both involve lots of conscious, explicit, and deliberate reasoning. But the processes are all unknowingly influenced by the implicit racial biases of the individual members as well. As a result, all of those selected to be interviewed turn out to be white; the committee overwhelming favored resumés from candidates with "white-sounding" names, even though there were equally well-qualified candidates of many racial and ethnic backgrounds in the application pool (the kind of outcome found in e.g. Bertrand and Mullainathan, 2004).

Now we are able to pose our paper's eponymous question: Who is responsible for this? The outcome is clearly unjust, as many deserving candidates were not given a fair shot at the job for reasons that had nothing to do with their

qualifications. Where does blame attach, and how blameworthy is each member of the committee for bringing about this outcome?

It seems to us that the earnest explicit racist is the most straightforward case, and thus the least interesting. On our construal, she knew full well what she was doing when she chose candidates with white-sounding names. She may have been helped along by her implicit biases, but she was not coerced in any way. Upon reflection, she would endorse her contribution to the committee, its procedures, and the hire in which they culminated. Her actions, assessments, and decisions, and the outcome they helped bring about are, in fact, an expression of her considered intentions. It is not clear that she would want recourse to any exculpating condition, and in any event does not satisfy either one. She is straightforwardly responsible and deserves considerable blame.

A much more interesting matter lies in what we think is an important difference between the old and new egalitarians. We described them as both having roughly the same psychological profile—the same avowed egalitarian ideals, but also the same implicit racial biases—and contributing equally to the same unjust outcome; so one might be tempted to say that they both deserve roughly the same amount of blame as well. We think this would be a mistake. Rather, we hold that the new school egalitarian is considerably more responsible than the old, and that more blame should be directed at him. Although neither knew they had implicit racial biases, and neither would endorse those biases or their manifestation in this case, given the relevant differences in external contexts and wider social circumstances, especially the psychological research that accumulated between 1980 and 2014, the new egalitarian *could* have and *ought* to have known about this, and *could* have and *ought* to have taken appropriate steps to nullify or counteract their influence on the decision process. Since the same cannot be said of the old-school egalitarian, she does not deserve nearly as much blame as the new. Times change; excuses wear thin.

To begin developing this intuition and what lies behind it, consider some arguments that, if successful, would absolve *both* egalitarians. One way to try to do this would be by appeal to a control-based exculpating condition, expressing a form of the Bypassing Worry we mentioned previously. This does not strike us as a promising way to go. First, we reject the claim that neither egalitarian had *any* kind of control over their decisions, or the resulting hire.[12] Indeed, for any of the

[12] The question of what kind of control is required for moral responsibility is a controversial one, to say the least. For some discussion of the different types of control that philosophers have thought to be relevant, see Holroyd and Kelly (forthcoming), who also defend a two-step argument that 1) implicit biases are in fact subject to what Andy Clark (2007) calls "ecological control," and that 2) ecological control is sufficient for moral responsibility.

three individuals on the committee, it seems utterly implausible to us that either person's explicit deliberative capacities were *completely* bypassed, or that what each individual experienced as conscious reflection over the course of the hiring process was purely epiphenomenal. Unlike the kind of snap decisions made by NBA referees, evaluating and ranking resumés is a case of a slow, temporally extended decision-making process. In such cases, conscious reflection is likely to be an important and causally efficacious element of the story, and so the resulting decision is certainly not a direct, unadulterated expression of implicit bias. Rather, the contribution of this aspect of the cognitive monster to the slow, temporally extended procedure is more like an illicit thumb on the scales of deliberation, contaminating the process. Moreover, it is a corrupting influence that, if recognized, could be neutralized in the type of case featured in our thought experiment, e.g. perhaps by removing or blinding oneself to the names from the resumés beforehand, or by using one of the other techniques that research has shown to be effective. The upshot of this is that if there is a problem for holding either of the egalitarians responsible, it is not that their conscious, deliberative, or other agential capacities were completely bypassed by the psychological processes resulting in the hiring decision. Therefore, if either can be found free of fault, the most plausible way to make the case is in terms of the Epistemic Worry rather than the Bypassing Worry.

Imagine an analogous (and perhaps dispiritingly familiar) case: you have a clueless student in one of your introductory classes who fails to show up to the midterm, but afterwards pleads his case to you. He claims that it was not his fault that he was absent because he was unaware that the examination was on that day. He begs for mercy, and asks to be allowed to retake the examination; he did not mean to blow it off, he just forgot to check the syllabus, and so did not know. Cast in our terms, the student is clearly making appeal to the knowledge-based exculpating condition. Moreover, we are willing to concede that he in fact satisfies the condition. But of course, all things considered, he is responsible for missing the examination (and you would be well within your rights to deny his plea to retake it). Even though the student genuinely did not know when the examination was taking place, his mere ignorance does not excuse his absence. He should have known when the midterm was; it was his responsibility to know. Unfortunately for him, he satisfies the exception clause, too.

Importantly, it is not *everyone's* responsibility to know this fact about the midterm. Rather, the student inherits this and other responsibilities in virtue of occupying a certain social role—in this case, being a member of the introductory class. In signing up for the class, it became the student's responsibility to know certain things, such as the date of the examination and the rest of the contents of

the syllabus. Particular responsibilities attach to other social roles as well. Consider a negligent doctor who fails to keep up with current medical findings and techniques, and so loses a patient who could have been saved easily by a relatively new, life-saving procedure of which she was unaware. Once again, the doctor genuinely does not know about the new life-saving procedure, but this ignorance alone does not excuse her. As a medical doctor it is her responsibility *to know certain things*, to keep abreast of the current state of medical knowledge. Not everyone is required or expected to have this knowledge, but certain people most definitely are, and if they do not, they are blamed for bad things that happen as a result of their ignorance. As in the case of the clueless student, we grant that in the scenario described, our negligent doctor satisfies the knowledge-based exculpating condition, but we maintain that she *also* satisfies the exception clause. She is not excused, and so is straightforwardly responsible for the situation, and deserves blame for her patient's death.[13]

With these examples in hand, we can say more precisely how knowledge of implicit bias affects the respective responsibility of our hiring committee egalitarians. But first note another commonality between the two: not only do they both share roughly the same psychological profile, but in being on the same hiring committee they also both occupy the same social role. Despite this, our assessments of the two diverge. While the old-school egalitarian did contribute to a morally problematic outcome, she herself is absolved of responsibility for that outcome, and bears little blame. She did not know that she harbored implicit racial biases, nor was she aware of implicit biases in general or how they were affecting her deliberations and judgments in assessing the resumés. In 1980, *no one* knew the unsettling psychological facts about implicit biases; the psychological research had not yet been done, and so today's wealth of empirical evidence simply did not exist. The old-school egalitarian is absolved in virtue of meeting the knowledge condition. She did not know about implicit biases, and could not have been expected to know about them.

The new egalitarian does not get off so easily. He has much in common with the old: he occupies the same social role as a member of the hiring committee,

[13] The negligent doctor case is similar to a case from Holly Smith's (1983 paper "Culpable ignorance." For Smith, an individual is culpably ignorant if they behave in a way that demonstrates an irresponsible willingness to take risks. We agree with Smith that there is a meaningful distinction between an individual who does not know any better and an individual who ought to have known better. In other words, for an individual to be culpable there must be a "benighting" act, in which irresponsible risk-taking occurs. But this does not mean that the benighting act is *all* that the individual is responsible for. We contend that individuals are responsible for their risk-taking *and* their later "unwitting" acts—the negligent doctor is responsible for not having kept up with her craft *and* for each subsequent patient she injures.

and contributes to the same morally problematic outcome. His deliberative process was not bypassed, nor was he coerced in any other way. He was likewise ignorant of his own implicit biases and their influence, so he satisfies the knowledge condition. But the new egalitarian also meets an exception clause, and therefore is responsible and bears more blame for the outcome than the old-school egalitarian does. This difference is a function of the external context and wider social circumstances he inhabits in virtue of being indexed to the year 2014. Today, the amount of empirical evidence collected on implicit biases is enormous, and it continues to mount. Much more is known in general, and that knowledge is much more widespread in the new egalitarian's epistemic environment than it was in the early 1980s of the old-school egalitarian.[14] The new egalitarian, like the old, does not know about his own implicit biases—but *unlike* the old-school egalitarian, he *should have been aware*. The differences in their wider social circumstances and informational environments are represented in the different relations each bears to the exception clause: only the new egalitarian satisfies the knowledge condition's exception clause, and thus bears considerably more blame than his old-school counterpart.

5 Taking a Step Back: Externalism and the Evolution of the Epistemic Environment

We hope readers share our initial intuition that there is something questionable about the new egalitarian when compared to his old-school counterpart, and we hope to have begun to unpack what lies behind that intuition in a way that helps clarify and strengthen the assessment it supports. In this section we will take a step back from the specifics of that case and reflect on some of the more general features of our approach, and point to some questions that it raises. While recent discoveries about a set of counterintuitive and unsettling facts of human psychology are at the forefront of our discussion, there is also something traditional about the approach we have taken to them. We understand ourselves to primarily be doing moral philosophy. We have not added to the evidence about implicit

[14] As we discuss in Section 6, facts about implicit biases are not yet a matter of common knowledge, but information about them continues to be disseminated into the wider culture. Since 1998, more than 16 million people have taken on online IAT (Brian Nosek, personal communication). Moreover, in the US, high-profile cases, such as those involving Trayvon Martin and Michael Brown, continue to bring media attention to racial bias, and implicit biases have been increasingly discussed in the popular press commentary on those cases (see e.g. <http://www.huffingtonpost.com/tag/implicit-bias> <http://www.motherjones.com/politics/2014/11/science-of-racism-prejudice> <http://www.nytimes.com/2015/01/04/upshot/the-measuring-sticks-of-racial-bias-.html?_r=0>

bias, nor argued for a new interpretation of the extant data. But neither have we urged that taking account of those unsettling facts will require extensive revision of our moral concepts or radical overhaul of the currently entrenched practices surrounding ascription of responsibility and blame.[15] Rather, we started with typical, recognizable patterns of moral reasoning, as represented in our exculpating conditions and exception clauses, and tried to show how they can be extended to deal sensibly with a class of cases that involve implicit bias-influenced behavior. Our discussions of the mundane examples of Cate the cookie thief, the clueless student, and the negligent doctor were designed to illustrate this point.

However, the putatively uncontroversial general premises that we take as our starting point—knowledge is relevant to moral responsibility, differences in knowledge can be reflected in differences in responsibility and blameworthiness, changes in knowledge can generate changes in responsibility and blameworthiness—can lead to less banal conclusions when combined with premises inspired by the thriving research on implicit biases. What is novel about cases involving implicit bias driven behavior is the psychology, and perhaps the epistemology of that psychology. As with the clueless student and negligent doctor, we are finding fault with our new egalitarian for failing to know something that he should have known. However, unlike the first two cases, some of the facts about which the new egalitarian is ignorant are facts about himself, his own mental processes and tendencies. Moreover, the new egalitarian cannot gain the relevant knowledge about these mental states—the character of implicit biases, and that such mental states are present and operative in his own psychological apparatus—simply by introspecting. But while the mind is not here transparent to itself, the new egalitarian can become aware of them. Knowledge of them just has to come via less direct, often external pathways (perhaps by taking an IAT online); in this sense, the epistemology of one's own implicit biases is non-Cartesian. Moreover, empirical research suggests that implicit biases cannot be well controlled via direct or immediate exercise of willpower. Rather, successfully curbing the influence of one's implicit biases will first require the acquisition of more and different knowledge from without. For not only does an individual need to know that she has implicit biases before she can even try to exert control over them, but doing so consistently

[15] Compare with Doris (2015), who considers a much larger array of counterintuitive and creepy facts revealed by recent psychological research, and argues that these present us with a dilemma. We must either radically revamp the conception of agency found in much of the philosophical literature, which gives pride of place to reflection and conscious deliberation, or, if we hold onto the reflectivist conception, we will be driven towards skepticism about persons, and the conclusion that genuine episodes of agency actually occur *much* less frequently than previously thought.

and effectively will also require a special kind of knowledge—specifically, knowledge of and facility with the kind of techniques and methods that are being shown effective by the empirical research on the malleability of implicit bias (see fn. 10).

Our line of reasoning dovetails with other anti-Cartesian trends in the philosophy of mind and cognitive science that often fall under the banner of externalism. There are many forms of externalism, but the common thread is an insistence that the boundaries of an individual's skin and skull are relatively unimportant when it comes to the nature and content of her mind, and the bases of her judgments and behavior. Some externalists have famously claimed that the content of mental states is in part determined by factors outside of the head, while others have gone even farther, arguing that mental states and cognitive processes themselves can extend beyond the borders of a person's physical, organic body.[16] Similarly, in our thought experiment, the most important difference between our old and new egalitarians is not something within the boundaries of their skin, but is rather in the wider social and cultural circumstances in which each is situated, as captured by the years to which each is respectively indexed. Indeed, one implication of the intuition we are pumping is that not all of the knowledge relevant to moral responsibility and exculpation need be "in the head" of the agent whose actions are being evaluated.

Even this externalist aspect of our view is somewhat traditional—it appears to be true of the kind of reasoning that applies to cases like the clueless student and the negligent doctor. In the second case, for instance, there was information about a new, life-saving procedure in her cultural environment, but she was just not aware of it; it was in the journals, clinics, and other medical operatives' heads,

[16] See Putnam (1975), Burge (1979), and Fodor (1987, 1994) for defenses of what has become known as passive or semantic externalism, Clark and Chalmers (1998) for the initial statement of the extended mind thesis and what has become known as active or vehicle externalism, and Dennett (2003), Clark (2007), Shapiro (2007), and Ismael (2007) for the development of similar ideas. Other approaches that have an externalist flavor emphasize different aspects of the extrabodily environment and the different roles they can play in human psychology, often creating new terminology to talk about them. For instance, see Doris (1998, 2002) for discussion of the underappreciated role of external situational factors, both physical and social, in driving behavior, and Merritt (2000) for a development of the core idea that emphasizes how properly structured environments can make a sustaining social contribution to ethical behavior. Sterelny (2003, 2012) extends the conceptual resources of niche construction theory to show how humans actively engineer the informational niches in which they live, learn, and raise children, and argues that this deliberate organizing of their own epistemic environment is a key factor in explaining human behavior and evolution. Defenders of gene culture coevolutionary theory stress the importance of social learning and the accumulation of cultural information, which is often contained in brains, but can also be manifest in behavior, realized in artifacts, written in books, and so on. The name of the theory indicates that it construes the repository of cultural information as an inheritance system that operates in tandem and interacts with the genetic inheritance system, and whose contents are subject to analogous kinds of selective pressures (see e.g. Richerson and Boyd (2005); Boyd and Richerson (2005); Henrich (2011)).

but it was not in *her* head. Had that information been absent not just from her head, but from the doctor's epistemic environment in general, however—if the new life-saving procedure had not yet been developed—then she would not have been to blame her patient's death. Again, we maintain that similar reasoning applies to the two egalitarians in our thought experiment.[17] Differences in knowledge can produce differences in responsibility and blameworthiness, and changes in knowledge can generate changes in responsibility and blameworthiness—even when those changes are in the informational content of the cultural and epistemic environment, rather than in the head of the agent being evaluated.

Perhaps somewhat oddly, the case involving implicit bias leads us to the position that not all of the knowledge relevant to moral responsibility and exculpation need be in the head of the agent whose actions are being evaluated, even when the *subject matter* of that knowledge is, in fact, in her very own head! One can take the slight whiff of paradox out of this if one considers a parallel case of knowing one's own cholesterol level or blood pressure. You might think that today, as a well-informed adult, part of taking responsibility for your own health and wellbeing is keeping track of these physiological features of your own body, and taking steps to keep them at acceptable levels.[18] But of course, this could not have been the case for, say, someone living in Shakespeare's day. It is dependent on advances in medical knowledge—a whole slew of discoveries made only in the last few centuries. No one innately grasps truths about cholesterol or blood pressure, nor knows intuitively that these are important indicators of health and should be monitored. Nor can anyone introspectively discern her own cholesterol level; you have to look without to gain that knowledge about yourself. For most of us this involves going to a doctor, who will use technologically sophisticated instruments to take measurements whose meaning she will report back to you. Moreover, no one can directly control her own blood pressure, or

[17] While our approach focuses on individual responsibility, we are also alert to the fact that the externalist orientation can inspire similar arguments about the responsibility that institutions have to take steps against bias and prejudice, and to structure the institutional environment of those individuals who operate within them. We briefly discuss this idea in Section 6, and also think that a more sustained look at the interaction of collective and individual responsibility is a worthwhile undertaking. Construing the project (of attempting to raise awareness, alter social norms, and reform institutions in the relevant ways) as a specific attempt at guided cultural evolution could provide useful insight into how best approach the task.

[18] Even today, this is only the case in some cultures, those with the readily available technology and properly disseminated medical information. The variability in cultures, and the resultant differences in what "taking responsibility for your own health and well being" amounts to, is compatible with the externalist orientation we favor, and the idea that differences in the informational environment and moral ecology can yield different kinds of responsibility on the individuals who inhabit them.

bring about an immediate and sustained change in it by direct act of will. To effectively control our blood pressure, most of us need to learn about and use the more roundabout, external methods that have been empirically verified. You will have to take slower, less direct steps to bring about the sought after change, even though the change is internal, and in your own physiological makeup and functioning. In these respects, our view is: as with cholesterol and blood pressure, so with implicit bias.[19]

6 A Glimpse Ahead: Objections and Open Questions

A series of further questions can be asked about comparative levels of responsibility. There are important questions not just about who is responsible, but also about how much responsibility we are ascribing to the different members of the hiring committee, and, for the two we did deem responsible, how blameworthy each is, respectively, and what form of punishment would be most appropriate (not to mention effective).[20] For the purposes of this paper, we have largely left our discussion at an intuitive level, and hope that, as such, it can be plugged into different, more sophisticated ways of answering such questions, and making the attendant notions more precise. For now, we will comment on a point that has been made to us a number of times: namely, that our approach seems to have assumed a qualitative notion of responsibility, but one that is comparative and graded as well; i.e. we say that the new egalitarian is less responsible than the explicit racist. Whether or not sense can be made of this way of construing responsibility ascription is an interesting question, but even if it turns out to be unworkable, we do not think it would threaten our main point. As far as we can see, everything we have said is also compatible with a view according to which responsibility is not graded, but rather a more binary, all-or-nothing notion—the

[19] Using Clark's (2007) useful terminology, we could make this point by saying that one needs to use *ecological control* to effectively influence one's own blood pressure and cholesterol levels; also see Holroyd and Kelly (forthcoming) for discussion of ecological control and implicit bias.

[20] While some alarmists worry that holding each other responsible for behaviors influenced by implicit bias is likely to provoke strongly negative, counterproductive reactions, and perhaps even an increase in biased tendencies, this is not always what happens. It turns out that some forms of finding and addressing fault can ultimately help to bring about more positive results. For example, the effects of interpersonal confrontation are much less straightforward than might be expected. In a series of papers, Alexander Czopp and colleagues have explored different aspects of the phenomenon, showing that confronting a person who expresses bias reduces that person's prejudicial behavior (Czopp et al., 2006). Even more intriguing, they also found that failing to confront prejudicial behavior that one witnesses can lead to an increase in one's own bias (Rasinski et al., 2013). Moreover, there could also be important individual differences here, with different techniques more likely to work on different individuals; for instance, in their reaction to and susceptibility to guilt.

assessment of an action in light of the relevant exculpating conditions will find the agent to be either responsible or not. The need for a more fine-grained spectrum of distinctions can still be met by graded notions of blame and blameworthiness, which are brought to bear only on actions to which the binary notion of responsibility applies. In these terms, the explicit racist and the new egalitarian are both equally responsible, but the former bears more blame than the latter, which might be reflected in a more severe form of punishment. Whether and how any of these qualitative, comparative notions can be made more precise by being interpreted in a quantitative framework are intriguing questions, but ones that we can fully address here (though see Kagan, 2012, for such a framework).

Another objection we have heard in presenting this material accepts our argument that the new egalitarian is responsible and blameworthy, but holds that we are being too easy on the old-school egalitarian. She is also responsible for the unjust job search because she too should have known better. Bias, prejudice and discrimination have been with us for a long time, goes the objection, and people have known about it for just as long. An observer of the human scene astute enough to appreciate prejudicial behavior for what it is would also notice that such behavior can be more or less overt. If the old-school egalitarian genuinely holds the values she professes to hold, she should have been alert to the possibility of prejudice covertly influencing her participation in the hiring decision, and taken steps to prevent it. In effect, this objection accuses us of overplaying the relevant differences between 1980 and 2014.[21]

We have no doubt that bias and discrimination have been a part of the human condition as long as there has been one, and that enlightened individuals have been able to recognize it and many of its forms as such. However, we also think that the advances in empirical psychology make a collective difference, and the sheer amount of evidence on implicit biases constitutes a crucial one. As captured by Bargh's invocation of the cognitive monster, there is a natural suspicion that these advances and evidence make a *moral* difference, that they have moral significance; indeed, we take ourselves to be showing one way to flesh out that suspicion. Someone living prior to 1980 could be sufficiently observant to discern that something funny and potentially biased was going on in the kind of hiring process we imagined. But she certainly could not have known, let alone been responsible for having known, anything terribly specific about the kinds of

[21] We are thankful to Alex Madva for pushing us to think about this; see Madva (this volume) and Brownstein and Madva (2012).

psychological processes that were driving it.[22] That implicit biases are not easily visible to folk psychology and the folk morality that depends on it, and that the details about them are so counterintuitive—they are so widespread, and can automatically influence a variety of behaviors and judgments, can co-exist with considered views to the contrary, are opaque to introspection and resistant to common forms of control—only strengthens our position on this. But now we do know, and that matters. There have been specific changes in the collective knowledge about *these* specific psychological processes, and changes in the collective knowledge of the possibility and likelihood that *these very psychological processes* will influence certain outcomes in *these very ways*, and that their influence can be mitigated with *these* kinds of methods but not *these*. More and more specific things are known about biases and the sources of prejudiced behavior now, and that, we maintain, should be reflected in shifts in how people should be held responsible for them. As our collective informational environment evolves in this way, our moral ecology should evolve with it.[23]

Of course, as impressive as the accumulated knowledge about implicit biases is, it has still not risen to the level of *common knowledge*, and it probably will not at any time in the immediate future. Different members of the population certainly have different levels of familiarity with the research on these unsettling facts. Judging from our own experience, people with the psychological profile of our new egalitarian (explicitly egalitarian, vaguely aware of the existence of implicit bias in general, but unaware of their own) remain common, and the percentage of people who have not heard of implicit bias at all is probably still quite high. Considering this observation in light of our evaluation of the members of the

[22] Another set of unsettling facts highlighted by recent empirical psychological work center on the normal human tendency to confabulate, and a susceptibility for believing self-serving rationalizations about what motivates our behaviors and judgments. Given this, we also think that asking someone like our old-school egalitarian to have enough clarity and self-awareness to see through her own rationalizations and glean what was actually happening in 1980, without benefit the empirical record, or even the clear idea of a mental state with a profile of characteristics like that of implicit biases, would have been asking unreasonably much of her, to put it lightly.

[23] We also see a fellow traveler in Miranda Fricker, and are sympathetic to much of her discussion about the relativism of blame (2010). She notes that "sometimes an agent maybe living at a time of transition," and that when we in the present are making retrospective judgments about agents who lived during such transitional periods in the past, there is a "form of critical moral judgment" we can use on those who are "slow to pick up on" the period's "dawning moral–epistemic innovation" that she calls *moral–epistemic disappointment*. Our old-school egalitarian seems to fit this bill quite well; he is on the cusp of an important transition in our understanding of the psychology of prejudice. As Madva emphasizes, suspicion that there is more to prejudice and bias than the full-blown explicit kind has been around for a long time, but as we emphasize, the accumulation of experimental evidence and detailed understanding of the specific character and operation of implicit biases has been a recent development. As such, we can say that one appropriate attitude we might take to our old-school egalitarian is that of moral–epistemic disappointment.

hiring committee can prompt the worry that our approach does something like penalize the wrong people: those admirably curious people who stay well informed enough to become alert to the existence of implicit biases become responsible for their own, and thereby open themselves up to blame if they fail to take appropriate measures to deal with them.

We have three things to say in response. First, our position is that now that knowledge about implicit bias is part of the informational environment, certain people can inherit the *responsibility to know* about it, and to know about and deal with their own. However, we are not suggesting that it is equally everyone's responsibility—not even everyone in 2014—to know about and take precautions again implicit biases. Responsibility to know about implicit bias, like responsibility to know about the contents of a syllabus or about advances in medical knowledge, is not (yet) distributed equally across all people in a population, nor across all agents in virtue of being agents. Rather, we think that as research about these unsettling facts and their unsavory influence mounts, the responsibility to take preventive measures against them accrues first, or at least more quickly and disproportionately, to occupiers of specific social roles. These include those involved in hiring decisions, obviously, but also teachers, social workers, and those in other "gatekeeper" positions whose activities can have the most amplified effects on various institutions and population level outcomes. Exactly which social roles this responsibility falls upon, and at what point during the accumulation of research on implicit bias and the evolution of the wider informational environment, is an important and interesting question. We take ourselves to have made progress in raising the question in this form, but do not yet have an answer to it.

Second, bracketing the idea of a role specific responsibility to know, we can better address the worry about "penalizing" the wrong people in virtue of what they actually do or do not know about implicit biases (rather than what they should know). In short, we are willing to bite the bullet on this point. Those who harbor implicit biases but have no knowledge of them (and do not occupy one of relevant social roles) can satisfy the knowledge-based exculpating condition without also meeting the exception clause. In such cases, they should not be held responsible for their actions, and are not blameworthy. In this sense, those who *are* in the know are indeed the first to inherit the "burden" of responsibility for their own implicit biases. This does not strike us as particularly surprising, or wrong. Someone has to lead the way, and the kinds of culturally sophisticated, ethically motivated, self-aware people who are likely to hear about implicit biases and take the time to find out if they harbor any themselves are exactly the type of people we would expect to be willing to act as examples for the rest.

This, however, brings us to our final point, which is just to emphasize that the situation continues to evolve. Indeed, those interested in progress need not just passively watch it happen, but can channel their efforts in informed ways to help guide the process. Of course, they can act as role models and vocal advocates in traditional ways. But the perspective we favor also shows the value of disseminating the psychological research, publicizing it, and making it part of the common knowledge of the population: raising awareness can raise the standards of responsibility as well. Activists can encourage people to take implicit association tests, and press institutions to require certain members to do so. As a greater segment of the population comes to know about their implicit biases, more and more people can be held responsible for dealing with them. As psychologists learn more, those interested in change can keep packaging the relevant parts of that research in ways that make it easiest for the uninitiated to understand, and continue pumping it out into the wider social environment. In this way we can continue in actively constructing the larger informational environment. In shaping our epistemic niche, we are also guiding evolution of the moral ecology that we all inhabit. One ideal to aim at is in engineering an environment in which everyone is expected to know about and deal with their own implicit biases—not because they occupy any specific social role, but because it is a responsibility one has simply in virtue of being a person.

7 In Place of a Conclusion

We certainly realize that there remains work to be done, but maintain that the ongoing project will benefit from being informed by the broadly externalist orientation we have been urging. Our main specific aim in this paper is to show that there are clear cases in which a person should be held responsible for behaviors influenced by implicit bias, even if she does not know she has the implicit bias. We have defended this claim by making explicit and illustrating patterns of moral reasoning that typically accompany currently existing norms and practices, and showing how they can be extended to deal with certain cases of behavior driven by these kinds of counterintuitive and unsettling psychological processes. The approach we have taken draws inspiration from the growing family of views emerging in philosophy of mind and the cognitive sciences united by their insistence that what is outside the head can be just as important as what is inside it. We have used all of these resources to map out one way that moral significance can be given to the accumulating empirical evidence about implicit biases.

We want to bring the discussion to a close by putting a positive twist on what could otherwise seem like a depressing or vindictive endeavor, as if our

motivation was primarily to find fault, identify targets for blame, and perhaps implying that they deserve punishment—that we are mainly out to get someone. To be sure, fully appreciating the potential power of implicit biases and the kinds of outcomes they help produce and sustain can be disheartening. But another, more uplifting way to look at the significance of the mounting empirical evidence and the consequent changes in the moral ecology that it can spark is in terms of the increases of freedom they might purchase. From this point of view, the trajectory of the changes in the larger epistemic environment on which we have been focusing (e.g. advances in empirical psychology from 1980 to 2014, and hopefully on into the future) have been largely in the direction of improvement, and provide reason for optimism. Not only do those changes constitute genuine *empirical progress*, but we hope to have begun to sketch out how they can be translated into *moral progress* as well. In better understanding, publicizing, and ultimately using what is being discovered about implicit biases and how to most effectively mitigate their influence, we can take more responsibility for ourselves, buying both freedom *from* the unwanted influence of these unsettling mental processes, and the freedom *to* be able to act in ways that we wish—ways that more closely fit with our considered ideals. Understanding our own minds can help us to take responsibility for ourselves more fully, and more often. Understood properly, empirical discoveries about implicit biases and other unsettling psychological facts can eventually, and perhaps unexpectedly, show us how to increase our agency, do better things, and be better people.

Acknowledgments

We would like to thank the following people for useful questions and feedback on earlier presentations and drafts of this material: Michael Brownstein, Luc Faucher, Jules Holroyd, Alex Madva, Chandra Sripada, the members of the Purdue Psychology Department Brown-bag talk series, Jennifer Saul and the attendees of the Implicit Bias and Philosophy workshops held at the University of Sheffield, Andreas de Block and the members of the Faculty of Business and Economics at KU Leuven, members of the Moral Psychology Research Group, Cathrine Felix and the members of the Filosofiska Föreningen at Lund University, and the audience members at the 2013 SSPP symposium on Implicit Bias and Moral Responsibility.

References

Banaji, M. and Greenwald, A. G. (2013). *Blindspot: Hidden Biases of Good People*. New York, NY: Delacorte Press.

Bargh, J. A. (1999). "The cognitive monster: The case against controllability of automatic stereotype effects." In Chaiken, S. and Trope, Y. (eds.), *Dual Process Theories in Social Psychology*. New York: Guilford Press: 361–82.

Bertrand, M. and Mullainathan, S. (2004). "Are Emily and Greg more employable than Lakisha and Jamal?: A field experiment on labor market and discrimination." *American Economic Review* 94(4): 991–1013.

Blair, I., Steiner, J. and Havranek, E. (2011). "Unconscious (implicit) bias and health disparities." *The Permanente Journal* 15(2): 71–8.

Boyd R. and Richerson, P. (2005). *The Origin and Evolution of Cultures*. New York: Oxford University Press.

Brownstein, M. and Madva, A. (2012). "The normativity of automaticity." *Mind and Language* 27(4): 410–34.

Brownstein, M. and Saul, J. (eds.) (2015). *Implicit Bias and Philosophy. Volume I: Metaphysics and Epistemology*. Oxford: Oxford University Press.

Burge, T. (1979). "Individualism and the mental." In French, P. A., Uehling, T. E., and Wettstein, H. K. (eds.), *Midwest Studies in Philosophy*, vol. 4: Minneapolis: MN: University of Minnesota Press: 73–121.

Carr, P. B. and Steele, C. M. (2010). "Stereotype threat affects financial decision making." *Psychological Science* 21(10): 1411–16.

Clark, A. (2007). "Soft selves and ecological control." In Spurrett, D., Ross, D., Kincaid, H., and Stephens, L. (eds.), *Distributed Cognition and the Will*. Cambridge, MA: MIT Press.

Clark, A. and Chalmers, D. J. (1998). "The extended mind." *Analysis* 58: 7–19.

Czopp, A., Monteith, M., and Mark, A. (2006). "Standing up for change: Reducing bias through interpersonal confrontation." *Journal of Personality and Social Psychology* 90(5): 784–803.

Dennett, D. (2003). *Freedom Evolves*. Penguin Books.

Dennett, D. (2013). *Intuition Pumps and Other Tools for Thinking*. New York, NY: W. W. Norton.

Doris, J. (1998). "Persons, situations, and virtue ethics." *Noûs* 32: 504–30.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. New York, NY: Cambridge University Press.

Doris, J. (2015) *Talking to Ourselves: Reflection, Ignorance, and Agency*. Oxford University Press.

Fischer, J. and Tognazzini, N. (2009). "The truth about tracing." *Noûs* 43: 531–56.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J. (1994). *The Elm and the Expert*. Cambridge, MA: MIT Press.

Fricker, M. (2010). "The relativism of blame and William's relativism of distance." *Aristotelian Society Supplementary Volume* 84(1): 151–77.

Haidt, J. (2006). *The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom*. New York, NY: Basic Books.

Haney-López, I. (2000). "Institutional racism: Judicial conduct and a new theory of racial discrimination." *The Yale Law Journal* 109(8): 1717–885.

Henrich, J. (2011). "A cultural species: How culture drove human evolution." *Psychological Science Agenda*. Science Brief: <http://www.apa.org/science/about/psa/2011/11/human-evolution.aspx>

Henrich, J., Heine, S., and Norenzayan, A. (2010). "The weirdest people in the world." *Behavioral and Brain Sciences* 33: 61–135.

Hirstein, W. (2005). *Brain Fiction: Self-Deception and the Riddle of Confabulation*. Cambridge, MA, and London: MIT Press.

Holroyd, J. (2012). "Responsibility for implicit bias." *Journal of Social Philosophy* 43(3): 274–306.

Holroyd, J. and Kelly, D. (forthcoming) "Implicit responsibility character and control." In Webber, J. and Masala, A. (eds.), *From Personality to Virtue*. Oxford: Oxford University Press.

Ismael, J. (2007). *The Situated Self*. Oxford: Oxford University Press.

Kagan, S. (2012). *The Geometry of Desert*. New York, NY: Oxford University Press.

Kawakami, K., Dovidio, J. F., and van Kamp, S. (2007). "The impact of näıve theories related to strategies to reduce biases and correction processes on the application of stereotypes." *Group Processes and Intergroup Relation* 10: 139–56.

Kelly, D. and Roedder, E. (2008). "'Racial cognition and the ethics of implicit bias." *Philosophy Compass* 3(3): 522–40. doi:10.1111/j.1747-9991.2008.00138.x

Machery, E., Faucher, L., and Kelly, D. (2010). "On the alleged inadequacies of psychological explanations of racism." *The Monist* 93(2): 228–55.

Madva, A. (ms.). "Biased against de-biasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice."

Merritt, M. (2000). "Virtue ethics and situationist personality psychology." *Ethical Theory and Moral Practice* 3: 365–83.

Nahmias, E. (2010). "Scientific challenges to free will." In O'Connor, T. and Sandis, C. (eds.), *A Companion to the Philosophy of Action.* New York, NY: Wiley–Blackwell.

Nisbett, R. E. and Wilson, T. D. (1977). "Telling more than we can know: verbal reports on mental processes." *Psychological Review* 84(3): 231–59.

Payne, B. K. (2005). "Conceptualizing control in social cognition: The role of automatic and controlled processes in misperceiving a weapon." *Journal of Personality Social Psychology* 81: 181–92.

Payne, B. K. (2006). "Weapon bias: Split-second decisions and unintended stereotyping." *Current Directions in Psychological Science* 15: 287–91.

Price, J. and Wolfers, J. (2010). "Racial discrimination among NBA referees." *Quarterly Journal of Economics* 125(4): 1859–87.

Putnam, Hilary (1975). "The meaning of meaning." In *Philosophical Papers, Vol. II: Mind, Language, and Reality*. Cambridge: Cambridge University Press: 215–71.

Rasinski, H., Geers, A., and Czopp, A. (2013). '"I guess what he said wasn't that bad': Dissonance in nonconfronting targets of prejudice." *Social Psychology Bulletin* 39(7): 856–69.

Richerson, P. and Boyd, R. (2005). *Not by Genes Alone*. Chicago, IL: University of Chicago Press.

Roskies, A. (2006). "Neuroscientific challenges to free will and responsibility." *Trends in Cognitive Science* 10(9): 419–23.

Saul, J. (2013.) "Implicit bias, stereotype threat and women in philosophy." In Jenkins, F. and Hutchison, K. (eds.), *Women in Philosophy: What Needs to Change*? New York, NY: Oxford University Press: 39–60.

Shapiro, L. (2007). "The embodied cognition research programme." *Philosophy Compass* 2(2): 338–46.

Shoemaker, D. (2011). "Attributability, answerability, and accountability: Toward a wider theory of moral responsibility.' *Ethics* 121(3): 602–32.

Smith, A. (2012). "Attributability, answerability, and accountability: In defense of a unified account." *Ethics* 122(3): 575–89.

Smith, H. (1983). "Culpable ignorance." *Philosophical Review* 92(4): 543–71

Smith, H. (2011) "Non-tracing cases of culpable ignorance." *Criminal Law and Philosophy* 5(2): 115–46.

Sommers, T. (2011). *Relative Justice: Cultural Diversity, Free Will, and Moral Responsibility*. Princeton, NJ: Princeton University Press.

Sripada, C. (forthcoming). "Self-expression: A deep self theory of moral responsibility." *Philosophical Studies*.

Staats, C. and Patton, C. (2013). *State of the Science: Implicit Bias Review 2013*. Columbus, OH: The Kirwan Institute.

Sterelny, K. (2003). *Thought in a Hostile World*. New York, NY: Blackwell.

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA: MIT Press.

Tavris, C. and Aronson, E. (2007). *Mistakes Were Made (But Not By Me)*. New York, NY: Harcourt.

Vargas, M. (2005). "The trouble with tracing." *Midwest Studies in Philosophy* 29(1): 269–91.

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Watson, G. (1996). "Two faces of responsibility." *Philosophical Topics* 24(2): 227–48.

Wilson, T. D. (2002). *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.