Implicit Bias, Responsibility, and Moral Ecology

Manuel R. Vargas
unfilosofo@gmail.com

Forthcoming in *Oxford Studies in Agency and Responsibility*.
10.07.16

ABSTRACT: Roughly, implicit bias is a partially unconscious and partially automatic (frequently negative) evaluative tendency directed at individuals, based on their apparent membership in a socially salient category or group. It is unclear what we should think about an agent's blameworthiness for actions produced in part by implicit biases, and there are reasons that weigh both in favor and against holding that such agents are blameworthy. There is also a more radical possibility lurking: implicit bias may reveal the limitations of a widespread conception of agency. That is, perhaps implicit bias (maybe along with various other results from the cognitive and neurosciences) reveals that our received views about agency are mistaken or confused in some important way. If so, then perhaps implicit bias is not merely some further phenomenon to which we can apply our pre-existing theories of moral responsibility and agency, but instead, a kind of challenge to those theories and the presumption that responsibility can be understood and characterized without appeal to context.

In response to the foregoing thoughts, there are two main questions this essay attempts to answer. First, are people morally responsible for actions that derive from their implicit biases? Second, is it possible to chart a middle way between the defense of common sense and the revolutionary import of phenomena like implicit bias that can sometimes suggest our received views of agency are mistaken? The view defended here is, respectively, sometimes yes, and yes. That is, there is an appealing way of thinking about the blameworthiness of actions caused by implicit bias that allows us to accommodate some of the radical aspects of the emerging scientific picture of agency, without entirely abandoning our commonsense picture of agency. The key is to recognize how a roughly "ecological" conception of moral agency can provide us with principled resources for distinguishing when agents are in circumstances that afford responsibility, and when they are not. On this approach, the status of social practices and norms is central for our being morally responsible.

KEYWORDS: implicit bias; moral responsibility; moral ecology; agency; moral psychology; blame; responsibility; capacities

1. Two pictures of responsible agency

Moral blame is a pervasive and familiar feature of our lives. It can arise in moments of great moral import—for example, in moral outrage at some public act of injustice—but also in more intimate moments, when a colleague neglects to have prepared for a meeting, a friend fails to display suitable empathy for one's trouble, or when one inwardly condemns the parenting of another.

The nature and import of these blameworthiness judgments, attitudes, and practices, are the traditional subjects of a theory of individual moral responsibility. Although philosophers have characterized the phenomena, structure, and normative foundations of moral responsibility in divergent ways, some presumptions are widely shared among contemporary theories.

For example, most accounts of moral responsibility maintain that an agent's responsibility for an action is grounded in two basic requirements. First, the action must have moral significance.[1] Second, the action must be suitably related to some internal feature of the agent, i.e., some bit of psychology arranged in this way rather than that, such that the agent identified with the action, or that it flows from the agent's values, or that it was a product of the agent's rational or normative capacities. That is, these accounts specify a set of capacities or states, the possession of which is settled entirely by features of the agent. If the considered action has a moral valence, and the agent has those internal features specified by the account, then we can settle questions about whether the agent was blameworthy in so acting. Call such accounts *atomistic*, because they treat responsibility for action as settled by intrinsic, or at least internal, features of agents.

Beyond atomistic presumptions, typical approaches to moral responsibility also share a methodological commitment. On these accounts, philosophical theorizing begins with the phenomena of our agency, and the reality of our everyday practices. We then reason about these

[1] Some accounts do away with the requirement that actions have moral valence. On Fischer and Ravizza's (1998) account, for example, one can be morally responsible for acts without moral significance. For reasons to resist this characterization, see Vargas (2013a: 307-09).

things. Sometimes, this process involves generating principles that explain responsibility judgments and practices. Other times we ask what it would take to justify our received views, and then we attempt then construct accounts that provide some basis for thinking that our practices can be justified. In either case, the primary metaphysical or normative toil involves checking for intuitive fit, addressing counterexamples, and adjusting principles in light of various identifiable pressures. Thus, most accounts accept that the ordinary image of our agency provides both the initial starting point, and the recurring measure of our normative and metaphysical speculations.

This happy state of limited (but genuine) philosophical consensus about presumptions and methodological starting points is threatened. Perhaps the most pressing difficulty arises from approaches to agency that grow out of scientific research on agency. The large and varied body of results from diverse experimental sciences suggests a picture of human agency that is vastly more complicated and puzzling than one typically finds in philosophical accounts of responsible agency. In the empirically studied kingdom of agency, sightings of the conscious, deliberative, plan-making rational agent of philosophical accounts are relatively rare. Experimental work suggests that a good deal of action is automatic, detached from conscious deliberative exercises of a self-aware psychology. With alarming frequency, what we do, and why we do it, turns out to be disconnected from the kinds of reasons we cite for our behavior (Doris 2002; Nahmias 2007; Nelkin 2005; Vargas 2013c).

On these matters, there is sometimes an impulse to assimilate the various scientific threats under the old rubric of debates about free will and determinism. Yet, these worries need not be a matter of determinism. The idea of a free and independent self, a self in control of what it does, is threatened by the fact that our choices are structured by institutions and forces over which, at best, we have extraordinarily limited influence. Moreover, the ways in which we respond to those structured choices are themselves products of histories and institutions. Even if these forces do not operate on us in a strictly deterministic fashion, they nevertheless suggest a kind of fragility or porousness to our agency that does not easily cohere with the ordinary image of agency.[2]

Perhaps it is not an oversimplification to say that we have a conflict between, on the one hand, philosophical conceptions of agency and responsibility that have sought to vindicate much of our everyday conception of human agency, and on the other hand, a considerably more skeptical, revolutionary picture of agency arising (in part) from work in psychology and neuroscience.[3]

In this context, implicit bias is an exceptionally interesting phenomenon. As a first approximation, implicit bias is a partially unconscious and partially automatic (frequently negative) evaluative tendency directed at individuals, based on their apparent membership in a socially salient category or group (Cf. Brownstein 2016). Apart from the obvious challenges it raises for our understanding of our own motives, and the injustices that may arise in the wake of invisible biases, what makes implicit bias exceptionally interesting for theories of agency and moral responsibility, is that aspects of implicit bias and our reaction to it provide reasons for blaming and not blaming biased agents. On the one hand, implicit biases can appear to be largely outside the direct control of agents, and not expressive of their values or real self. That many people remain unaware that such bias is possible seems to only exacerbate the problem of direct control. On the other hand, first personal attitudes suggest a different picture. It is hard to shake the sense that my discovery of some particularly odious bias in myself would and should

[2] This thought is at the root of many arguments for responsibility skepticism, both philosophical and scientific. See, for example, Galen's Strawson's (1994) Basic Argument, Derk Pereboom's (2001) Four Case Argument, but also the "Boys from Brazil" idea in Greene and Cohen (2004).
[3] For examples of the chorus of skeptical voices drawing the conclusion that various strands of scientific work overturn conventional and philosophical understandings of agency and responsibility, see Cashmore (2010), Montague (2008), Wegner (2002), Pockett (2013), and Bargh (2008). I have replied to some of these concerns in Vargas (2013b).

make me feel guilty. Moreover, the plausibility of some amount of indirect control over our biases may suggest some reason, grounded in moral improvement, to hold blameworthy those acts caused by bias.

Perhaps more interestingly, the challenge presented by implicit bias does not solely depend on the ordinary image of our responsible agency. Suppose we are tempted by the thought that implicit bias reveals the limitations of a naive, pre-philosophical conception of agency. Were we to conclude that implicit bias is one more reason to doubt the ordinary image of our agency, we might still wonder if something can be said on behalf of our responsibility practices.

So, there are two questions I attempt to answer in this essay. First, are people morally responsible for actions that derive from their implicit biases? Second, can we chart a middle way between the defense of common sense and the revolutionary import of the scientific picture of agency? I argue that there is an appealing way of thinking about the blameworthiness of actions caused by implicit bias that allows us to accommodate some of the radical aspects of the scientific picture of agency, without entirely abandoning our commonsense picture of agency.

2. The challenge of implicit bias

The precise contours of what we call implicit bias remains a matter of ongoing empirical investigation. However, it is plausible that the phenomenon of implicit bias involves a diverse set of psychological mechanisms with distinct functional profiles.

For example, some forms of bias may operate on primarily semantic associations, others will operate primarily through affective phenomena (Holroyd and Sweetman 2016; Levy 2014; Levy 2015). Moreover, the probes used to disentangle the difference between *awareness* of a stereotype (or stereotype activation) and *exercise* of that stereotype (stereotype application) remains a matter under development (Krieglmeyer and Sherman 2012).[4] Additionally, there may be variation in the awareness and degree of control agents may have over their biases and expressions of biases. How much variation there is across and within agents, and whether there are especially self-aware or especially skilled self-regulators remains largely unknown. Nevertheless, clear, ready and reliable first-personal occurrent awareness of one's own biases does not regularly seem to be the case.

Despite the evolving state of the empirical literature, there is an obvious reason to think carefully about its implications. If bias-induced action is unwarranted and unjust, then this requires our attention and deserves our efforts to ameliorate its effects. Indeed, one of the most pernicious features of implicit bias is that most people do not even know to look for it. If, in addition, we lack an adequate account about whether bias-caused action is morally blameworthy, it becomes more difficult to know what to say when we encounter it. Thus, absent an account of blameworthiness, an important aspect of the moral significance of implicit bias remains unclear.

My interest is in responsibility for bias-caused action, and what follows is largely silent on whether agents can be responsible for just the bare having of biased attitudes. For present purposes, I will work with a relatively coarse-grained conception of implicit bias, one that

[4] The Implicit Association Task has been something of a lightning rod for dissatisfaction about empirical research on implicit bias (Blanton et al. 2009; Greenwald et al. 2009; Jost et al. 2009; Oswald et al. 2013; Greenwald et al. 2015; Oswald et al. 2015). A representative concern can be found in Levitin (2013), who notes that the Greenwald et al.'s own putatively pro-IAT metaanalysis finds that the IAT was weakly correlated with other measures, failing to account for more than 93% of the data in a review of 184 independent samples covering 15,000 experimental subjects. My sense is that qualms about the IAT and the extent to which it identifies a phenomenon that has significant behavior consequences are unlikely to go away soon. Nevertheless, the more general finding that people can have biases of which they are plausibly unaware and/or that at least sometimes they are unaware of these biases affecting behavior seems to have good support. For discussion of some of the behavioral data that favors this reading of the empirical data, see Saul (2013b) and Brownstein (2016).

encompasses biased valuations, stereotypes, and attitudes. An important idea in what follows is that the moral and epistemic significance of psychological phenomena will tend to cut things in places other than the psychological joints. I will also assume that the challenge is not just that we are unaware of bias. Rather, there is a web of issues here about moral culpability where many people are unaware of the bias *and* it is unclear whether and when some bit of action is a product of bias, even for those aware of the possibility of implicit bias in their own case.

A related issue concerns the degree to which a given instance of action is bias-affected. Sometimes, the formation of a particular intention to act will be partly affected by bias at the moment of its formulation. Other times, bias might leak into the details of how some general action plan (itself conceived of without bias) comes to be filled in. So, it is natural to think that culpability for bias-affected action will oftentimes be a matter of degree (Nelkin 2016), rather than blame full-stop. Finally, my focus here is on the form of responsibility that is sometimes characterized as *accountability*, a kind of responsibility that implies culpability and appropriateness as a target of the reactive attitudes (Watson 1996; Shoemaker 2011).

The standard way to investigate responsibility for bias-caused action brings to mind the metaphor of a toaster. The toaster is some characterization of moral responsibility (whether philosophical or intuitive). We take some bread—the phenomenon of implicit bias—and then put it in. We wait a few minutes, and out pops the toast, delivering a verdict on the phenomenon.[5]

[5] So, for example, one prima facie case for thinking that one is not responsible is that, on first blush, implicit bias doesn't satisfy standard conditions on responsibility, including conditions of awareness and control— e.g., Saul (2013a) Washington and Kelly (2016) and in some passages, Kelly and Roedder (2008). However, there is an alternative case to be made in the opposite direction, in support of the claim that one can be responsible for implicit bias. First, one can accept the awareness control-focused account of responsibility and insist that biased action satisfies those conditions—see, for example, Madva (In preparation)—or one can argue that the relevant sense of awareness can be satisfied in many cases of implicit bias (Holroyd 2012; Holroyd 2015). Second, one can argue to responsibility for biased action on the basis of responsibility for non-volitional actions. The operative idea there is that, as a matter of our actual

We have learned a good deal from these discussions. Nevertheless, such approaches risk underplaying the deeper philosophical challenge of implicit bias, and in particular, its significance in the context of larger challenges to our ordinary image of agency. When we apply theories of responsibility that have been weighed and measured in light of our ordinary image of agency, we presume that such accounts are in good standing. This presumption is dubious because the phenomenon of implicit bias, and the picture of agency it implies, call into question the picture of agency that funds traditional accounts.

Above, I gestured at one concern with the presumption. That is, philosophical accounts of responsibility—especially those typically invoked in accounts of culpability for implicit bias-affected action—have typically been generated in relative isolation from the empirical data on moral psychology. Yet, the empirical literature on agency and psychology has suggested that human psychology is importantly different than it appears to us. Implicit bias looks like one of many ways in which our agency is importantly disunified. Conscious deliberation plays a smaller role and oftentimes different role than we tend to think (Doris 2015a; Arpaly 2003; Doris 2002; Wegner 2002; Nichols 2015; Nahmias 2007). What we regard as an agent's values seems to be subject to framing effects in the context of evaluation (Knobe and Roedder 2009; Doris 2015a). We exaggerate our independence from context, not recognizing the role of influences that are not readily apparent to us, and that we would even disavow or regard as deliberatively irrelevant

practices, we tend to find fault in people when they give evidence of a wide range of non-volitional behaviors, including negligence, bad characters, and failures of appropriate reaction (Brownstein 2015; also suggested in Kelly and Roedder 2008: 527). If implicit bias is structurally similar to culpable, non-volitional shortcomings, then one might conclude that we can be responsible for implicit bias, even if it doesn't have the control features we paradigmatically associate with responsibility. There is a third route, recently pursued in recent work by Joshua Glasgow (2016). Rather than proceeding in a top-down fashion of fitting theories together, Glasgow aims to build a partial account of responsibility for implicit bias by starting with the thought that we can feel guilt or shame upon discovering that we are afflicted with biases. The issue then is how to accommodate that thought.

(Doris 2002; Doris 2015a; Huebner 2016). Finally, there is little reason to think that we have robust, cross-situationally stable capacities (whether for endorsement, acting from values, or from reasons) that figure in theories of responsibility (Doris and Murphy 2007; Nelkin 2005; Vargas 2013c). Instead, what we have are psychological dispositions surprisingly circumscribed by context (Vargas 2013a).

The more global concern here is that the current state of psychological research on human agency suggests a much more contextually and socially embedded form of agency than is presupposed by going accounts of moral responsibility. So, it seems at least prima facie worrisome to evaluate implicit bias in light of a theory of moral responsibility the foundations of which seem threatened by the larger body of psychological research of which implicit bias is a part.

Most of the philosophical literature on moral responsibility has not seriously engaged with these concerns. (Although, thankfully, the tide may be starting to turn.) If the foregoing is correct, though, on one way of reading the significance of implicit bias, its main upshot is that it lends further weight to the view that standard accounts of responsibility are in want of an empirically adequate account of agency. That is, beyond raising puzzles about the blameworthiness of bias-caused action, the deeper challenge of implicit bias is that it suggests that a mistaken picture of agency lurks at the heart of standard accounts of responsibility.

Once we allow the empirical concerns to get purchase, various conceptual puzzles begin to seem more pressing. For example, given that our behavioral dispositions are often situationally fragile, we might want to know how to accommodate this fact in our accounts of the capacities of responsible agents. A natural solution is to decide that capacity talk about agency displays a certain degree of interest-relativity. That is, whether a certain degree of musculo-skeletal precision counts as being control depends on whether we are performing surgery or casually playing a video game. If that is right, at least some everyday notions of capacity talk flout the atomistic presumptions built into standard accounts of responsible agency. The idea that an occupation, activity, or practice—all things external to a given agent—can contribute to the truth of capacity ascriptions hint that something may be wrong with atomistic assumptions for moral responsibility.

These varied thoughts—the emerging psychological picture of ourselves, the surprising contours of our dispositions, and the interest-relativity of capacity talk—seem to have no natural home in conventional, atomistic accounts of responsibility. This is not to say that conventional accounts have nothing to offer, or that they cannot be adapted to accommodate these phenomena. On the contrary, any adequate account of responsibility will almost surely draw from the resources of existing conventional accounts. The point here is about the larger dialectic situation. If we seek to account for responsibility in domains where our agency turns out to be radically different than we have supposed, we should be cautious of appealing to theories that proceed on the assumption that our agency is roughly the way we thought before the 20th century flowering of the sciences of the mind.

Ideally, we would have an empirically plausible, normatively adequate picture of agency that takes seriously the circumstance-dependent nature of our abilities and the possibility of interest-relative conceptions of agential capacities. In the next section, I offer a picture of such an account.

3. Outline of a theory of moral responsibility

There are distinct forms of evaluation we can adopt with respect to agents. We can, for example, "grade" the quality of *actions* in light of some or another standard, without imputing fault or culpability in an agent. Similarly, we can use evaluative language, even thick normative terms, to characterize the dispositions or qualities of *agents*, without imputing blame. For example, I can refer to both a pet and a person as aggressive or gentle without thereby making an assessment of culpability in either case. To be sure, "aggressive" is not devoid of moral and practical significance; in either case, though, the assessment of blame is something further. Depending on our aspirations or given the norms we suggest, these not-necessarily-culpability-imputing judgments can have special interpersonal significance. Where agents seem capable of suppressing or enhancing those interpersonally significant traits, we have reason to encourage those adjustments. Importantly, though, one can make and accept these assessments while simultaneously rejecting the idea that anyone merits blame. That is, we can make these axiological assessments of actions and agents without inferring the aptness of blame in the sense at stake here.

Aretaic (or characterological) assessments and more broadly axiological evaluations are distinct from robustly culpability-imputing judgments and attitudes. This latter class of evaluations are the stuff of moral responsibility in perhaps the most familiar usage. The present account is focused on moral responsibility in the sense of blameworthiness. That is, when agents are morally responsible, they are (other things equal) worthy of moralized blame, of the sort that is paradigmatically invoked when we condemn someone for their actions.[6]

[6] Recall that my concern here is with responsibility, of the *accountability* variety, i.e., on moralized blaming of the sort that characteristically involves the reactive attitudes and ascriptions of culpability. (I'm putting to the side the question of whether moralized praising plays the analogous positive notion; and, admittedly, some have thought that notion of blameworthiness itself decomposes into distinct varieties, associated with distinct attitudes—see, for example, Erin Kelly (2009) on the requirements for the distinctively retributive attitudes.) In focusing on accountability, I do not mean to take a stand on whether there are questions about implicit bias that are usefully pursued in connection with a concern for *attributability*, *answerability*, and other notions including *strong self-governed agency*, *autonomous agency*, and *authentic agency*. I do worry that sometimes the terminology is not consistently deployed—for example, some understandings of

By blame, I mean two things: (1) a judgment or judgment-like attitude, and (2) a class of characteristic reactions. Blame *judgments* are judgments of a pro tanto license to express a class of characteristic interpersonal reactions. This is blame in the largely cognitive mode. There is also the affective mode of blame, what I call blaming *reactions* which include both the Strawsonian reactive attitudes and the expression of those attitudes in characteristic ways. Those expressions range from calls for censure or condemnation, to retractions of interpersonal warmth and proximity (Vargas 2013a, 116-21).

The present challenge is to explain how agents can be worthy of blame judgments and blaming reactions, in a way that does not rely upon a commitment to atomism about agency, or that does not put the justification of blaming at odds with a broadly scientific view of agency. I call the proposal that follows the *agency cultivation model* of responsibility.

The agency cultivation model has two features that are especially significant in the present context: a distinctive account of the social and normative function of the responsibility practices, and a novel account of the capacities required for responsible agency.

The first idea is that the social and normative function of responsibility practices is best understood as enabling the functioning of a particular form of self-governance. Roughly, the idea is that we can ground a normative account of responsibility (including an account of the justification of responsibility practices) in terms of the effects of those practices on an agent's ability to self-govern in light of moral considerations. This is the idea of *social self-governance*.

"responsibility as attributability" strikes me as ambiguous between a distinct *variety* of responsibility and a distinct package of commitments about the requirements on some independently specifiable notion (such as blameworthiness). However, Shoemaker (2011) has done a nice job of isolating some of the varied notions around which the literature has clustered; see also detailed taxonomies by Fischer and Tognazinni (2011) and Vincent (2011). It is an open question how far the ideas pursued here would go through on different target notion of responsibility and agency.

Responsibility practices derive part of their justification from the manner in which they aid individual agents in self-governance, or roughly, from the way they help us become better beings.

The second idea is a *circumstantialist* picture of the capacities required for moral responsibility. The idea here is that the capacities required for responsibility are higher-order capacities that are distinct from the physiological or metaphysically basic abilities of agent. To make sense of this idea, I appeal to Sidgwickian capacities, that is, a construal of the required capacities that is based on the normative function of the responsibility practices.

So, the first part of what follows provides an account of, roughly, the modestly teleological element of responsibility and how it structures responsibility norms, and the second part provides an account of the agential requirements on responsibility.

I'll start with social self-governance and the normative foundations of responsibility.

Social self-governance

As agents, we have a deep and abiding interest in distinguishing between those domains or circumstances where our agency is reliable and competent in its functioning, and those where it is not. Typically, we mark this distinction in terms of some domain of behavior being in or out of the agent's control. However, we use a range of terms to pick out this idea, including talk of competence, mastery, or maturity in this or some other context.[7] These spheres of reliable agency, however we characterize them, need not be static. The typical case of a life is one marked by a mix of slowly expanding and retreating domains of greater and lesser control.[8]

[7] Cf. Joseph Raz (2011) on responsibility, although he characterizes the notion in terms of "domains of secure competence" rather than domains of control. For discussion, see Vargas (2014).

[8] Being subject to blame, social estrangement, and the like, has an educational function, in both early developing agents and later, more sophisticated agents. Our ability to learn about both the good and the right, in part, by the moralized feedback we receive from our ways of manifesting

In some domains it is not entirely up to me, or even a matter of my intrinsic features, whether my agency is competent or reliable. Some activities, roles, and the domains of their operation are incorrigibly social in their nature. Whether I am reliably competent at being a good scholar, a good union member, a politically engaged citizen, an upstanding member of a religious community, and so on, is partly a matter of what the shared construal of those roles comes to.

Which domain I am interested in being competent at may not be the same domains of interest to you. However, both of us typically have an investment in being viewed as competent in a wide range of domains with social significance. And, ordinarily, the most reliable way to be viewed as competent is to in fact be competent in that sphere of activity. Most of us, most of the time, aspire to be reliably capable of control in the ways that shared cooperative living presumes.

A familiar and powerful way to characterize the idea of agential control is in terms of responsiveness to reasons. Of course, we are not perfectly rational agents. Our capacities to recognize and respond to one set of considerations do not seem to guarantee a comparable ability across another class of reasons.[9] Moreover, we sometimes seem to respond to reasons without being conscious of it, or conceiving of things in that way (Arpaly 2003). Nevertheless, it is

our agency in the world. This is typically, but not exclusively through the vehicle of our choices. In childhood, judgment of culpability and the expression of characteristic blaming reactions are sometimes feigned until the child has a baseline of moral knowledge and the requisite dispositions to use that knowledge seem present. After that point, judgments of culpability have reasonably identifiable conditions for appropriateness, i.e., that the agent can recognize and respond to moral considerations, and that in the considered circumstances, the agent has violated some standard or demand of moral significance.

[9] Varieties of reasons here should be understood in a largely nominal sense, delineating varieties only according to familiar or intuitive social, practical, and theoretical kinds. This characterization is intended to be neutral on more foundational questions about the ontology of reasons.

plausible that we respond to reasons or considerations of various kinds, at least sometimes.[10]

We can be better and worse at recognizing reasons. Not everyone is maximally good at recognizing the force of, say, considerations rooted in mathematics, sartorial achievement, or pub game contests. However, there are classes of considerations about which we demand that others widely exhibit some threshold of ability to recognize and respond. In particular, responsiveness to moral reasons or moral considerations looms large in our practices—indeed, it is ordinarily a prerequisite for being rightly subject to moral blame.[11]

Thus, domains of reliable control are to be understood in terms of the capacity to recognize and respond to reasons. Recognizing and responding to moral reasons looks like an important part of responsibility practices and social life. Here, though, a justification question looms large. Our practices of moral blame—responsibility practices, in short—are not *just* "ways we do things around here." What these practices, judgments, and attitudes express is, among other things, a demand that agents conduct themselves in some ways and not others. Merely arbitrary and parochial norms would be inadequate grounds for meting out serious social penalties—in some cases, even death. Moreover, such norms would be poor grounds on which to take seriously a minimal threshold of your interests and values across the entirety of social

[10] Even the usually contentious partisans of the free will debates seem to agree on this much—humans evidently have the capacity to recognize and respond to reasons, and this is so even if determinism is true. While there is disagreement concerning how best to characterize the power of being able to recognize and respond to reasons, whether we have it in greater or lesser frequency, and whether such a power is sufficient to support our responsibility characteristic practices and/or ascriptions of free will, there is little disagreement with the claim that we at least sometimes have this power. To be sure, there is serious disagreements concerning how to understand what reasons are, and the proper characterization of the relationship between reasons, reasoning, and our affective or emotional states and dispositions. For present purposes, however, little turns on those matters.

[11] The "ordinarily" bit of this formulation is intended to distinguish cases of original responsibility, where control is plausibly required, from cases of derivative responsibility, where occurrent control may be absent.

space.[12]

So, if responsibility practices are to be what they present themselves as being, they must be *justified*. We cannot simply find ourselves with them, for their normative import requires justification if we are to retain them.[13] So here is a conjecture. When we hold moral considerations-responsive agents responsible (minimally, when we evaluate them in culpability-entailing ways) we participate in a system of practices, atttitudes, and judgments that support and improve our responsiveness to moral considerations.

On the present account, responsibility characteristic practices are properly directed only at responsible agents, i.e., agents that have a certain threshold of ability to recognize and respond to moral considerations. What praise and blame do (and what our responsibility practices collectively do), is (over time) sustain and further develop those capacities to recognize and respond to moral considerations. If that is right, we have an account of the justification of moral responsibility practices.

To be sure, the *norms* governing responsibility practices do not themselves have an instrumentalist content. Instead, they are better understood along familiar Strawsonian lines, i.e.,

[12] Were we to perceive responsibility norms as arbitrary and parochial, our concern for responsibility would be closer to norms governing sports fandom. That is, our assessments might be locally significant, and sometimes license for a kind of exhortation and evaluative assessment. However, they would remain entirely optional, and vulnerable to easy substitution. That is, we might replace an interest in the 49ers with an interest in the Raiders, were the 49ers to leave the San Francisco Bay Area for less excellent environs, or if the team was struck by irreparable calamity. In contrast, our concern for responsibility does not seem to be like that at all. Our concern for whether someone is responsible or not cannot be substituted for an interest in etiquette, or simply dissolved if the passion fades. Of course, dedicated fans of sport might disagree. Such disagreement, however, seems more likely a way of expressing the thought that their fandom does have a greater than parochial and random significance; it smacks of moral conviction.

[13] Our psychological dispositions are not unimportant. Our practices are shaped by our psychological dispositions, our interests, and the things we find ourselves regarding as morally salient. Those psychological dispositions provide a kind of constraint, internal to the practices, or what patterns of behavior we can demand and reasonably expect adherence to.

as demands for due moral concern. On this account, what makes someone responsible is that they have certain capacities, and that they have violated a norm of responsibility. Our moral outrage is triggered by that—the violation of a norm we have internalized and accepted. So, this is not a picture according to which agents understand themselves to be trying to influence other agents. Instead, it is a two-tiered account, where the content of the norms ("act with due moral concern") make no appeal to effects, but the justification for continuing to enact those norms does appeal to the effects (Cf. Rawls 1955; Hart 1959). As such, it permits retrospective, desert-entailing retributive content in ordinary judgments (Doris 2015b; Vargas 2015).

We now have several interconnecting pieces: (a) the idea that control matters for agents; (b) that control can be understood in terms of the capacity to suitably recognize and respond to reasons; (c) that such capacities, especially with respect to moral considerations, are of special importance in our social lives; (d) that the capacity to recognize and suitably respond to moral considerations is ordinarily a necessary requirement on being subject to moral blame; and (e) that practices of moral blame are such that our participation in them enhances our individual and collective abilities to recognize and respond to moral considerations.

To some, the instrumentalist cast of the account thus far will be grounds for objecting. Here is one way of giving voice to that concern: this account offers mere expediency for genuine responsibility. Whatever this is an account of, the critic might say, it cannot be an account of responsibility. As useful as it might be to praise or blame agents, the issue is whether people can be rightly judged to deserve these moralized reactions.

There is some force to this objection, but the account is incomplete. Recall, again, the social aspect of responsibility practices. Blame (at least in the Western, post-industrial world) ordinarily entails a loss of social status, and separates us emotionally from our communities. In creatures like us, acceptance of blame is typically connected to the offending agent experiencing guilt. The experience of that guilt provides a powerful (but not infallible) motivational impetus to moral repair, at least in agents concerned with either moral demands or social standing. If I go unblamed, it is harder for me to experience the guilt that motivates moral self-improvement. Correlatively, it is more difficult to undertake moral repair with those I have wronged (Bennett 2002; McKenna 2012; Vargas 2013a: 261-66).

This picture can accommodate the familiar idea that blame must be deserved. Blame is deserved when an offending agent is the right kind of agent (a responsible agent, i.e., an agent with capacities above the relevant Sidgwickian thresholds), and that agent has violated some relevant moral norm. This picture captures the phenomenological character of ordinary, first-order, potentially retributive, desert-entailing judgments about blame. But unlike conventional accounts, the explanation for *why* we have reason to be in the business of making those judgments, and what that system of practices achieves for us, defers to a further normative basis. Desert judgments earn their keep, normatively speaking, in light of the role that such judgments have in a practice that supports moral considerations-sensitive agency. However, the reason one has for making a given desert judgment is always, at the first order, about the agent being the right sort of agent and having evinced some failure of due moral concern.

An important difference between this account and simple-minded consequentialist justifications of moral blame is that the basis of blaming in a particular case on the agency cultivation model is not grounded in the expedience, utility, or the maximization of pro-social behaviors of blaming of so blaming in that case. On the present account, blameworthiness and responsible agency are not settled by a local assessment of whether it is useful to hold an offending agent responsible in this case. Rather, it is a matter of specific statuses determined within the practice.

To be sure, the practice gets its deepest normative basis from the way it supports and sustains forms of agency that are valuable to us both as individual agents and as communities of agents. So, there is a teleological structure to the normative foundations of a practice that frequently presents itself as backward-looking. However, the familiar backward-looking surface structure of the normative practice is retained. Blame is deserved in cases where the offender is the right sort of agent and has failed to act with the kind of moral concern we are justified in demanding. So, if one wishes to object to the teleological element of the account, it cannot be on the basis that it fails to capture the general features of our ordinary responsibility practices. Instead, the account must be rejected on subtler grounds having to do with the reasons for favoring or disfavoring teleological justifications of norms and normative practices more generally. On this matter, it is not difficult to find vigorous defenses of the basic approach (Cf. Parfit 2011: 371-403; Hooker 2016).

A related but distinct worry is that this approach is *merely pragmatic*, or ultimately prudential, rather than moral. I take it that this is best understood as a worry about the kind of consideration that rationalizes the norms of moral responsibility. The worry here is that talk of social feedback, the utility of condemnation, and so on, is not a moral reason—i.e., a reason connected to the nature of the offending agent and the offense—for thereby holding such agents to account.

First, I want to resist this characterization. It runs together the distinction between the proximal reasons we have for blaming (namely, that there is a violation of due moral concern by an agent of the relevant sort) and the further story about why those proximal reasons are the proximal reasons we have. At the first order, the reasons we have for blaming are clearly moral: they are concerned with whether someone is a responsible agent and whether they have acted in a way disfavored by justified moral norms. This is not a pragmatic matter, grounded in self-interest or prudence. Rather, it concerns what morality demands of responsible agents.

Second, the worry about my proposal being "merely pragmatic" may tacitly reflect a widespread but tendentious picture of moral agency. Suppose that one starts from the presumption that we have all the tools we need for navigating the moral world, and these tools are already part of the intrinsic powers of a mature mind considered in itself. If this is your picture, then my emphasis on the social scaffolding of our moral agency will appear to be, at best, a bit of gratuitous outrigging to what is morally central about responsibility and responsible agency. If so, then the machinery that is central to my account—including the mechanisms of social feedback and their justification—cannot be central parts of the moral basis of responsibility.

This social outrigging is hardly peripheral, however, on a different picture of our agency. On the account I am proposing, we start with the idea that our agency is socially embedded—structured, constrained, and influenced—in ways that are both fundamental and ineradicable to who and what we are. Given this picture, the crucial insight is that responsibility practices provide us with a regular and reliable feedback loop without which a morally significant form of self-governance would be impossible. To disparage the social scaffolding of our agency as a merely pragmatic addendum to moral responsibility is to fail to appreciate the kinds of creatures we are. It is not just that this feedback is convenient for us. Rather, it is the way we learn moral norms, acquire moral knowledge, and the way we develop dispositions of seeing moral considerations, we *require* this feedback. Absent these forms of communal moral shaping, we cannot be the sorts of agents we aspire to be. Making this picture plausible is too much for a theory to do by itself. However, the resources this account provides for thinking about phenomena like implicit bias should be regarded as part of a larger argument for what is powerful and instructive in taking the social scaffolding of our agency as central to responsible agency.[14]

Suppose one grants all of the foregoing. What the account still needs, however, is some story about what it means to have a capacity to recognize and respond to moral considerations. Without it, the account founders where so many others have, i.e., on the vexed nature of the capacities required for moral responsibility.

Circumstantialism

Thus far, I have argued that we can understand the normative foundations of responsibility in terms of the effects that responsibility practices, attitudes, and judgments have in enhancing our capacity to recognize and respond to moral considerations.

If we are to arrive at a satisfactory notion of capacity, the first thing to note is that capacity talk is strongly interest-sensitive. Whether I am capable of juggling a soccer ball 15 consecutive times depends, in part, on your interest in asking, and what background assumptions we take to be operative in the question. For instance, whether my being asleep matters for my ability to juggle depends on whether you are asking about my suitability for providing a juggling demonstration next week, or whether instead the interest is in my entertaining you at that exact moment. In the former case, my sleeping state is no deterrent to my being able to juggle the soccer ball, and in the latter, it is.

A metaphysics of capacity does well to be able to capture this feature of ordinary discourse. One way to do this is to allow that the metaphysical facts of my juggling capacity are given, in no small part, by our interests. There may well be important notions of capacity that are not interest-sensitive in this way. Nevertheless, the idea here is that agential capacities, especially

[14] For more on these ideas, see Vargas (forthcoming).

in connection with moral concerns, are very much sensitive in this fashion. What we need, however, is an account of *whose* interests are determinative—and why those interests and not some others.

At this point, the idea of a *Sidgwickian capacity* is helpful. A Sidgwickian capacity is a capacity that is identified by an ideal observer with some specified interest or interests. For present purposes, let us suppose that our ideal observer is in the actual world, fully informed, and ideally rational.

Here, the socio-normative foundations of our responsibility practices can do some work for us, for we can specify the observer's interest by an appeal to it. What justifies the responsibility practices is the effects of those practices in sustaining and enhancing moral considerations-sensitive agency. Thus, our ideal observer's interest is in selecting a notion of capacity that would be at least co-optimal for (1) ensuring that agents in the actual world recognize and suitably govern themselves in light of moral considerations, and (2) ensuring agents have wider rather than narrower ranges of context of action and deliberation in which agents so deliberate and act, so long as it does not conflict with (1).

In selecting the relevant Sidgwickian capacity, our observer is concerned to respect features of our current psychological dispositions, the cultural and social circumstances of our agency, our interest in resisting counterfactuals we deem deliberatively irrelevant in the actual world (think: finks[15], and the need for agents to internalize norms of action and deliberation concerning moral considerations at a level of granularity that is useful in ordinary deliberative

[15] And Frankfurt-style cases. I take it that in the actual world, Frankfurt cases are infrequent and not a possibility with deliberative significance in the ordinary course of things. In a world in which we had reason to think Frankfurt cases were common, rather than exceptional, it is conceivable that such cases could have a different significance for the relevant construal of capacities.

and practical circumstances.[16] On this account, an agent has the responsibility-relevant capacity to recognize and respond to moral considerations if he or she recognizes and appropriately responds to the relevant moral considerations, or, if in a suitable proportion of relevantly similar worlds, the agent recognizes and responds to moral considerations, by the standard specified by the observer.

This picture gives us a systematic account for thinking about the capacities that matter for responsibility. First, capacities are indexed to circumstances. That is, we relinquish appeal to global, cross-situationally stable agential capacities to recognize and respond to considerations. Second, in a circumstance, the capacity required for blameworthiness is that capacity that, in the actual world, supports and extends our ability to recognize and respond to moral considerations. These capacity facts are in some sense "higher-order" or constructed facts, rather than facts about interest-independent metaphysical features of agents. This is desirable. As I noted at the outset, the point is to show that our interest in responsibility and the metaphysics of agency are tied to something of value. That is precisely what the Sidgwickian capacity specifies.

An important consequence of this approach is that the relevant Sidgwickian capacity makes sense of the idea of unexercised capacities. That is, there is ample room for the truth of judgments that someone could have done differently, but did not. Indeed, one way our capacities to recognize and respond to moral considerations can be extended to new concerns and contexts is for us to be vulnerable to blame because we had an unexercised capacity.[17]

[16] As one might suspect, this picture of capacities admits of a much more detailed characterization, involving possible worlds and context individuation. I pursue those details elsewhere (Vargas 2013a: 213-28).

[17] Notice that this account provides an explanation for nonvoluntary culpability of the sort that has motivated "attributionists": nonvolitional cases are cases of having the responsibility-relevant capacity to recognize and suitably respond to some set of moral considerations centered on due concern for others, but where the agent has failed to do so.

Sidgwickian capacities are unlikely to characterize our powers exactly as would scientific or metaphysical inquiries detached from the present normative concerns. For example, the notion of ability here will not map on to a notion of ability concerned with accessible worlds, as allowed by the actual past and the laws of nature. Whatever the intuitive appeal of a restricted notion of ability, it is not required to support the social and normative concerns that structure the proposed account of responsibility. Moreover, the capacities significant for responsibility are plausibly more coarsely grained than that. The reason is simple: the failure to possess such capacities is normally of tremendous significance to individual agents, and to their participation in shared cooperative life.[18] Thus, the relevant capacities will be structured by the fragility of our actual psychological possibilities, but those capacities will be neither so finely grained as the capacities we get from scientific inquiries into humans nor so coarsely grained as we get from philosophical theories that presume cross-situational stable rational capacities in agents.

[18] Thanks to Steven Wall for calling my attention to this point. This dovetailing of personal interest and collective interests is, I maintain, a core aspect of moral responsibility. What about first-personal concerns about implicit bias, of the Glasgowian sort? There is usually a prima facie first personal concern about failures of self-governance. Part of membership in the moral community is the presumption that one can appropriately self-govern in light of moral considerations, and discovery of one's implicit biases casts doubt on that. The first personal case works somewhat differently than other cases, because the relationship to the fault is more personal: the defect is ours by our own lights, and our agency is revealed to us as less excellent than we thought it to be. Even when others don't hold that such defects are culpable, we can still think poorly of ourselves for not meeting our own standards. When an agent feels guilty about implicit bias, it is not insensible that such an agent feels guilty about it. It is, after all, a breakdown of our self-governance in light of moral considerations, and that can matter first-personally, even if the community at large doesn't find that breakdown telling with respect to the moral demands we collectively recognize. In such a case, guilt works in the customary way: it increases the agent's disposition to moral repair and stokes efforts to improve one's own moral agency. In light of this, it seems to me reasonable to hold that it can make good sense for an individual find fault with him or herself for implicit bias, and to welcome condemnation for it, without it following that such blame is more generally licensed.

4. Responsibility and implicit bias

On the present account, individual moral responsibility is less a question of metaphysics than social and normative functioning. Here, I turn to what implications this account has for responsibility for implicit bias. I argue that (1) there is no uniform answer here, across all contexts, but that (2) most people are not currently responsible for action caused by implicit bias, but that they will (perhaps soon) be responsible for implicit bias-caused action.[19]

I will begin by canvassing some of the more salient considerations pro and con for holding agents responsible for action derived from implicit bias.

There are at least three reasons to insist on holding agents responsible for implicit bias-derived action. First, there is the matter of indirect responsibility. That is, even if agents do not have direct control over their biases, they may have sufficient indirect control over them to ground responsibility attributions. Second, there is the characteristic benefit of blame that would follow in the wake of holding responsible. That is, blaming would bring into play the moral feedback upon which our socially self-governed agency normally depends. Third, there is also the possibility that narratives about responsibility and self-control can support the acquisition of and sustain the possession of the requisite capacities. Scholars in a number of fields have argued that narratives about capacity can work to enable or disable those capacities.

For example, if we tell young girls that they are not good at math and science, we stand a much better chance of making a generation of women who are in fact less capable at math and science. Similarly, if we tell children that their socially significant qualities are not fixed, but

[19] I say "most" because there may well be pockets of the social world in which people are responsible, even now. To anticipate: if there are social contexts in which virtually everyone is aware of the existence of implicit bias, it is widely regarded as morally undesirable, and there are widely recognized norms prohibiting implicitly biased behavior, then for those agents, if they regularly operate in those contexts, then they may count as morally responsible for bias-derived action in those and nearby contexts.

instead subject to improvement through effort and practice, such qualities are likely to be subject to more self-development (Dweck and Molden 2008). So, perhaps responsibility for implicit bias-based action will work in a similar fashion: by promulgating narratives of control and responsibility for bias, we might make it the case that agents come to have more control and responsibility for their biases.

This last possibility is worth more consideration than I can give it here. I will say this, however: although the promulgation of the control idea is powerfully appealing, it raises special difficulties. Norms do not operate in a vacuum. For norms to work, they require a kind of buy-in on the part of agents, a willingness to internalize and enforce the norm on the part of both the blamed and the blamers. One cost to maintaining that people are responsible for their implicit biases is that it seems plausible that nearly everyone is subject to them. If we begin insisting on universal culpability for implicit biases, the risk is that we provoke widespread defensiveness and hostility (Cf. Saul 2013a: 55). Defensiveness and hostility might slow the successful internalization of the relevant norms, and might even undercut the moral force of concern for bias, its effects, or even practices of moral blame.

The difficulties here are not just on acceptance of the norms by those being blamed. To see why, reflect on what we might think of the "gossipy" dimensions of blaming (Cf. Vargas 2016). That is, blaming invites others to express commitment to shared norms, to express solidarity with those affected by some offense, and to identify offenders as violating our shared norms. If that is right, the social and communicative dimensions of blame, including its protestive goods (McKenna 2013, Cf.; Smith 2013), are less likely to play their customary roles when the blamer is viewed as expressing norms that are not already shared. When blamers have a commitment to a standard of conduct without widespread currency in the blamer's society, it is plausibly costlier to express and enforce the norm. In contrast, the cost of blaming plausibly goes down when one is perceived to be enforcing a norm that others accept. So, there are special challenges for pursuing blame in an environment where the norm is not already in play. To be sure, matters here are complicated.[20] However, U.S. experiences of Prohibition suggest one reason for caution: norm advocacy that gets too far ahead of internalized practices can be extraordinarily costly. Under Prohibition, the costs of the new legal norms were plausibly not limited to unwillingness to give up drinking. Instead, the effects of resistance to the law arguably spread to a number of affiliated domains, undermining the authority of the law in a broader way (Bilz and Nadler 2009).

There is a distinct but related worry here, concerning indifference. One response to the discovery of a widespread shortcoming is to shrug one's shoulders. Sometimes, a problem for all is regarded by most as a problem for none. Rather than something in need of effortful remedy, the absence of clear narratives about the effects of bias in action may cause many people to regard bias as something to which we should just be resigned. This possibility highlights the need for social, material, and normatively-structured circumstances that reinforce and support the wrongfulness of bias-produced actions. I'll return to this idea—what I've elsewhere called the idea of "moral ecology"—below.

Appealing to the advantages of extending blaming practices into the domain of assessments of implicit bias is problematic as well. In the first place, there is the quasi-empirical

[20] For example, there is some evidence that suggests that confronting people about their bias decreases biased attitudes Czopp et al. (2006). Moreover, it is plausible that here there are familiar norms of reproach about certain forms of bias, and there is notable risk of attention, knowledge of anti-bias norms plausibly structure behavior. For a fascinating example of this in a game show context see Levitt and Dubner (2009, 75-81). There are also really interesting philosophical issues concerning the appropriateness of blame and reproach under conditions that systematically disable the culpability of wrongdoers. For a rich and fascinating discussion, see Calhoun (1989). I'm less sanguine than Calhoun about disconnecting the appropriateness of blame from blameworthiness, but her account merits more attention than I can give it here.

question of whether agents really are sensitive to moral considerations in the operations of implicit bias.[21] Second, even if they are suitably sensitive to agent-level control, there is still the question about whether holding one another responsible for such exercises of agency would, given the ubiquity of implicit bias, cut against the buy-in or efficacy of the practice.

What about the possibility that we have sufficient indirect control over our biases to support blameworthiness? It depends on the details. Of course, if implicit bias turns out to be a plural phenomenon—and I noted at the outset that it likely is—then how indirect control operates may itself be sufficiently diverse that we cannot give easy, sweeping answers about it. I remain skeptical, because (as I will argue below) some of the social scaffolding of our responsibility practices are not in place in the case of implicit bias.

Suppose that agents have some awareness of their own biases, and that there is some sense in which it is not impossible for agents to guide their behavior in light of this fact.[22] Even so, agents are often not blameworthy for implicit bias-caused action, and aspects of the agency cultivation model help bring out why. First, because of their social nature, responsibility norms are dynamic, not static. As social needs change, as our cultural scripts or narratives about agents' powers shift, the responsibility-relevant capacities we have will shift, as will the content of our justified responsibility norms. Second, consider the thought that, on the agency cultivation model, the circumstances of responsibility matter for whether agents are responsible. What this account shows us is that if we are interested in moral responsibility, we also need to be interested in the circumstances of moral responsibility, or the moral ecology of responsibility.[23]

[21] Levy (2015) advances reasons to doubt that implicit biased-caused action is typically sensitive to reasons in this way.

[22] Recall that this matter is contested. See, for example, Saul (2013a), Holroyd (2015), and Levy (2014; Levy 2015).

[23] I confess that I am sometimes tempted by the thought that it is simply an indeterminate matter whether people are responsible for implicit bias-caused actions. There may be no best set of

How the ecology of responsibility matters becomes apparent if we contrast it with how we might think about implicit bias in conventional atomistic terms. Suppose, for example, that we wished to show that implicit bias satisfies some picture of non-Sidgwickian capacities to recognize reasons and respond to them. On this approach, one could argue that implicit bias satisfies such conditions if, for example, agents could be shown to be aware of such biases and had some control, even if only indirect, over them.

For example, some have argued that non-specialists, or the folk, have been aware of the phenomenon of implicit bias for some time, even if not under the specialist label (Madva in preparation). That is an important insight. But recognizing a phenomenon is not the same thing as having widespread social expectations to monitor and control behavior arising from that phenomenon. Without the latter—norms of monitoring and controlling behavior rooted in implicit bias—it does not seem that the social components required for blameworthiness are in place in the case of implicit bias.

Many of our social contexts do *not* support robust moral blame for implicit bias in the customary ways. To see why, consider that virtually all the salient facts about implicit bias are invisible outside of specialized literature in the academy, human resources departments, and legal counsel concerned with employment issues. In the broader public, there is little recognition that:

i. everyone has implicit biases

ii. implicit biases affect behavior

iii. we have reason to monitor our implicit attitudes, or at least implicit-attitude behavior, when those implicit attitudes are such that expressions of the attitude

norms concerning responsibility for implicit bias, because both the considerations for and against holding agents responsible in our current circumstances have considerable independent force. If so, this is a domain in which the facts of responsibility, such as they are, remain to be constructed by us. Today, however, I am inclined to resist the temptation offered by that thought.

disproportionately impose costs on others.

iv. self-monitoring helps us control and mitigate behavioral effects of implicit bias

These information deficits make a difference in the ecology of responsibility. Responsibility norms are partly a matter of our being awareness that others expect certain things of us. If we are not generally cognizant of the existence of implicit bias, and that these biases really do affect behaviors, it is difficult for such things to give rise to reasons to monitor those attitudes and to take ownership of them.

These considerations do not preclude the possibility that there are social contexts in which the requisite social or institutional scaffolding is in place to underpin culpability for particular forms of bias-produced action. For example, in many larger companies, human resources departments have, or should have, sufficient awareness of the relevant facts to require hiring practices that are intended to counteract or preclude gender and racial bias in hiring. Smaller groups—say, a small firm of employment discrimination attorneys, or private practice of a group of clinical psychologists—may be suitably situated to take up and implement such practices as well. However, the requisite knowledge and norms are less plausibly in place in, for example, the foul calls of a weekend soccer game referee, the hiring of a bartender at the corner bar, or even a university student's evaluation of the teaching qualities of their instructors.

If we have learned anything from 20th century psychology, it is that the picture of selves as possessing an impregnable inner citadel, immune to social influence, is largely a myth. Once we give up an atomistic picture of responsibility, we have to look to whether social contexts support the exercise of rational agency. With respect to eradicating implicit bias, our current contexts oftentimes fail to provide the relevant support (Huebner 2016). Our control over ourselves depends in part on reliable feedback from others and our regular exposure to risk of others' disapprobation. Where such things are absent, I am less reliably able to ensure that I act on moral principles that we would all accept (or, at least not reject). So, even if we have a kind of indirect control over implicit bias—indeed, especially if we have only indirect control over our biases—without widespread attendant norms concerning implicit bias, there are limited grounds for insisting on responsibility.

Thus, from the standpoint of a concern for moral considerations-sensitive agency, we do better to think of agents in many circumstances as not typically morally responsible for the biased elements of their bias-caused actions.

In rejecting responsibility for (the biased elements in or flowing from) bias-affected actions, we do not thereby lose all resources for making moral evaluations of agents and bias-tainted acts. Aretaic and axiological evaluations remain largely intact, even when there is no responsibility. From the standpoint of an ideally virtuous person, or from the standpoint of whether we satisfy values that we aspire to, we can still critique biases and biased agents. Moreover, the absence of responsibility would not absolve us of the burden to ameliorate its effects, and to undertake transformation of the social features that give rise to it. All non-responsibility means is that we cannot readily blame most people for the shaving of these attitudes.

What we can do, though, is to blame ourselves and others for failing to take steps to begin changing the institutional and social circumstances that give rise to these biases. Importantly, implicit biases do not operate in vacuum. To a significant extent, they seem to form in response to statistical patterns in one's environment (Gendler 2011). For example, if one only encounters members of a minority group in, for example, service industry professions, one is less likely to regard individual members of that group as well-suited for cognitive labor. So, the challenge is to change the circumstances that foster such attitudes.

These thoughts—the absence of suitable social scaffolding to support the appropriateness of blaming, the thought that agents are oftentimes not responsible for the acquisition of implicit biases, and pressure to change the circumstances that give rise to implicit bias–motivate the following view. First, there is a presumption that, absent particular institutional contexts, most agents are not morally responsible for bias-caused action in our current circumstances. Second, this presumption will be overturned in the not-too-distant future. The motivating idea is this: given that we are currently under some obligation to fight implicit bias and its effects—and assuming that we do so by highlighting its existence, its operations, and its epistemic and moral costs—then the social context of future bias will be different. We can hope that in some not-entirely distant future, agents in those circumstances will have the advantages of structural and institutional transformations that reduce the risk of implicit bias and that correspondingly improve our capacity for social self-governance.

Most of the time, now, we should avoid blame for implicit bias-caused action. We should also make our circumstances such that full-throated blame for implicit bias becomes a live option, and so that the benefits of social self-governance ensue. After remarking positively about the potential for effective interventions on individuals' biases, Huebner (2016), concludes that "we will be unable to moderate or suppress all of our problematic biases until we eliminate the conditions under which they arise" (71). This seems exactly right. However, the situation may be even more pressing than Huebner was supposing: recent data suggests that the standardly touted interventions to reduce implicit bias have no long term success (Lai et al. 2016). If implicit bias is a product of a biased social world, and it is incorrigible as the Lai et al. data suggests, large-scale remapping of the bias-reflecting and bias-producing norms, expectations, and social patterns becomes even more pressing. Plausibly, this puts special—although not exclusive—pressure on those who have outsize effects on our social context. That is, there is a normative pressure, perhaps even an obligation, to foster circumstances of moral responsibility. So, if one can shape the statistical patterns that create the basis for bias (as, for example, education systems and media plausibly do), there is special reason to do so. As always, though, there are many morally fraught paths lurking here, and we should not assume that the way forward is obvious, and nor should we assume it is without moral and material cost.

5. Conclusion

I have argued for several claims. First, traditional approaches to the question of responsibility for implicit bias-caused action face methodological worries. Second, we can make use of an approach to responsibility—the agency cultivation model—that avoids atomistic presumptions and provides a principled way of capturing some interest-relative features of talk about agent-level capacities. In doing so, we can cast some light on the matter of responsibility for implicit bias-derived action. Finally, on this account, there is some reason to hold that, frequently, people are not responsible for implicit bias in a range of ordinary social contexts, but that we should nevertheless try to change the individual sources and social milieu in which implicit bias arises. If we do, we can make it the case that most agents are blameworthy for a wide range of implicit bias-caused action.

Two further upshots are worth highlighting. First, it is evident that we have been asking the wrong questions about moral responsibility and implicit bias. Rather than focusing on whether agents are morally responsible for bias-caused action, we should be instead focused on the question of how our norms ought to be, or the form we wish our moral ecology to have. If we can reshape our contexts and social practices so that they do not dispose us to bias, that will be no small victory. Indeed, it might make the question of the blameworthiness of implicit bias-caused action less central, given the foundational matter of establishing a suitable moral ecology.

The second upshot is this: there is a middle path between the ordinary image of our agency and the more revolutionary and skeptical pictures of agency and responsibility that have emerged from the scientific literature. This middle path has distinctive challenges, too, but it also offers a template for reconciling the manifest and scientific images of our agency.[24]

[24] My thanks to participants and audience members at a variety of places where I floated these ideas. They include the 2013 Implicit Bias, Philosophy, and Psychology conference at the University of Sheffield, the philosophy department of Lewis and Clark College, the School of Law at the University of San Francisco, the 2015 Austin J. Fagothey, S.J. Philosophy Conference at Santa Clara University organized by Erick Ramirez; and the 2015 New Orleans Workshop on Agency and Responsibility organized by David Shoemaker. Special thanks to John Doris, Jules Holroyd, Bryce Huebner, Dan Kelly, Alex Madva, Elinor Mason, Jay Odenbaugh, Adina Roskies, Grant Rozeboom, Yonatan Shemmer, Nicholas Smith, and Natalia Washington for particularly helpful feedback on aspects of the paper, as well as two anonymous referees for this volume. Thanks also to The Leverhulme Trust for its support of the Implicit Bias and Philosophy International Research Project (grant IN-058), and especially the organizer, Jennifer Saul, for the forum that gave rise to many of these thoughts.

Works Cited

Arpaly, Nomy. (2003). *Unprincipled Virtue*. (New York: Oxford.)

Bergh, John A. (2008). "Free Will is Un-Natural." In *Are We Free? Psychology and Free Will*, edited by John Baer, James C. Kaufman, and Roy F. Baumeister, 128–54. (New York: Oxford University Press.)

Bennett, Christopher. (2002). "The Varieties of Retributive Experience." *The Philosophical Quarterly* 52 (207): 145–63.

Bilz, Kenworthey, and Janice Nadler. (2009). "Law, Psychology, and Morality." In *Moral Cognition and Decision Making: The Psychology of Learning and Motivation*, edited by D. Medin, L. Dkitka, C.W. Bauman, and D. Bartels, 101–31. (San Diego, CA: Academic Press.)

Blanton, Hart, James Jaccard, Jonathan Klick, Barbara Mellers, Gregory Mitchell, and Philip Tetlock. (2009). "Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT." *Journal of Applied Psychology* 94 (3): 567–82.

Brownstein, Michael. (2016). "Implicit Bias." *The Stanford Encyclopedia of Philosophy* Spring 2016 Edition <http://plato.stanford.edu/archives/spr2016/entries/implicit–bias/>.

Brownstein, Michael. (online 2015). "Attributionism and Moral Responsibility for Implicit Bias." *Review of Philosophy and Psychology* 1–22.

Calhoun, Cheshire. (1989). "Responsibility and Reproach." *Ethics* 99 389–406.

Cashmore, Anthony R. (2010). "The Lucretian Swerve: The Biological Basis of Human Behavior and the Criminal Justice System." *Proceedings of the National Academies of Sciences* 107 (10): 4459–504.

Czopp, Alexander M., Margo J. Monteith, and Aimee Y. Mark. (2006). "Standing Up for a Change: Reducing Bias Through Interpersonal Confrontation." *Journal of Personality and Social Psychology* 90 (5): 784–803.

Doris, John. (2002). *Lack of Character*. (New York: Cambridge University Press).

Doris, John. (2015a). *Talking to Ourselves: Reflection, Skepticism, and Agency*. (New York: Oxford University Press).

Doris, John. (2015b). "Doing Without (Arguing About) Desert." *Philosophical Studies* 172 (10): 2625–34.

Doris, John, and Dominic Murphy. (2007). "From My Lai to Abu Ghraib: The Moral Psychology of Atrocity." *Midwest Studies in Philosophy* 31 25–55.

Dweck, Carol S., and Daniel C. Molden. (2008). "Self-Theories: The Construction of Free Will." In *Are We Free? Psychology and Free Will,* edited by John Baer, James C. Kaufman, and Roy F. Baumeister, 44–64. (New York: Oxford University Press).

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. (New York: Cambridge University Press).

Fischer, John Martin, and Neal Tognazinni. (2011). "The Physiognomy of Responsibility." *Philosophy and Phenomenological Research* 82 (2): 381–417.

Gendler, Tamar Szabó. (2011). "On the Epistemic Costs of Implicit Bias." *Philosophical Studies* 156 (1): 33–63.

Glasgow, Joshua. (2016). "Alienation and Responsibility." In *Implicit Bias and Philosophy: Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, edited by Michael Brownstein, and Jennifer Saul. (New York: Oxford University Press).

Greene, Joshua, and Jonathan Cohen. (2004). "For the Law, Neuroscience Changes Everything and Nothing." *Philosophical Transactions of the Royal Society of London B* 359 1775–85.

Greenwald, Anthony, G., Mahzarin R. Banaji, and Brian A. Nosek. (2015). "Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects." *Journal of Personality and Social Psychology* 108 (4): 553–61.

Greenwald, Anthony, G., T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji. (2009). "Understanding and Using the Implicit Association Test: Iii. Meta-Analysis of Predictive Validity." *Journal of Personality and Social Psychology* 97 (10): 17–41.

Hart, H. L. A. (1959). "Prolegomena to the Principles of Punishment." *Proceedings of the Aristotelian Society New Series* 60 1–26.

Holroyd, Jules. (2012). "Responsibility for Implicit Bias." *Journal of Social Philosophy* 43 (3): 274–306.

Holroyd, Jules. (2015). "Implicit Bias, Awareness, and Imperfect Cognitions." *Consciousness and Cognition* 33 511–23.

Holroyd, Jules, and Joseph Sweetman. (2016). "The Heterogeneity of Implicit Bias." In *Implicit Bias and Philosophy: Volume 1: Metaphysics and Epistemology*, edited by Michael Brownstein, and Jennifer Saul, 80-103. (New York: Oxford University Press).

Hooker, Brad. (2016). 'Wrongness, Evolutionary Debunking, Public Rules', *Ethics and Politics*, XVIII (1), 135-49.

Huebner, Bryce. (2016). "Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition." In *Implicit Bias and Philosophy: Volume 1: Metaphysics and Epistemology*, edited by Michael Brownstein, and Jennifer Saul, 47-79. (New York: Oxford University Press).

Jost, John T., Laurie A Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, and Curtis D. Hardin. (2009). "The Existence of Implicit Bias is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies That No Manager Should Ignore." *Research in Organizational Behavior* 29 39–69.

Kelly, Daniel, and Erica Roedder. (2008). "Racial Cognition and the Ethics of Implicit Bias." *Philosophy Compass* 3 (3): 522–40.

Kelly, Erin I. (2009). "Criminal Justice Without Retribution." *The Journal of Philosophy* 106 (8): 440–62.

Knobe, Joshua, and Erica Roedder. (2009). "The Ordinary Concept of Valuing." *Philosophical Issues* 19 131–47.

Krieglmeyer, R, and J.W. Sherman. (2012). "Disentangling Stereotype Activation and Stereotype Application in Stereotype Misperception Task." *Journal of Personality and Social Psychology* 103 (2): 205–24.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., Hu, X., McLean, M. C., Axt, J. R., Asgari, S., Schmidt, K., Rubinstein, R, Marini, M., Rubichi, S., Shin, J. L., & Nosek, and B. A. (2016). "Reducing Implicit Racial Preferences: II. Intervention Across Time." *Open Science Framework* <https://osf.io/v36wf/> Draft.

Levitin, Daniel. (2013). What You Might Be Missing. *The Wall Street Journal*, C5.

Levitt, Steven D., and Stephen Dubner. (2009). *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. (New York: William Morrow).

Levy, Neil. (2015). "Neither Fish Nor Fowl: Implicit Attitudes as Patchy Endorsement." *Noûs* 49 (4): 800–23.

Levy, Neil. (2014). "Consciousness, Implicit Attitudes, and Moral Responsibility." *Noûs* 48 (1): 21–40.

Madva, Alex. (In preparation). "Implicit Bias, Moods, and Moral Responsibility." Unpublished Manuscript.

McKenna, Michael. (2012). *Conversation and Responsibility*. (New York: Oxford University Press).

McKenna, Michael. (2013). "Directed Blame and Conversation." In *Blame: Its Nature and Norms*, edited by Justin D. Coates, and Neal A. Tognazzini, 119–40. (Oxford: Oxford University Press).

Montague, P. Read. (2008). "Free Will." *Current Biology* 18 (14): R584–85.

Nahmias, Eddy. (2007). "Autonomous Agency and Social Psychology." In *Cartographies of the Mind: Philosophy and Psychology in Intersection*, edited by Massimo Marraffa, Mario De Caro, and Francesco Ferretti, 169–85. (Berlin: Springer).

Nelkin, Dana. (2005). "Freedom, Responsibility, and the Challenge of Situationism." *Midwest Studies in Philosophy* 29 (1): 181–206.

Nelkin, Dana. (2016). "Difficulty and Degrees of Praiseworthiness and Blameworthiness." *Noûs* 50(2) 356-378.

Nichols, Shaun. (2015). *Bound: Essays on Free Will and Responsibility*. (New York: Oxford University Press).

Oswald, Frederick, Derek Mitchell, Hart Blanton, James Jaccard, and Philip Tetlock. (2013). "Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies." *Journal of Personality and Social Psychology* 105 (2): 171–92.

Oswald, Frederick, Gregory Mitchell, Hart Blanton, and James Jaccard. (2015). "Using the IAT to Predict Ethnic and Racial Discrimination: Small Effect Sizes of Unknown Societal Significance." *Journal of Personality and Social Psychology* 108 (4): 562–71.

Parfit, Derek. (2011). *On What Matters*. Vol. 1 (New York: Oxford University Press).

Pereboom, Derk. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

Pockett, Susan. (2013). "If Free Will Did Not Exist, it Would Be Necessary to Invent it." In *Exploring the Illusion of Free Will and Moral Responsibility*, edited by Gregg Caruso, 265–72. (Lanham, Maryland: Lexington Books).

Rawls, John. (1955). "Two Concepts of Rules." *Philosophical Review* 64 3–32.

Raz, Joseph. (2011). *From Normativity to Responsibility*. (Oxford: Oxford University Press).

Saul, Jennifer. (2013a). "Unconscious Influences and Women in Philosophy." In *Women in Philosophy: What Needs to Change?*, edited by Fiona Jenkins, and Katrina Hutchison, 39–60 (New York: Oxford).

Saul, Jennifer. (2013b). "Scepticism and Implicit Bias." *Disputatio* 5 (37): 243–63.

Shoemaker, David W. (2011). "Attributability, Answerability, and Accountability: Towards a Wider Theory of Moral Responsibility." *Ethics* 121 602–32.

Smith, Angela. (2013). "Moral Blame and Moral Protest." In *Blame: Its Nature and Norms*, edited by D. Justin Coates, and Neal A. Tognazzini, 27–48. (New York: Oxford University Press).

Strawson, Galen. (1994). "The Impossibility of Moral Responsibility." *Philosophical Studies* 75 5–24.

Vargas, Manuel. (2013a). *Building Better Beings: A Theory of Moral Responsibility*. (Oxford, U.K.: Oxford University Press).

Vargas, Manuel. (2013b). "If Free Will Does Not Exist, Then Neither Does Water." In *Exploring the Illusion of Free Will and Moral Responsibility*, edited by Gregg Caruso, 177–202. (Lanham, Maryland: Lexington Books).

Vargas, Manuel. (2013c). "Situationism and Moral Responsibility: Free Will in Fragments." In

Decomposing the Will, edited by Till Vierkant, Julian Kiverstein, and Andy Clark, 325–49. (New York: Oxford University Press).

Vargas, Manuel. (2014). "Razian Responsibility." *Jurisprudence* 5 (1): 161–72.

Vargas, Manuel R. (2015). "Desert, Responsibility, and Justification: Reply to Doris, McGeer, and Robinson." *Philosophical Studies* 172 (10): 2659–78.

Vargas, Manuel. (2016). "Responsibility and the Limits of Conversation." *Criminal Law and Philosophy* 10 (2): 221–40.

Vargas, Manuel. (Forthcoming). The Social Constitution of Responsible Agency: Oppression, Politics, and Moral Ecology. In *The Social Dimensions of Responsibility*, edited by Marina Oshana, Katrina Hutchinson, and Catriona Mackenzie.

Vincent, Nicole. (2011). "A Structured Taxonomy of Responsibility Concepts." In *Moral Responsibility: Beyond Free Will and Determinism*, edited by Nicole Vincent, Ibo van de Poel, and Jeroen van den Hoven, 15–35. (Dordrecht, The Netherlands: Springer).

Washington, Natalia, and Daniel Kelly. (2016). "Who's Responsible for This? Implicit Bias and the Knowledge Condition." In *Implicit Bias and Philosophy: Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, edited by Michael Brownstein, and Jennifer Saul. (New York: Oxford University Press).

Watson, Gary. (1996). "Two Faces of Responsibility." *Philosophical Topics* 24 227–48.

Wegner, Daniel M. (2002). *The Illusion of Conscious Will.* (Cambridge, MA: MIT Press.)