Chapter Six

Social Psychology, Phenomenology, and the Indeterminate Content of Unreflective Racial Bias

Alex Madva

This chapter is about implicit bias. I began writing this, however, in the United States of 2017, during a period of burgeoning explicit bigotry and intergroup hostility. The dominant narrative surrounding implicit bias has been *aversive racism*,[1] according to which most Americans are sincerely opposed to racial discrimination at the explicit level but biased at the implicit level. What, then, to make of the resurgence of full-throated self-ascriptions of white supremacy? Should we say, with apologies to Virginia Woolf, that on or about November 2016, human nature changed? Or that so many of us just got human nature completely wrong?

This chapter argues that implicit bias actually helps to explain this resurgence of bigotry. But properly appreciating implicit bias's explanatory power requires rewriting the dominant narratives about its content, conscious accessibility, and context sensitivity. One source of confusion is that "implicit" racial bias—the construct measured with tools like the Implicit Association Test (IAT)[2]—is often described as entirely unconscious. Perhaps describing implicit bias as completely unconscious helps people to acknowledge that they are "part of the problem" without becoming defensive. Yet the evidence consistently suggests that individuals are aware of their implicit biases, albeit in partial, inarticulate, or even distorted ways. These biases form part of the "background" of social experience, exerting a pervasive influence on attention, judgment, and action, even though they are often felt without being noticed, or noticed without being understood.[3] To help make sense of these findings, this chapter develops a thought suggested by Linda Martín Alcoff and Gail Weiss, that implicit bias paradigmatically operates at the *intermedi-*

ate level of awareness, between total nonconsciousness and articulated self-knowledge, that has long been a central concern of the phenomenological tradition. Implicit bias dwells in the "unthematized, taken-for-granted . . . pre-reflective habits"[4] that "structure affect, perception, and interpretation."[5]

Another source of confusion is that implicit biases are often glossed as *mere associations* between groups and traits, which lack intentional content. Take the race–weapon implicit bias. Most Americans (including many African Americans) more easily and quickly identify images of dangerous weapons when they are paired (e.g., share the same button on the keyboard) with black faces than with white faces. This tendency to "associate" blacks and weapons correlates with a bias toward "shooting" unarmed black men in a video game[6] and with regional U.S. trends involving disproportionate police shootings of blacks.[7] That is, in regions where participants (most of whom are neither police officers nor victims of brutality) display stronger black–weapon associations, unarmed black people are also more likely to be shot by the police.[8] Regional IAT data predicted these shooting disparities more than any other tested variable, including self-reported racial attitudes and regional levels of residential segregation, violent crime, and unemployment. Now what does it mean to interpret this race–weapon implicit bias as a mere contentless association? It is, for one thing, to deny that individuals who demonstrate this bias tend to *believe*, either consciously or unconsciously, that blacks are more likely to carry weapons. The idea instead is that these individuals simply associate "black" and "weapon" in much the same way that they associate "salt" and "pepper" or "doctor" and "nurse": thoughts of one activate thoughts of the other.

This chapter argues that implicit biases are neither mere associations nor fully articulated, propositionally structured beliefs or emotions.[9] Implicit biases are contentful—they take the world to be a certain way—but, in paradigmatic cases, their content is *indeterminate*. I defend content indeterminism about implicit bias in the metaphysical (rather than epistemic; see below) sense. For example, when a white person experiences a "gut feeling" of discomfort during an interaction with a black person, there is a question about the meaning or nature of that discomfort. Is it a fear of black people? Is it mere anxiety about appearing racist? There is, I'll argue, no general, determinate answer. The contents of our unreflective racial attitudes are fundamentally vague and open-ended, although they take on particular shapes and implications—that is, they *become* determinate—depending on contextual features including character traits, background assumptions, and structural power relations. In other words, they are indeterminate when interpreted individualistically, in isolation from context, but determinate (or at least less indeterminate) when understood holistically and relationally, as part of broader cognitive-bodily-social-environmental systems.[10]

In addition to the specific theses defended, I hope to offer a case study in bridging diverse approaches to race and racism. First, we must merge individualistic and structural perspectives.[11] Understanding individuals' racial attitudes requires situating those individuals in broader social contexts; conversely, understanding racially oppressive structures—and envisioning emancipatory alternatives—requires populating those structures with embodied, biased minds. Second, following the twentieth-century phenomenologists who integrated ongoing social-scientific developments with the philosophy of lived experience, I hope to spur greater cross talk between psychologists and phenomenologists studying race. My sense is that some phenomenologists have been unduly dismissive of implicit bias research, raising criticisms more aptly directed at popularized depictions of the research than at in-the-weeds empirical developments.[12] Conversely, many psychologists (and the philosophers under their influence) are too quick to infer, from various *specific* findings that upend *specific* "commonsense" assumptions about experience, that phenomenological investigation is altogether wrongheaded. To the contrary, phenomenologists of race may have much to offer social psychologists, both in assisting theoretical interpretations and identifying underexplored directions for research.

The roadmap is as follows. In the first section, I say a bit about the phenomenology of indeterminacy. In section two, I make the case that implicit bias has indeterminate content. In section three, I draw out further implications of my argument, rejecting more radically constructionist and existentialist approaches and defending an enriched understanding of person–situation relations. In the final section, I consider two alternative views.

PRIMER ON PHENOMENOLOGICAL INDETERMINACY

Maurice Merleau-Ponty argues that all experience is shot through with indeterminacy. Indeterminacy is, for him, not an epistemic flaw or practical limitation in our relation to the world, but a fundamental condition underlying our basic abilities to know and navigate physical and social environments. "We must recognize the indeterminate as a positive phenomenon," he writes; we must identify and understand "the presence in the perceived of a positive indeterminacy."[13]

What kind of indeterminacy is Merleau-Ponty interested in? What makes it positive? He is focused on a range of fleeting experiences that we encounter in everyday contexts, such as difficult-to-make-out street signs or scribblings on a blackboard. When a sighted person, in her ordinary comings and goings, comes across such indeterminate-looking percepts, her experience often has an *affective* and *action-oriented* character: their blurriness presents itself as a *problem* for her to solve. These "vague something-or-others . . .

invite further exploration,"[14] perhaps compelling her to lean forward or approach the something-or-others to see them more clearly, or at least to squint or tilt her head to ease the felt sense of tension induced by the indeterminacy. Some of the most easily replicable and shareable experiences of indeterminacy may be auditory, such as the shared difficulty we have in hearing song lyrics. There is, for example, an infamous line in Aretha Franklin's rendering of "Respect" that is persistently difficult to make out.

What makes this indeterminacy "positive"? First, these experiences of indeterminacy are functional and norm sensitive: they induce a sense of unease that motivates certain behavioral responses, as when we tilt our ears to the speaker, reach to turn up the volume, or briefly stop singing along to listen more attentively. It is *because* the lyric is perceived indeterminately that we feel compelled to discover its determinate properties. In short, the experience of indeterminacy motivates us to make matters more determinate, to get a better epistemic and practical grasp on our environment. Merleau-Ponty thus writes that "a sensible that is about to be sensed poses to my body a sort of confused problem. I must find the attitude that *will* provide it with the means to become determinate. . . . I must find the response to a properly formulated question."[15]

Moreover, in paradigm cases, we must bring to bear a range of perceptual, bodily, and cognitive *skills* in order to resolve the indeterminacy. The transition from vague something-or-other to determinate perception is not a passive process that just happens to us but a cognitive-affective-behavioral accomplishment. It is also typically a social accomplishment. Many indeterminacy-resolving skills are learned via interaction with others, and many indeterminacies are resolved collaboratively, as when we pause the conversation during a song's chorus in order to collectively discern the garbled lyrics. Consider also how drivers are more likely to get in accidents when they talk on the phone (even if they are using hands-free devices to communicate) than when they talk to a passenger, because passengers and drivers jointly attend to the road.[16] Situational awareness and indeterminacy resolution are often socially shared. We collaborate to disambiguate.

According to Merleau-Ponty, the structures of perceptual indeterminacy, and indeterminacy resolution, resemble the figure–ground structure of pictures. As countless perceptual illusions demonstrate, our perception of the figure (the foregrounded point of focus) is shaped by the background against which it is situated. Taken in isolation, the figure may be ambiguous, but the cues surrounding it, together with our skills for understanding those cues, guide us to perceive the figure in determinate ways. Two papers on the interlinked perception of race and emotion exemplify this figure–ground structure nicely.[17] In the first, participants were more likely to identify ambiguous emotional expressions as angry if they belonged to a black face, but happy if they belonged to a white face. Here the figure (the foregrounded problem to solve) was the emotion being expressed, while the ground (the context covertly guiding judgments about the figure) was constituted, in part, by perceptions of race. In the second paper, figure and ground were reversed: participants were now more likely to identify *racially ambiguous* faces as black when the faces looked angry, but white when they looked happy. In both cases, racial biases in perceptual judgment were predicted by participants' performance on the IAT (but not by their self-reported racial attitudes). Implicit bias, in other words, was also part of the background, shading interpretations about otherwise indeterminate objects of attention. Nor is the indeterminacy-resolving power of implicit bias restricted to split-second judgments. Implicit bias also affects reflective deliberation, leading, for example, mock jurors to judge that ambiguous evidence is more incriminating when defendants are dark skinned.[18] Implicit biases thus figure among the set of "skills" we develop for drawing on contextual cues to resolve indeterminacies. I put "skills" in scare quotes because, although these dispositions are socially learned, they are obviously (and sometimes tragically) biased and misleading.

Phenomenologists use the notion of horizon to characterize these features of experience. There are distinct (but related) uses of this idea, two of which are relevant here. Loosely following Husserl,[19] we can call the first the *internal horizon*, to refer to the range of possible interpretations of a particular percept, with some more central and intuitive, others somewhat strained but still in the ballpark, and still others decisively out of bounds. (Consider trying to identify a blurry letter on an optometrist's chart; perhaps it looks most like a *P*, but it *might* be an *F*, and it's *definitely not* an *E* or a *Z*.) Call the second the *external horizon*, referring to the broader context or field within which each particular percept is experienced, including other percepts as well as background expectations, bodily postures, moods, and so on. Internal horizons refer to particular entities, that is, the range of possible interpretive options of a given percept, while external horizons refer to the context making particular interpretive options more or less salient and determinate.

Heidegger and Gadamer invoke interpretive horizons to understand not just perception but also our relationships to texts and art, and Alcoff appeals to interpretive horizons to understand visible social identities like race. She writes, "The concept of horizon helps to capture the background, framing assumptions we bring with us to perception and understanding, the congealed experiences that become premises by which we strive to make sense of the world, the range of concepts and categories of description that we have at our disposal."[20] Alcoff's view accounts for the open-endedness and freedom involved in self-interpretations of identity, while at the same time explaining how this range of plausible self-interpretations is constrained by experience, embodiment, and social relations. For example, I (a white American) can understand my social identity as a symbol of supremacy and power, as a source of guilt or privilege, as a descendant of immigrants from diverse national origins, and in numerous other ways. But I cannot, at least in current sociopolitical circumstances, understand myself as black; that option is out of bounds. Racially mixed individuals may understand themselves, and be perceived by others, as nonwhite in some contexts and white in others. There is, according to Alcoff, an "indeterminacy of racial categories,"[21] comparable in broad but important strokes to the indeterminacies of perceptual experience and textual interpretation.

I claim the same applies to the experiential contents of our implicit biases: they have an indeterminate character. As Lee, Lindquist, and Payne put it, "implicit affect toward outgroups serves as an *ambiguous signal* that is available to be conceptualized as different discrete emotions based on the context."[22] Implicit bias exists in a holistic relationship with a range of other factors, any of which may, when taken in isolation, be indeterminate—neither white nor black, neither angry nor happy, and neither biased nor unbiased—but each of which can become determinate in context, via relations to the others.

IMPLICIT BIAS FROM BACKGROUND TO FOREGROUND

One of the most prominent debates about implicit bias in social psychology and philosophy has been about cognitive structure. On the received view, implicit biases are stored in long-term memory in a network of semantic associations. Part of what inspired and sustains this view is that leading *measures* of implicit bias are associative in nature. They assess, in various ways, how quickly or likely participants are to pair stimuli, such as images of racially typical faces with images of weapons. In recent years, however, an alternative interpretation has gained ground, that implicit biases are language-like, propositional structures, which can update swiftly in light of the evidence. This propositional interpretation has been buoyed by a raft of studies demonstrating, for example, that an isolated piece of relevant information is sometimes sufficient to shift individuals' performance on implicit measures, in patterns consistent with the rational revision of belief but harder to square with the intensive reconditioning presumably required for rewiring ingrained associations.[23]

Phenomenologists may recognize in this debate the echoes of traditional disputes between empiricism or behaviorism on the one hand and rationalism, intellectualism, or cognitivism on the other. Thinkers such as Merleau-Ponty sought to transcend these disputes, emphasizing each approach's insights and oversights. We find ourselves similarly poised with respect to implicit bias. Propositionalists are right that there must be more to implicit biases than mere associations between concepts. Something more substantive must be said about the intentional relations in which the concepts stand. For example, a black–weapons association on the IAT might reflect the belief, perhaps unconscious, that blacks are violent, but couldn't it equally well reflect the (justified and true) belief that blacks are more likely to be *stereotyped* as violent, or indeed that blacks are more likely to be *victims* of weapon-related violence? The sheer fact of the association doesn't distinguish between these interpretations. So if the association were the only evidence we had, we could not say that it amounted to *bias* against blacks rather than, say, an acute acquaintance with the realities of black oppression. As it happens, of course, the association is not the only evidence we have. Myriad studies correlate performance on these associative measures with discriminatory behavior. Even the most clamorous critics of the race IAT grant that it predicts behavior and is at worst comparable in average predictive power to more traditional self-report measures.[24] But once we have evidence tying implicit measures to behavior, we also have evidence that the intentional contents of implicit biases are more than mere associations. The evidence that a black–weapon association predicts a bias toward shooting unarmed black people is also evidence that this association is more closely tied to a racial attitude along the lines of *black people are threatening* than *black people are threatened.*

Yet propositionalists take this insight too far and overestimate implicit bias's determinacy. The full range of behavior predicted by implicit bias is surprisingly broad and mercurial, much more so than in the case of propositional attitudes like belief and desire (at least as they are traditionally understood). In some conditions, an ostensibly "biased" IAT score correlates with prosocial and arguably ethical, rather than discriminatory, behavior. The associative approach is therefore right to suggest that the relations between concepts are open-ended but wrong to leave them *too* open-ended, as if implicit biases were altogether devoid of intentional content. By contrast, the propositional approach is right that implicit biases are contentful but wrong to portray their contents as more precise than they actually are.

The most tried-and-true method for knowing the contents of people's minds (what they want, believe, etc.) is to ask them. Individuals who sincerely believe that *P* will be disposed to assert that *P* when asked (other things equal and in appropriate conditions, e.g., when they want to tell the truth). This strategy might seem unavailable for implicit biases because they are often glossed as opaque to introspection. Yet it has been relatively clear for some time that implicit biases are conscious—or at least no less conscious than so-called explicit attitudes. Leading theorists in both the associative and propositional camps argue that conscious awareness has numerous roles to play in a full accounting of the causes, effects, and nature of implicit bias. How, then, to distinguish explicit from implicit? One influential theory distinguishes between propositional and associative processes, rather than conscious and unconscious representations.[25] This theory describes the outputs of associative processes as spontaneous "affective reactions" to stimuli or, more colloquially, as "gut feelings," which are qualitatively felt, as in the immediate sense of discomfort a white person might feel during an interracial encounter. The tendency for race-related gut feelings to go unreported does not reflect their being unconscious, unintentional, or automatic. Rather, whether they are reported depends on how they are interpreted in a given context, which in turn depends on a host of further facts about the *interpreter*. Here propositional (i.e., reflective, inferential) processes enter the scene.

On this view, when people with negatively valenced spontaneous reactions toward blacks are asked about their attitudes, they will, holding all else equal, say something along the lines of, "I dislike black people" or "Blacks are unpleasant." Most six-year-old children, for example, readily report racial preferences, while ten-year-olds are less likely to do so, and adults are much less likely still.[26] Why do people become less likely to report racial biases as they age? Research suggests that adults' willingness to report their spontaneous reactions depends in part on their other beliefs and values—that is, their interpretive horizons. People who recognize that "Black people represent a disadvantaged minority group" and that "negative evaluations of disadvantaged minority groups are wrong" will infer upon reflection that negative evaluations of black people are wrong—unjustified, inaccurate, or immoral.[27] This means that their own spontaneous negative reactions are wrong, and many individuals resolve the perceived inconsistency (cognitive dissonance) between their biased feelings and egalitarian commitments by reporting that they like blacks and nonblacks equally. Note that this "failure" to self-report racial preferences need not involve intentional misreporting or self-deception. They may be pristinely aware of the prima facie problematic implications of their gut feelings, yet sincerely reject those feelings on the grounds that they do not represent their considered opinion, much as someone can be aware of their phobic or superstitious impulses but recognize that they are not all-things-considered justified. Thus, when given the opportunity to separately report both their "gut feelings" and their "actual feelings," participants' self-reported gut feelings correlate more strongly with implicit than explicit measures.[28]

By contrast, "old-fashioned" supremacists, who believe that negative evaluations of low-status racial minorities are entirely appropriate (e.g., because they believe racial minorities are actually inferior or are otherwise threats to their way of life), generally do not hesitate to say so. Such individuals have, in effect, a direct line of communication between their immediate affective dispositions and their explicit reports. Similarly, more "modern" racists—who nominally accept that negativity toward the disadvantaged is wrong but who deny, as a factual matter, that blacks continue to be disadvantaged—are also more open about their negative feelings. These individuals might attribute the low social status of many African Americans to problematic aspects of their “culture” or “values” or to faulty “personal choices” rather than to oppression. Perceiving blacks as responsible for their own hardships is then taken to license explicit negative evaluations, such as blame or condescension. In sum, individuals who don’t believe, or simply don’t care, that blacks continue to be oppressed will be more likely to report than disavow their gut feelings. Only those who *both* endorse antiracist values *and* believe in the persistence of racial oppression will refrain from reporting negative racial sentiments.

Now is as good a time as any to note that such findings have straightforward implications for meta-analyses on the correlations between implicit and explicit measures of bias, and between either measure and “real world” behavior.[29] Whether people report their biases, or act on them, depends fundamentally on how those biases interact with other psychological and contextual factors. Ignoring such factors would be analogous to running a meta-analysis of studies examining whether striking a match leads it to catch fire without keeping track of whether, in the preponderance of experiments, there was any oxygen in the room or the matches were soaking wet. There is no “pure” correlation to expect between match striking and match lighting without accounting for such unassailably essential background conditions. Meta-analyses of implicit bias that ignore psychological and social context are, therefore, largely uninformative. Nevertheless, there is nothing inherent to meta-analytic research that precludes coding for context, and meta-analyses that do so find that correlations between implicit measures, explicit measures, and real-world behavior vary to a significant extent in keeping with theory-based predictions.[30]

I have yet, however, to discuss the best evidence for indeterminism. The thrust of the aforementioned studies might even seem to run in the other direction. Psychologists claim that participants’ social-affective reactions “imply” concrete evaluative judgments, most naturally expressed with statements of (dis)liking. Of course, a mere association of “black” with “bad” cannot, on its own, imply anything, because it lacks intentional content. So it might seem that, to make sense of these claims, we have to grant that implicit biases characteristically do have determinate content, with a canonical or default articulation along the lines of, “I dislike members of group G.”[31] Now, on my view, this sort of self-ascription of disliking may frequently be among the more central, intuitive options for interpreting implicit bias, but there are grounds for questioning whether it or anything else constitutes the precise, canonical articulation. Why, in particular, should the dispositions at issue be exclusively associated with (dis)liking rather than other forms of affect or motivation? Consider, for example, the finding that antiblack implicit bias did not correlate with any particular self-reported emotion toward black people (whether fear, anger, guilt, etc.), but that it did correlate with the average of all these negative emotions taken together.[32] When implicitly biased people report their racial attitudes, they are more likely to say *something negative* than something positive, but little beyond that is settled. Implicit bias lacks a unique emotional signature and instead reflects a generic, vague negativity. Evidence associating implicit bias with this sort of vague negativity may be why open-ended dislike often seems a natural interpretation, but under certain conditions, this vague negativity can be channeled into distinctive emotional reactions not best interpreted that way.

This brings us to the studies most suggestive of indeterminism. First, participants completed an implicit measure of their spontaneous affective reactions to images of white versus black faces. One group of participants was then told that the gut feelings they may have had during the measure reflected *fear* of blacks; another group was told that these feelings reflected *sympathy* toward blacks. Participants were then asked to generate two or three reasons why they might have felt fear or sympathy, respectively. Subsequently, those who both tested high in implicit bias and who were instructed to interpret their feelings as fear now tended to agree with statements like "Blacks are scary" and "Blacks are threatening." However, implicitly biased participants in the sympathy condition tended *not* to report explicit fear of blacks. (Another study found the same pattern simply by measuring, rather than manipulating, participants' antecedent beliefs about whether their gut feelings reflected fear versus sympathy.) Negative affect per se did not, just as such, correlate with self-reported fear of African Americans, unless participants *noticed and interpreted* their negative affect as fear and felt permitted to say so.

As a phenomenologist, I interpret these studies as illustrating the open-endedness, and pervasive potential for distortion, when we step back and reflect on experience, switching gears from our habitual, "ready-to-hand" mode of being-in-the-world to a more theoretical posture. These studies exemplify the extent to which our unthematized social experience is up for interpretive grabs. Compare Merleau-Ponty's account of the transition from indeterminacy to determinacy that results when reflective attention is directed upon prereflective experience:

> Attention, then, is neither an association of ideas nor the return to itself of a thought that is already the master of its objects; rather, attention is the active constitution of a new object that develops and thematizes what was until then only offered as an indeterminate horizon. . . . The object only gives rise to the "knowing event" that will transform it through the still ambiguous sense that it offers to attention as needing-something-to-be-determined, such that the object is the "motive" of and not the cause of this event. . . . This passage from the indeterminate to the determinate, this continuous taking up again of its own history in the unity of a new sense, is thought itself.[33]

Specifically, these studies highlight the meaningful but indeterminate relations between our immediate affective dispositions and our concrete, articulated emotions, and walk in lockstep with accounts of prereflective affectivity offered by phenomenologists. For example, citing Husserl and Merleau-Ponty, Alia Al-Saji writes that "the realm of affectivity is wider than what can be called *emotion*, since emotion is an intentional, sense-giving relation (to an object) that is built on affect, whereas affect is the preintentional tendency or force (attraction, repulsion, pain, pleasure, etc.) that can motivate and support this intentional turning toward an object."[34] It is hard to imagine a more apt description of how social psychologists are coming to understand the relations between implicit affect, explicit prejudice, and discriminatory behavior.

Moreover, the power of attention to make concrete meaning out of vague feelings (or, as Merleau-Ponty puts it, to actively constitute new objects out of indeterminate horizons) was not limited to swaying participants' verbal reports. Participants who interpreted their gut feelings as fear were, on a subsequent task, more likely to perceive emotionally ambiguous black faces as angry. This result recalls, but also contextualizes, the findings mentioned above on implicit bias and the perception of emotion and race. Those results suggested, at first glance, that implicit bias *just as such* influenced judgments about ambiguous percepts, but the horizon of holistic interconnections is evidently more complex: the effects of gut feelings on perceptual judgment are (always already) mediated by self-interpretations of what those gut feelings mean. Self-interpretation forms part of the background shaping the meaning of implicit bias, which in turn forms part of the background shaping the interpretation of perception.

FURTHER TAKEAWAYS

Constructionism, Existentialism, and the Bounded Horizons of Interpretation

There are further notable findings and takeaways. First, although these studies indicate the wide horizon of possible implicit bias interpretations, they simultaneously reveal that this horizon is bounded. Participants who were instructed to interpret their gut feelings as sympathy were less likely to report fear of blacks, but they were not, it turns out, more likely to report sympathy. The experiment *reduced* self-reports of fear but did not *increase* self-reports of sympathy. In this context, then, implicit biases were sufficiently flexible as to be interpreted as either fear or not-fear, but not so completely up for grabs as to be interpreted as any emotion whatsoever. This evidence of nontrivial constraints on the range of feasible interpretations speaks against radically constructionist or existentialist approaches that would attribute unlimited freedom to the mind's self-interpreting and self-constituting powers. This evidence, for example, qualifies the apparently radical constructionist implications of Schachter and Singer's (1962) notorious (and notoriously difficult to replicate) studies, which found that participants injected with adrenaline reported profoundly different emotional experiences depending on available contextual cues. We must disagree with the Hamlet- or Sartre-inspired interlocutor who would assert that "nothing's [fear or sympathy] but thinking makes it so."

These findings resonate instead with a model of implicit (and explicit) bias as indeterminate, interpretive horizon. As Alcoff claims about racial identity and history, so it is with the content of implicit bias: "dynamic and unstable, but its meanings are not completely indeterminate or infinitely flexible, and they are not forged by any individual alone."[35] For a particular individual experiencing a particular gut feeling, certain interpretations and actions are more straightforwardly afforded by the context, while others are off limits, with a gray border area in between.

Power, Situation, and Individual Bias

Social psychology also makes vivid how the passage of bias from indeterminate to determinate is shaped by the broader situation, outside the individual. In several studies, participants are (overtly or covertly) encouraged by authority figures into conceiving of their gut feelings in particular ways, thereby illustrating the intimate, self-interpretive depths to which power relations can reach. (Interpretations of racial gut feelings are not forged by any individual alone.) These studies thus recall the classic situationist experiments by the likes of Milgram, Zimbardo, and Sherif, as further lessons in the power of authority, norms, and the like.[36] Typically, the authority figure in these experiments is a scientist (and participants themselves tend to be psychology majors, who are especially likely to value psychological research and defer to psychologists' summaries of what that research means). But there is no reason to assume that scientists are alone in occupying positions of influence over others' interpretive horizons; parents, professors, politicians, and religious leaders presumably also share the power to shape how we interpret our own minds. Authoritative others have the power to activate, legitimize, and even mold our inchoate gut feelings, transforming vague feelings of social discomfort into concrete emotional experiences of fear or indignation, and indeterminate implicit biases into explicitly endorsed discrimination. It is not just that authority figures "take advantage of" or "exploit" preexisting biases, but that they play a significant role in making those biases what they are.

Yet even as these studies demonstrate the profound power of the situation, they simultaneously undermine the conventional situationist narrative, according to which factors "external" to the individual are somehow *more* powerful drivers of behavior than factors "internal" (personality traits, moral commitments, etc.). This approach to situational influence is foundationally flawed, and these studies help explain why. For example, only those participants who tested *high* in implicit bias were influenced by the manipulations telling them how to interpret their gut feelings. Participants who demonstrated little or no implicit bias were immune to demagogic testaments to the validity of their negative feelings, for the obvious reason that they didn't have negative feelings to validate. Implicit biases thus constitute an important *individual-difference variable*, determining whether and how situational factors shape thought and action. Numerous other personal beliefs and traits also have decisive roles to play. For example, telling (implicitly biased) participants that their gut feelings represent their "real" attitudes and "genuine" selves can also lead them to report explicit prejudice and support for discriminatory policies;[37] however, this manipulation primarily affects only those (implicitly biased) participants who are also high in self-esteem, that is, those antecedently disposed to have positive views of their "selves." Thus, although studies like this are clearly reminiscent of classic situationist findings, they also represent a decisive departure from the narratives passed down about those findings. The core contrast between person and situation, or in sociological contexts between agency and structure, is confused. Situations do not operate by themselves to shape self-interpretation and action, but only in conjunction with individuals' other (potentially idiosyncratic) attitudes, habits, traits, experiences—and implicit biases. The power of situations depends, fundamentally, on the minds of those in them.

In a final exemplification of indeterminacy, some studies even find that, in the right metacognitive context, implicit biases become cues to act virtuously.[38] Specifically, individuals who feel acutely aware and even guilty about their biased feelings, but who are earnestly committed to being unprejudiced, can effectively learn to reinterpret those feelings, not as invitations to be biased, but as palpable, internal signals to be just. Such findings speak to strategies for ameliorating injustice, which I explore more fully elsewhere.[39] Notably, for example, when diversity trainers stress that the "vast majority of people" harbor implicit biases, trainees become *more* biased, but when trainers stress that the "vast majority of people try to overcome" their implicit biases, trainees became *less* biased.[40] Collectively telling ourselves that our biases are inaccurate, unintentional, unrepresentative of our underlying commitments, and ultimately to be overcome, may be key to discouraging the implicit from bubbling up into the explicit and motivating us to unlearn our biases altogether.

OBJECTIONS AND ALTERNATIVE VIEWS

Epistemicism

I take such findings to reveal the profound indeterminacy inherent in our implicit biases. An alternative interpretation is that individuals simply don't *know* what the contents of their attitudes are.[41] This alternative locates the problem in epistemology rather than metaphysics, concluding that introspective access to implicit bias is limited, and corresponds to Williamson's "epistemicist" take on vagueness, which argues that apparent borderline cases (e.g., between being bald and having hair) are actually determinate; we just don't, perhaps *can't*, know whether they are one way or the other.[42] There is, on this line of thinking, a fact of the matter about whether implicitly biased individuals do or don't fear black people, but they (and we) may not know either way.

Let me first note my agreement that these studies highlight various obstacles to self-knowledge. A central theme of the phenomenological tradition, after all, is that mental life is not transparent to introspection. "Nothing is more difficult," Merleau-Ponty writes, "than knowing precisely *what we see*."[43] My view is that many of us fail to know that the contents of our racial attitudes are indeterminate and context sensitive. This is not because our attitudes are unconscious, only difficult to interpret. It is, however, unclear what empirical evidence or metaphysical ur-facts could settle that our biases have one precise content rather than another. If we take a broadly functionalist approach to individuating mental content (and what other approach is there?), then their content is manifestly underspecified. Can we appeal to self-report, and attribute beliefs based on people's sincere assertions? Not when the relations between implicit bias and sincere assertion are, to put it mildly, a big mess. The same is true if we look beyond self-report, for example to prereflective nonverbal behavior or perceptual judgment. White people with antiblack implicit biases are not, simply thereby, more likely to see black faces as angry; they must also interpret their bias as fear. So I am skeptical that there are pure, determinate facts about the content of implicit bias, which we then fail to know. What we *can* know is that this bias in this mind in this body in this environment is, say, fear of African Americans. But once we appreciate that each potential context plays a key role in shaping the meaning of our racial attitudes, then we should acknowledge that there is no neutral context that could, even in principle, reveal those attitudes in some unalloyed form.

Disjunctivism

Emphasizing the context-specificity of implicit bias, however, suggests another interpretation: disjunctivism. On this line (which is compatible with epistemicism), giving a full accounting of a given individual's attitudes toward blacks might require us to say that they feel *fear-in-context-C* but *sympathy-in-context-D*, and so on. This individual might not *generally* fear black people—suppose, for example, he reliably marches and votes against policies that criminalize denizens of the black ghetto—but he does (determinately) fear tall dark-skinned men wearing hoodies when walking alone at night. Disjunctivism rests satisfied with a mere listing of all these context-indexed dispositions and denies that there is anything substantive to say about his racial attitudes that remains true across contexts.

As with epistemicism, I have some sympathies with disjunctivism. I believe our racial attitudes (often) become determinate once embedded in context, so a sufficiently long and complex disjunction could perhaps describe an individual's overall dispositional profile. But would this disjunction constitute the most accurate and perspicuous way to represent the contents of his racial attitudes? I doubt it. First, since disjunctivism refuses to say anything about what underlies or unifies the individual's diverse race-related dispositions, it seems tantamount to abandoning an explanatory or illuminating account of racial bias altogether. Second, since the full disjunction may well be infinitely long, the disjunctivist tack makes even the *descriptive* project of capturing the phenomenology and psychology of racial bias come to seem hopeless. Disjunctivism is, then, not just compatible with, but may in fact entail, epistemicism. There are indefinitely many unknown and perhaps even unknowable contexts that could shape and be shaped by our implicit biases, making the true contents of those biases in principle unknowable as well. So it seems preferable to sum up individuals' racial attitudes with an admirably short and sweet, but regrettably vague and open-ended content (namely, that they are implicitly biased!), with the understanding that this vagueness will often be determinately resolved in specific cases.

That disjunctivism devolves into quietism and ignorance about both self and other may seem either a virtue or a vice, depending on one's other philosophical leanings. My grounds for preferring indeterminism are also practical and similar in spirit to Haslanger's critical-theoretical approach to race and gender, which asks which metaphysical views of these categories best serve our political aims.[44] Similarly, we might ask how best to conceptualize implicit bias for the purpose of resisting injustice. Which view is most useful, whether for generating new empirical research, or for motivating social change? Both epistemicism and disjunctivism seem to me to make the role for individual and collective responsibility obscure. How can we take responsibility for our biases if we can't know what they are, or if what we do in one context is wholly irrelevant to what we do in another?

My hunch is that with great indeterminacy comes great responsibility. Indeterminism makes salient that the biases we harbor hold the potential to guide us toward both just and unjust action—depending on our other habits, beliefs, and values, and the contexts in which we find ourselves. As Emily Lee writes, "why value openness and ambiguity? Open and ambiguous knowledge permits the possibility of becoming and change."[45] So it is with indeterminism about implicit bias, which brings into sharpest relief the potential for change, both for better and for worse. Each of us is individually responsible for interpreting, reining in, and ultimately eradicating our implicit biases, as well as for holding others accountable for doing the same. We are, moreover, collectively responsible for structuring social environments that, among other things, encourage more accurate and virtuous self-interpretations and discourage vicious ones.

But if I have convinced you of nothing else, I hope it is at least clear that the received aversive racism narrative, which portrays most Americans as unambiguously antiracist on the explicit level and irredeemably biased on the implicit level, is due for a rewrite. Our conscious egalitarianism is far more fragile and subject to contextual variation than commonly appreciated (or at least than was appreciated until about June 2015), and our implicit biases are also up for contextual grabs. Are these social gut feelings conscious or unconscious, reflectively endorsed or disavowed, representative of our real selves or alien implantations by external cultural forces? Are we afraid of the racial other or just afraid of appearing racist? Depending on context, the answer can be any or none of the above. We must own up to as much, and in particular own up to disowning our biases.[46]

NOTES

[1] See Pearson, Dovidio, and Gaertner, "Nature of Contemporary Prejudice."

[2] This chapter focuses on the spontaneous reactions driving performance on measures like the IAT. Yet our minds are populated with an array of arguably "implicit" cognitive structures that contribute to discrimination but may evade detection on IATs. See Del Pinal, Madva, and Reuter, "Stereotypes, Conceptual Centrality and Gender Bias."

[3] Madva, "Implicit Bias, Moods, and Moral Responsibility."

[4] Weiss, "Sedimented Attitudes and Existential Responsibilities," 97 n. 14; see also 93, 99–100 n. 43.

[5] Alcoff, "Sotomayor's Reasoning," 130 n. 17.

[6] Glaser and Knowles, "Implicit Motivation to Control Prejudice."

[7] Hehman, Flake, and Calanchini, "Disproportionate Use of Lethal Force in Policing."

[8] On implicit bias and criminal justice, see Cholbi and Madva, "Black Lives Matter and the Call for Death Penalty Abolition," sec. II.

[9] See also Brownstein and Madva, "Normativity of Automaticity."

[10] Eric Schwitzgebel defends another kind of indeterminacy in Americans' racial attitudes, such that aversive racists *neither believe nor fail to believe* that all races should be treated

equally, in "Acting Contrary to Our Professed Beliefs." I argue that the very contents of our racial attitudes are indeterminate, although much of the evidence to follow may be amenable to Schwitzgebel's analysis also.

11. Alcoff, *Future of Whiteness*, 74–90; Madva, "Plea for Anti-anti-individualism."

12. E.g., Ngo, "Racist Habits," 854.

13. Merleau-Ponty, *Phenomenology of Perception*, 7, 12.

14. Romdenh-Romluc, *Routledge Philosophy GuideBook to Merleau-Ponty*, 18.

15. Merleau-Ponty, *Phenomenology of Perception*, 222.

16. Drews, Pasupathi, and Strayer, "Passenger and Cell Phone Conversations in Simulated Driving."

17. Hugenberg and Bodenhausen, "Facing Prejudice" and "Ambiguity in Social Categorization." In future work, I plan to discuss numerous subsequent studies finding similar patterns.

18. Levinson and Young, "Different Shades of Bias."

19. Husserl, *Experience and Judgment*, sec. 8.

20. Alcoff, *Visible Identities*, 95.

21. Ibid., 179.

22. Lee, Lindquist, and Payne, "Constructing Bias," 4; emphasis added.

23. Cone, Mann, and Ferguson, "Changing Our Implicit Minds."

24. Oswald et al., "Predicting Ethnic and Racial Discrimination."

25. Gawronski and Bodenhausen, "Associative and Propositional Processes in Evaluation."

26. Dunham, Baron, and Banaji, "Development of Implicit Intergroup Cognition."

27. Gawronski et al., "Understanding the Relations between Different Forms of Racial Prejudice," 650.

28. Ranganath, Smith, and Nosek, "Distinguishing Automatic and Controlled Components of Attitudes."

29. Brownstein, Madva, and Gawronski, "Understanding Implicit Bias."

30. Cameron, Brown-Iannuzzi, and Payne, "Sequential Priming Measures of Implicit Social Cognition."

31. This account of the inferential relations between spontaneous affective reactions and propositional judgments is puzzling, given that these theorists also argue that implicit biases are stored as mere contentless associations (e.g., Gawronski et al., "Understanding the Relations between Different Forms of Racial Prejudice," 650, 660). How can a mere association "lead to" a gut feeling that in turn "implies" anything unless the original association also has intentional content? Are social psychologists running afoul of the "Myth of the Given"? Thanks to Gabby Johnson for discussion here, which I hope to address in future research.

32. Lee, Lindquist, and Payne, "Constructing Bias."

33. Merleau-Ponty, *Phenomenology of Perception*, 55.

34. Al-Saji, "Phenomenology of Hesitation," 162 n. 2.

35. Alcoff, *Visible Identities*, 115.

36. I had hoped to say more here about implicit bias, norms, and hermeneutical injustice, but space limitations require deferring that discussion to future work.

37. Cooley et al., "Who Owns Implicit Attitudes?"

38. Burns, Monteith, and Parker, "Training Away Bias."

39. Madva, "Virtue, Social Knowledge, and Implicit Bias"; Madva, "Biased against Debiasing"; Madva, "Inevitability of Aiming for Virtue."

40. Duguid and Thomas-Hunt, "Condoning Stereotyping?"

41. Thanks to Andreja Novakovic for discussion.

42. Williamson, *Vagueness*.

43. Merleau-Ponty, *Phenomenology of Perception*, 59.

44. Haslanger, *Resisting Reality*.

45. Lee, "Towards a Lived Understanding of Race and Sex," 85.

46. For insightful feedback, I am grateful to Michael Cholbi, Peter Ross, and especially Emily Lee, as well as to audiences at Washington University's "Psychology of Prejudice" workshop (November 2017), Wake Forest University (April 2018), and Cal Poly Pomona's "brown bag" workshop (April 2018).

WORKS CITED

Alcoff, Linda Martín. *The Future of Whiteness*. Malden, MA: Polity, 2015.

———. "Sotomayor's Reasoning." *Southern Journal of Philosophy* 48 (2010): 122–38. https://doi.org/10.1111/j.2041-6962.2010.01005.x.

———. *Visible Identities: Race, Gender, and the Self.* Oxford: Oxford University Press, 2006.

Al-Saji, Alia. "A Phenomenology of Hesitation: Interrupting Racializing Habits of Seeing." In *Living Alterities: Phenomenology, Embodiment, and Race*, edited by Emily S. Lee, 133–72. Albany: State University of New York Press, 2014.

Brownstein, Michael, and Alex Madva. "The Normativity of Automaticity." *Mind & Language* 27 (2012): 410–34. https://doi.org/10.1111/j.1468-0017.2012.01450.x.

Brownstein, Michael, Alex Madva, and Bertram Gawronski. "Understanding Implicit Bias: Putting the Criticism into Perspective." Unpublished paper.

Burns, Mason D., Margo J. Monteith, and Laura R. Parker. "Training Away Bias: The Differential Effects of Counterstereotype Training and Self-Regulation on Stereotype Activation and Application." *Journal of Experimental Social Psychology* 73 (2017): 97–110. https://doi.org/10.1016/j.jesp.2017.06.003.

Cameron, C. Daryl, Jazmin L. Brown-Iannuzzi, and B. Keith Payne. "Sequential Priming Measures of Implicit Social Cognition: A Meta-analysis of Associations with Behavior and Explicit Attitudes." *Personality and Social Psychology Review* 16 (2012): 330–50. https://doi.org/10.1177/1088868312440047.

Cholbi, Michael, and Alex Madva. "Black Lives Matter and the Call for Death Penalty Abolition." *Ethics* 128 (2018): 517–44. https://doi.org/10.1086/695988.

Cone, Jeremy, Thomas C. Mann, and Melissa J. Ferguson. "Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised." *Advances in Experimental Social Psychology*, 56 (2017): 131–99. https://doi.org/10.1016/bs.aesp.2017.03.001.

Cooley, Erin, B. Keith Payne, Chris Loersch, and Ryan Lei. "Who Owns Implicit Attitudes? Testing a Metacognitive Perspective." *Personality and Social Psychology Bulletin* 41 (2015): 103–15. https://doi.org/10.1177/0146167214559712.

Del Pinal, Guillermo, Alex Madva, and Kevin Reuter. "Stereotypes, Conceptual Centrality and Gender Bias: An Empirical Investigation." *Ratio* 30 (2017): 384–410. https://doi.org/10.1111/rati.12170.

Drews, Frank A., Monisha Pasupathi, and David L. Strayer. "Passenger and Cell Phone Conversations in Simulated Driving." *Journal of Experimental Psychology: Applied* 14 (2008): 392–400. https://doi.org/10.1037/a0013119.

Duguid, Michelle M., and Melissa C. Thomas-Hunt. "Condoning Stereotyping? How Awareness of Stereotyping Prevalence Impacts Expression of Stereotypes." *Journal of Applied Psychology* 100 (2015): 343–59. https://doi.org/10.1037/a0037908.

Dunham, Yarrow, Andrew S. Baron, and Mahzarin R. Banaji. "The Development of Implicit Intergroup Cognition." *Trends in Cognitive Sciences* 12 (2008): 248–53. https://doi.org/10.1016/j.tics.2008.04.006.

Gawronski, Bertram, and Galen V. Bodenhausen. "Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change." *Psychological Bulletin* 132 (2006): 692–731. https://doi.org/10.1037/0033-2909.132.5.692.

Gawronski, Bertram, Kurt R. Peters, Paula M. Brochu, and Fritz Strack. "Understanding the Relations between Different Forms of Racial Prejudice: A Cognitive Consistency Perspective." *Personality and Social Psychology Bulletin* 34 (2008): 648–65. https://doi.org/10.1177/0146167207313729.

Glaser, Jack, and Eric D. Knowles. "Implicit Motivation to Control Prejudice." *Journal of Experimental Social Psychology* 44 (2008): 164–72.

Haslanger, Sally, *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press, 2012.

Hehman, Eric, Jessica K. Flake, and Jimmy Calanchini. "Disproportionate Use of Lethal Force in Policing Is Associated with Regional Racial Biases of Residents." *Social Psychological and Personality Science* 9, no. 4 (2017): 393–401. https://doi.org/10.1177/1948550617711229.

Hugenberg, Kurt, and Galen V. Bodenhausen. "Ambiguity in Social Categorization: The Role of Prejudice and Facial Affect in Race Categorization." *Psychological Science* 15 (2004): 342–45. https://doi.org/10.1111/j.0956-7976.2004.00680.x.

———. "Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat." *Psychological Science* 14 (2003): 640–43. https://doi.org/10.1046/j.0956-7976.2003.psci_1478.x.

Husserl, Edmund. *Experience and Judgment.* Evanston, IL: Northwestern University Press, 1973.

Lee, Emily S. "Towards a Lived Understanding of Race and Sex." *Philosophy Today* 49 (2005): 82–88.

Lee, Kent M., Kristen A. Lindquist, and B. Keith Payne. "Constructing Bias: Conceptualization Breaks the Link between Implicit Bias and Fear of Black Americans." *Emotion* 18, no. 6 (2018): 855–71. https://doi.org/10.1037/emo0000347.

Levinson, Justin D., and Danielle Young. "Different Shades of Bias: Skin Tone, Implicit Racial Bias, and Judgments of Ambiguous Evidence." *West Virginia Law Review* 112 (2010): 307–50.

Madva, Alex. "Biased against Debiasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle against Prejudice." *Ergo, an Open Access Journal of Philosophy* 4, no. 6 (2017): 145–79. http://dx.doi.org/10.3998/ergo.12405314.0004.006.

———. "Implicit Bias, Moods, and Moral Responsibility." *Pacific Philosophical Quarterly* 99, no. S1 (2017): 53–78. https://doi.org/10.1111/papq.12212.

———. "The Inevitability of Aiming for Virtue." In *Overcoming Epistemic Injustice*, edited by by Benjamin R. Sherman and Stacey Goguen. Lanham, MD: Rowman & Littlefield, forthcoming.

———. "A Plea for Anti-anti-individualism: How Oversimple Psychology Misleads Social Policy." *Ergo, an Open Access Journal of Philosophy* 3, no. 27 (2016): 701–28. https://doi.org/10.3998/ergo.12405314.0003.027.

———. "Virtue, Social Knowledge, and Implicit Bias." In *Implicit Bias and Philosophy: Metaphysics and Epistemology*, vol. 1, edited by Michael Brownstein and Jennifer Saul, 191–215. Oxford: Oxford University Press, 2016. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198713241.001.0001/acprof-9780198713241-chapter-8.

Merleau-Ponty, Maurice. *Phenomenology of Perception*. Translated by Donald Landes. Abingdon, UK: Routledge, 2013.

Ngo, Helen. "Racist Habits: A Phenomenological Analysis of Racism and the Habitual Body." *Philosophy & Social Criticism* 42, no. 9 (2016): 847–72. https://doi.org/10.1177/0191453715623320.

Oswald, Frederick L., Gregory Mitchell, Hart Blanton, James Jaccard, and Philip E. Tetlock. "Predicting Ethnic and Racial Discrimination: A Meta-analysis of IAT Criterion Studies." *Journal of Personality and Social Psychology* 105, no. 2 (2013): 171–92. https://doi.org/10.1037/a0032734.

Pearson, Adam R., John F. Dovidio, and Samuel L. Gaertner. "The Nature of Contemporary Prejudice: Insights from Aversive Racism." *Social and Personality Psychology Compass* 3, no. 3 (2009): 314–38. https://doi.org/10.1111/j.1751-9004.2009.00183.x.

Ranganath, Kate A., Colin Tucker Smith, and Brian A. Nosek. "Distinguishing Automatic and Controlled Components of Attitudes from Direct and Indirect Measurement Methods." *Journal of Experimental Social Psychology* 44 (2008): 386–96.

Romdenh-Romluc, Komarine. *Routledge Philosophy GuideBook to Merleau-Ponty and Phenomenology of Perception*. Abingdon, UK: Routledge, 2010.

Schwitzgebel, Eric. "Acting Contrary to Our Professed Beliefs or the Gulf between Occurrent Judgment and Dispositional Belief." *Pacific Philosophical Quarterly* 91, no. 4 (2010): 531–53. https://doi.org/10.1111/j.1468-0114.2010.01381.x.

Weiss, Gail. "Sedimented Attitudes and Existential Responsibilities." In *Body/Self/Other: The Phenomenology of Social Encounters*, edited by Luna Dolezal and Danielle Petherbridge, 75–102. Albany: State University of New York Press, 2017.

Williamson, Timothy. *Vagueness*. New York: Routledge, 1994.