A Justice-Oriented Account of Moral Responsibility for Implicit Bias

by

Robin Zheng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2015

Doctoral Committee:
Professor Elizabeth S. Anderson, Chair
Professor Sarah Buss
Professor Shinobu Kitayama
Professor Peter A. Railton

2015

To my friends, my family, and the fight, in the broadest senses

ACKNOWLEDGMENTS

As an avid reader of acknowledgements, hungry for the human stories behind the work, I regret that I can but barely evoke all the stories of the people mentioned in these pages, without whom I could not have done any of this. My first thanks must go to my first philosophy teachers—Melina Bell, Nathaniel Goldberg, James Mahon, and Angie Smith—as I continue working toward the hopes they had for me, and to my committee—Elizabeth Anderson, Sarah Buss, and Peter Railton—whose conversations on a snowy April drew me here and have never since ceased to enthrall, challenge, and guide me these past six years. To the other wonderful philosophers with whom I have had the privilege of being a colleague, especially my letter-writers, and most especially Jim Joyce, for his valiant labors on the market, and Laura Ruetsche, under whose care as department chair I never felt more able to flourish (and chuckle). Also to my other mentors, Jenny Saul and Kristie Dotson, leading the revolution and keeping us alive. And to the department staff—Linda, Jude, Jen, Kim, and Molly—without whom I officially would never have completed my program requirements.

Many of my teachers, however, I have found amongst my graduate student colleagues. My heartfelt thanks, then, to Daniel J. Singer and J. Dmitri Gallow, Chloe Armstrong and Anna Edmonds, and especially to Nathaniel ~~Coleman~~, Shen-yi Liao, Lina Jansson, Nate Charlow, and Eduardo García Ramírez—my fellow skydivers and mentors, my first year and beyond. To Chip Sebens, interior decorator extraordinaire and fellow Peter-the-Cat-lover. To Sara and Reza, purveyors of fine food, accommodation, and even finer conversation. To Fown and to Nils, my housemate, cohort-mate, co-author, and the best mate anyone could ask for. To Annette and Vittorio, especially Annette, whom I admire more than I know how to express, and who I know will continue to do great things.

I am also grateful for feedback on these chapters from audiences at the 2012 Implicit Bias and Philosophy Project Workshop at the University of Sheffield, the 2013 Eastern APA Division Meeting, the 2014 Bouchet Conference on Diversity and Graduate Education at Yale University, the 2014 Implicit Bias Mini-Conference at Washington and Lee University's Roger Mudd Center for Ethics, the 2015 Conference on Interdisciplinary Perspectives on Moral Responsibility at Utah Valley University, and the Departments of Philosophy at North Carolina State University, the University of Kentucky, Carnegie Mellon University, the University of Nottingham, and Yale-NUS College.

The more I learned to do philosophy, the more I realized I needed to learn and do things outside philosophy. I am indebted to everyone who indulged and encouraged me in these endeavors. To Jeanine Delay, whose partnership at A2Ethics has been invaluable. To Ian Robinson and the Huron Central Valley Labor Council, for inviting me to share their vision. To those same people I kept seeing everywhere—L.E. Hunter, Cass Adair, Liz Rodrigues, and Jennifer Alzate González—who inspired me to do more. Especially to Austin, who taught me to appreciate history, and how to make it. To Will, my buddy in all things and not just lunch. To Naim, for his sunshine (my favorite of nature's blessings), and through the rain. To the stewards and the 2013-14, 2014-15 Officer/Staff corps of the Graduate Employees' Organization (GEO), especially the Boston trio: Persephone, Batia, and John Ware. Most of all to the staff, Denise, Lynne, Dom and Jim, who welcomed me to the labor movement in a coffee shop three years ago. To Simeon, living proof that every last minute matters and the world holds more wonderful surprises than I could have known—for showing me how to be better.

Finally, to my family and my friends who are family. Anqi and Faye, whose beautiful weddings this year reminded me what really matters. To Sisi, Cherry, and Dina. To my sister. And at the end and beginning of everything, my parents, without whom I would be nothing.

TABLE OF CONTENTS

Dedication	ii
Acknowledgments	iii
List of Figures	vi
List of Tables	vii
Abstract	viii
Chapter 1: Implicit Bias and Moral Responsibility	1
Chapter 2: Attributability, Accountability, and Implicit Bias	34
Chapter 3: Expanding the Moral Repertoire: Oughts, Ideals, and Appraisals	61
Chapter 4: Transformative Accountability for Structural Injustice	94
Chapter 5: Concluding Remarks	123

List of Figures

Figure 1: IAT, from Green et al. (2007), "Implicit Bias Among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients" 7

Figure 2: Sequential Priming, from Amodio (2014), "The Neuroscience of Prejudice and Stereotyping" 8

Figure 3: Example of "Don't Shoot," from Corell et al. (2002), "The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals" 11

Figure 4: Example of "Shoot," from Corell et al. (2002), ibid. 11

Figure 5: Two Routes to Responsibility 39

Figure 6: Attributability and Accountability 126

List of Tables

Table 1: Etymologies of Words Denoting Moral Criticism 63

Table 2: Critical Moral Responses 125

Abstract

I defend an account of moral responsibility for implicit bias that is sensitive to both normative and pragmatic constraints: an acceptable theory of moral responsibility must not only do justice to our moral experience and agency, but also issue directives that are psychologically effective in bringing about positive changes in judgment and behavior. I begin by offering a conceptual genealogy of two different concepts of moral responsibility that arise from two distinct sources of philosophical concerns. We are morally responsible for our actions in first sense only when those actions reflect our identities as moral agents—that is, when they are *attributable* to us as manifestations of our character, attitudes, ends, commitments, or values. On the other hand, we are responsible in the second sense when it is appropriate for others to enforce certain demands and expectations on those actions—in other words, to hold us *accountable* for them. I argue that we may sometimes lack attributability for actions caused by implicit bias, but that even then we are still accountable for them. Next, I expand beyond individual actions at a particular time to patterns of action across time, to consider what we can reasonably be expected to do when it comes to avoiding and eliminating implicit bias in our selves. By thinking of these expectations as grounded in imperfect duties, I show, we can expand our moral repertoire to include *non-appraising* critical moral responses, in addition to *appraisal-based* responses such as blame and punishment (which are often counterproductive). Finally, I move beyond the actions of individuals to address the question of responsibility for eliminating the social conditions that breed implicit biases in the first place. I argue that accountability requires us not only to *conform* to a system of demands and expectations, but also to collectively organize to *reform* the system itself. By elaborating these multiple dimensions of moral responsibility—attributability versus accountability, particular versus patterns of action, individual versus collective—I demonstrate that the project of developing better practices of moral responsibility is continuous with, and thus contributes to, larger struggles for social equality and justice.

Chapter 1

Implicit Bias and Moral Responsibility

1. Introduction

In recent decades, a surge of new findings across the social and mind sciences of social and cognitive psychology, neuroscience, organizational behavior, and behavioral economics has threatened to sweep away many of our most cherished beliefs about ourselves.[1] We now know more about our minds than we ever thought there was to know—perhaps more than we are comfortable knowing. We have learned that we are prey to heuristics and biases that shape our reasoning in ways we cannot detect and would not endorse (Tversky and Kahneman 1974, Greenwald et al. 2009). We have learned that tiny, irrelevant features of our situation—finding a dime (Levin and Isen 1972), holding a warm drink, (Williams and Bargh 2008), seeing a briefcase rather than a backpack (Kay et al. 2004)—can have marked effects on our subsequent judgment and behavior.[2] And we have learned that our moral intuitions are often shaped by emotions (like disgust) and rationalized afterward (Haidt 2001, Schnall et al 2008). In short, we are not so in control of our thinking and behavior as we might have thought we were, and our moral selves are not so pure as we would like them to be.

During this same period, our social landscape has also undergone major shifts; however, striking social inequalities remain. Black and Latino elementary and high school students still trail

[1] For a clear and concise statement of this threat, see Nahmias (2009).

[2] These are all examples of situational *priming*, in which some feature of a person's immediate environment (the "prime") activates a concept, mood, or attitude that subsequently makes related concepts, etc. more accessible for use in judgment and behavior. Levin and Isen (1972), (1975) found that "feeling good" after finding a dime in a telephone booth made people more likely to engage in helping behavior. Williams and Bargh (2008) found that among study participants who were asked to judge the character traits of a (fictional) personal, those who had previously held a cup of hot rather than iced coffee rated the person more highly as having a "warm" personality; in a second study, participants who held the hot coffee were more likely than those who held the iced coffee to choose a reward for a friend rather than keep it for themselves. And Kay et al. (2004) found that being in a room with items associated with business, e.g. a briefcase, led participants to perceive more competitiveness in an ambiguous social interaction and to propose higher sums of money for themselves rather than their partners in an Ultimatum Game.

their White classmates by two grade levels (NCES 2009), and are more than twice as likely to drop out of high school (NCES 2011). The wage gap between women and men has also persisted, with White women earning 82 cents to the dollar as compared with White men; Black and Latina women earn only 70 and 61 cents, respectively (AAUP 2012). And the highest levels of inequality ever reported put the median wealth of White households at 20 times greater than Black households and 18 times greater than Hispanic households (Taylor et al 2011). At the same time, public opinion remains sharply divided along these lines: Black and White Americans disagree about whether and how much more should be done to achieve racial equality, and women and men disagree about why there are so few women in positions of leadership (Desilver 2015, Pew Research Center 2015). Political polarization by party has steadily increased to the highest levels on record (Pew Research Center 2014), and policies around women's reproductive rights, race-conscious admissions programs, and collective bargaining rights have come increasingly under fire.

I write about these two circumstances in conjunction because the former has offered significant new insights into the latter.[3] As U.S. Supreme Court Justice Ruth Bader Ginsburg emphasized in her recent dissent from the Court's decision to strike down a major provision of the Voting Rights Act, overt efforts to block African-Americans from voting have given way to "more subtle second-generation barriers" that continue to prevent them from full participation as equal citizens (Shelby County, 2013). Justice Ginsburg was referring to practices such as gerrymandering and vote dilution, in contrast to earlier practices like poll taxes, literacy tests or the outright denial of legal rights to vote. However, we now know that second-generation barriers can be explicit or implicit: erected through conscious undertakings, but also built into our non-conscious, automatic cognitive processes. In light of research from the social sciences, we can add various forms of *unintentional* discrimination involving implicit biases, microaggressions, and lagging informal practices (e.g. "old boys' networks") to the list of second-generation barriers to equal opportunity faced by groups historically disadvantaged by race, gender, class, sexuality, disability, etc.

As moral and political philosophers, we do well to pay attention to this social scientific

[3] I do not make any claims about just how many of these second-generation barriers how many are implicit as opposed to explicit but merely undercover (e.g. gerrymandering and vote dilution). However, I take it that implicit bias can explain at least some cases that the latter cannot. I am grateful to Elizabeth Anderson for discussion of this point.

literature, especially when these findings raise challenges for traditional philosophical conceptions of our capacities for reasoning and acting. At the same time, we would also do well to remain mindful of the social, political, and historical context in which our moral and political theorizing is situated. Philosophical work on critical moral responses to moral wrongs, with a few notable exceptions[4], has overwhelmingly focused on blame, resentment, and punishment. In our current climate, however, such responses are not always likely to fare well. When discrimination is unintentional, it is at best ineffective and at worst morally unjustified to blame, resent, and punish. Moreover, in climates of intergroup distrust and hostility, attempts to denounce more subtle forms of discrimination run the risk of provoking threat, denial, and resistance, with both sides pointing fingers and talking past one another.[5] Yet cooperation is required to enact the large-scale, transformative change that is needed to fully address wide-ranging structural injustice. That is the project to which this dissertation contributes.

My focus is on moral responsibility for implicit bias: by "moral responsibility" I mean the features of our actions that make certain responses appropriate (especially when those actions involve wrongdoing), and by "implicit bias" I mean automatically-activated associations between social categories and negative or stereotypical traits that influence people's judgment and behavior (often without their awareness or control). In what follows, I will use the terms "critical moral response" and "moral criticism" more or less interchangeably[6]. Following Elise Springer (2013), I intend the term "critical" to refer to the ways in which we as moral agents "notice and address what others are doing with their moral agency" (2). However, I also mean to evoke the sense of "critical" that is at play in various traditions of critical theory, e.g. critical legal studies, critical race theory, and critical philosophy of race, where the aim of such theory in the broadest sense is to dismantle existing structures of oppression. My overarching aim, then, to find and develop ways of responding to newly-discovered implicit cognitive processes such as implicit bias, which represent second-generation *psychological* barriers to social equality. Faced with the far

[4] Here I am referring to such work as Elise Springer's (2013) *Communicating Moral Concern*, Iris Marion Young's (2011) *Responsibility for Justice*, Margaret Urban Walker's (2006) *Moral Repair,* and Linda Radzik's (2011) *Making Amends*.

[5] It has been well documented by social psychologists that one and the same event can be interpreted by different individuals in different ways. Incompatible "subjective construals" of this sort can provoke prolonged and enduring interpersonal conflict when both sides are similarly afflicted by "naïve realism," a set of assumptions on which people believe themselves to be unbiased perceivers and that disagreement from others can only be the result of bad information, incompetent reasoning, or bias. See for example Griffin and Ross (1991) and Ross and Ward (1996).

[6] While the latter reads more smoothly, it may be more likely to prompt the sense that *agent* is being morally assessed; so where I believe there to be risk of this interpretation, I will use the former instead.

richer and more complicated model of human action that we have gained from the social sciences, I contend, we are in need of correspondingly more nuanced modes of critical moral response that go beyond our usual understanding of blame, resentment, and punishment as the standard practices of moral responsibility. My own account seeks to be sensitive to both normative and pragmatic constraints: an acceptable theory of moral responsibility must not only do justice to our moral experience and agency, but must also issue directives that are psychologically effective in bringing about positive changes in judgment and behavior.

To give away the punchline: I will argue that the problem of implicit bias forces us to take seriously the distinction between two different concepts of moral responsibility, which I call "responsibility as attributability" and "responsibility as accountability" (Chapter 2). We are morally responsible for our actions in first sense only when they reflect our identities as moral agents—that is, when those actions are *attributable* to us as manifestations of our character, attitudes, ends, commitments, or values. On the other hand, we are responsible in the second sense when it is appropriate for others to enforce certain demands and expectations on those actions—in other words, to hold us *accountable* for them. Distinguishing between the two philosophical origins of these two concepts, I show, puts us in a position to distinguish between two different families of moral criticism that they underwrite, which I call "appraisal-based" and "non-appraising" responses. Blame, resentment, and punishment are examples of appraisal-based responses. However, non-appraising responses—though common in ordinary moral life—have not hitherto been properly theorized. As I characterize them, non-appraising responses have important temporal and contextual dimensions, pertaining to individual's patterns of actions over time (Chapter 3) as well as their roles in wider social and institutional contexts (Chapter 4). These dimensions, while not wholly absent, are less salient for appraisal-based responses, which focus primarily on evaluations of the agent on the basis of her present actions. I argue that my new category of non-appraising responses gives us new tools for addressing the problem of implicit bias, as well as the structural injustice that gives rise to them.

In this introduction, I will give an overview of relevant background research, both empirical and moral, on implicit biases and on moral responsibility. Since both of these literatures are immense, I will focus on providing a "morally annotated" review of empirical research on implicit bias that highlights what I take to be its most morally relevant features. Similarly, I will provide only brief summaries of recent work on moral responsibility for implicit

bias, since disagreement in first-order judgments about moral responsibility for implicit bias, on my view, is traceable to more fundamental issues involving our concept(s) of moral responsibility. Throughout, I will flag the relevance of this work for my upcoming chapters and the overarching aims of my project, which I describe in the final section.

2. A Morally Annotated Review of Research on Implicit Bias

In what follows I describe some of the history and methods used to measure implicit bias, evidence for its ability to predict behavior, its prevalence across different social groups, research on learning, unlearning, and intervention, and interpretations of implicit bias as a (extra)personal phenomenon.

2.1 Historical Themes

Work on implicit bias can be located within the "cognitive turn" in stereotyping and prejudice research. Early research on stereotyping and prejudice from the 1930s to 1950s proposed mainly motivational explanations such as differences in individual personality (especially the "authoritarian personality"), which shifted in the 1960s to 1970s to sociocultural explanations such as intergroup competition (Monteith, Woodcock, & Lybarger 2013). In the 1980s, however, *social cognitive* explanations rose to dominance, in which stereotyping and prejudice "were no longer seen necessarily as the result of affective and motivational distortions or based on a history of intergroup conflict, but rather as resulting from normal information processing" (Monteith, Woodcock, & Laybarger 2013).

This shift was itself part of a more general cognitive turn within social psychology, driven by insights gained from the "cognitive revolution." Also in the 1970s and 1980s, two different strands of cognitive psychological work on memory converged under the heading of *implicit social cognition* (Payne and Gawronski 2010). The first derived from work on attention and short-term memory, focusing on a distinction between automatic and controlled processes of memory retrieval. What matters here, to put it in terms most relevant for moral philosophers, is *control*. While controlled processes are "voluntarily initiated and altered" in ways that require the use of our finite cognitive resources, automatic processes are "difficult to suppress voluntarily" in virtue of having been learned so well that they require virtually no cognitive resources. This is reflected, for example, in the difference between deliberately searching for some poorly learned memory,

on the one hand, and having some memory surface effortlessly or unbidden, on the other (Payne and Gawronski 2010, p. 2). As Fazio et al (1986) put it, "the key feature of such automatic activation, then, is inescapability"—in other words, lack of control (p. 229). The second strand of thought derived from work on a separate distinction between explicit and implicit memory; "implicit memory" refers to the way which traces of past experience remain influential on cognitive processing even when no conscious memory remains. Priming effects, for instance, are obtained when implicit memories of the experience of some prime—the backpack or briefcase, for instance—subsequently influence people's judgments and behavior without their consciously attributing those decisions to their prior experience of that prime. Here, by contrast, what matters is *awareness* or lack thereof. (I will return to the relevance of control and awareness for judgments of moral responsibility in Sections 3.1.1-3.1.2.) In crude terms, then, we could characterize implicit social cognition as information processing about social objects—oneself, others, and social groups—that occurs (in some sense) without an agent's control or awareness. Implicit bias is a species of such implicit social cognition.

2.2 Methods of Measurement

"Implicit bias" itself is a term of art (Brownstein 2015). While much of the psychological literature has focused on "implicit attitudes" or "implicit associations," I use the term "implicit bias" both because it is the term that has been more popularly adopted outside psychology—including in philosophy—and also in order to refrain from taking a stance on more detailed issues of ontological and theoretical interpretation (are implicit biases really attitudes? are they best understood as aliefs, beliefs, or "patchy endorsements"? do they have associative or propositional structure? are they purely cognitive or affect-laden?) which I largely leave aside.[7] My preferred strategy is to lean heavily on the set of representative studies that I present here, thereby emphasizing methods over interpretation. (I describe these methods in rather great detail, in the hopes of providing the reader a clear sense of what exactly is being measured.) Although the exact nature of these explicitly unendorsed associations remains under investigation, their existence is strongly supported by convergent evidence deriving from their measurement by a multiplicity of distinct though related implicit methods, indirect detection

[7] See, respectively, Arkes and Tetlock (2004) and Banaji, Greenwald, & Nosek (2004), Gendler (2011), De Houwer (2014), Levy (2015), Gawronski and Bodenhausen (2011) and Mandelbaum (2015), Amodio and Ratner (2011). For a review, see Brownstein (2015).

through field studies, and non-scientific, ordinary observations of social experience.

2.2.1 The Implicit Association Test

By far the most well-known and commonly-used method of measuring implicit bias is the Implicit Association Test (IAT), pioneered by Greenwald, McGhee, & Schwartz (1998) and further developed by Greenwald, Nosek, & Banaji (2003). The IAT compares subjects' accuracy and reaction times in performing a sorting task across different sets of trials. In one set of trials, participants are asked to sort together (as quickly and accurately as they can) members of some social category together with positive words, and members of some other social category together with negative words; in another set of trials, the two social categories are switched. For example, in the sample IAT at left, the top-most trials (a) and (b) request participants to categorize together pictures of Black patients and negative words: the correct response for the Black patient in (a) would be to sort it to the left, while the correct response for the positive word in (b) would be to sort it to the right. In the bottom-most trials (c) and (d), however, the categories have been switched. Participants are now requested to categorize together pictures of White patients and bad words. The correct response for (c) would be to sort the picture of the Black patient to the right, and to sort the positive word in (d) to the right as well. If a person is faster and more accurate at categorizing together pictures of Black patients with negative words than with positive words than she is at performing the same task with pictures of White patients, that suggests that she has a more negative association with Black people than with White people (Green et al. 2007).

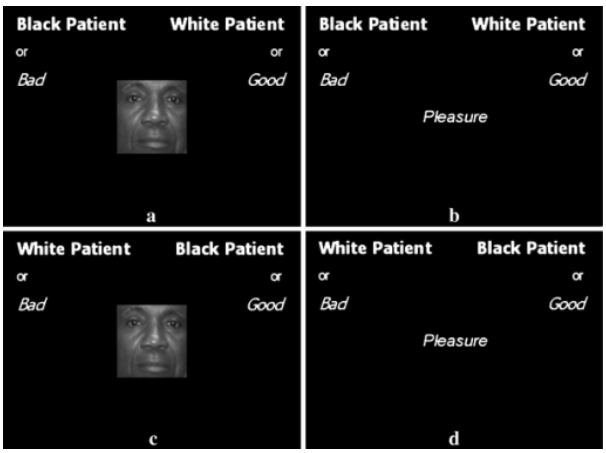

Figure 1: IAT, from Green et al. (2007), "Implicit Bias Among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients"

2.2.2 Sequential Priming

A somewhat earlier measure of implicit bias, used by Fazio et al. (1995), is the sequential priming task, which also involves categorization and response times. It compares the time needed to categorize a target word or picture after being exposed for a very short time—short enough to

require automatic rather than controlled processing—to a prime. For example, Figure 2 illustrates a (a) sequential priming task and (b) weapons identification task where the primes are pictures of Black and White faces and the targets are either pleasant/unpleasant words or pictures of guns/tools. If a person is faster at categorizing a word as unpleasant or at identifying a gun after being primed with a picture of Black person than she is after a picture of a White person, that suggests she has a negative association with Black people, or that she associates them with weapons (threat).

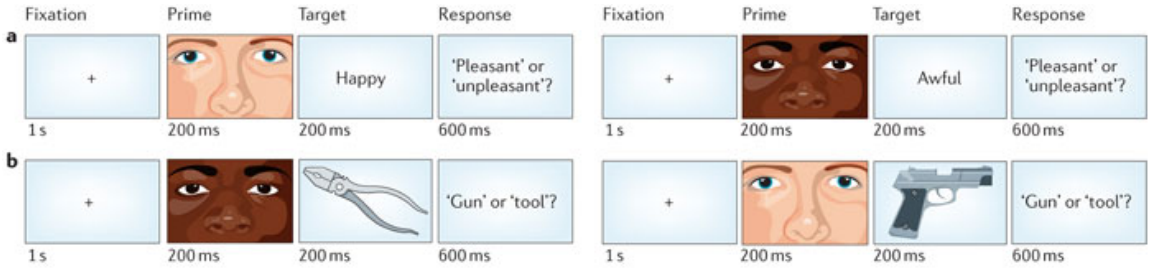

Figure 2: Sequential Priming, from Amodio (2014), "The Neuroscience of Prejudice and Stereotyping"

2.2.3 The Go/No Go Association Test

The Go/No Go Association Test (GNAT) was developed by Nosek and Banaji (2001) to improve on one of the drawbacks of the IAT. The IAT can only be used to detect relative, but not absolute valences, e.g. that a White person has a *stronger* negative association with Black faces than White faces, but not whether this is due to negative out-group bias against Black faces or positive in-group favoritism toward White faces. The GNAT, by contrast, provides a single target category whose positive or negative valence is assessed independently, using the same response action (e.g. pressing the same key). It does so by asking participants to respond when they see positive words and instances of a target category—those that "belong"—and to do nothing when they see distracter words (e.g. negative words and words from another category) which do not "belong"; this is then repeated in another set of trials where the same target category is paired with negative rather than positive words. If a person is more accurate at picking out the target category together with negative words than she is with positive words, this suggests she has a negative association with the target category.

2.2.4 The Affect Misattribution Procedure

Payne et al. (2005) developed an Affect Misattribution Procedure (AMP) for measuring

implicit bias. In the AMP, like the sequential priming task, participants are presented with a prime, e.g. Black or White faces, that appears quickly enough to demand automatic processing; they are subsequently asked to rate the "visual pleasantness" of, e.g. a Chinese character (that was presumably meaningless to non-Chinese participants). If a person is more likely to judge a meaningless Chinese character as pleasant after viewing a White face and unpleasant after a Black face—in other words, if they *misattribute* their affective response to the face as an affective response to the Chinese character—then this suggests that they have a positive association with White people and a negative association with Black people.

2.2.5 Field Studies

Finally, one last category of research that deserves mention is field research performed in real-world contexts. In one of the best-known examples of an "audit study," Bertrand and Mullainathan (2004) performed a field experiment in which they headed identical CVs with two different names: a stereotypically "White-sounding" name and a stereotypically "African-American-sounding" name. When they mailed CVs to actual job listings, they found that their imaginary White applicants needed only to send off 10 applications before getting a callback while Black applicants needed to send off 15; in effect, a White name conferred the same benefit as an additional 8 years' of work experience (Bertrand and Mullainathan, p. 992). A recent replication by Gaddis (2015) found similarly that White college graduates from state schools had the same chances of callback as did Black college graduates from Ivy League universities. Other audit studies—which send out paired testers matched on all relevant characteristics but differing in a target feature like race—have found significant racial differences in people's prospects for renting houses, buying cars, and being reported for shoplifting, as well as more mundane activities such as selling items on eBay and receiving help from faculty.[8]

Another type of field study takes advantage of natural experiments that occur outside the laboratory. For instance, Goldin and Rouse (2000) found that symphony orchestras who adopted a policy of holding auditions behind a screen subsequently increased their selection of female musicians by 50%. To take another example, Anwar, Bayer, & Hjalmarsson (2015) found that all-white juries were 16% more likely to convict Black defendants than White defendants—an effect that disappeared when juries included at least one Black juror.

[8] These studies are reviewed in Blank, Dabady, & Citro (2004) and in Mullainathan (2015).

While such studies do not measure implicit bias *per se*, one likely explanation for these findings—given the plausible assumption that people are not consciously discriminating on the basis of gender and race—is that people's implicit associations involving racial and gender categories affect their subsequent judgments about others' competence, talent, trustworthiness, and guilt.

2.3 Implicit Bias as Predictive of Behavior

A 2009 meta-analysis undertaken by Greenwald, Poehlman, Uhlmann, & Banaji covered 122 studies that included measures of the correlation between IAT scores, direct self-report, and behavioral measures, where behavioral measures included a variety of "behaviors" such as physiological and nonverbal responses, evaluative judgments, preferential choices, and physical actions. The authors found a moderate correlation between IAT scores and behavioral measures which was on the whole smaller than the correlation between direct self-report and behavioral measures. However, they found that both IAT scores and self-report demonstrated incremental validity relative to one another, i.e. that they provided new predictive information that was not redundant. More significantly, they found that the predictive power of self-report decreased significantly with respect to socially sensitive topics such as interracial relations; in these cases, IAT scores were able to predict behavior better than direct self-report.

A number of the behavioral measures tested were morally important. Jost et al. (2009), for instance, collect together what they claim are "ten studies that no manager should ignore." Among them is the Green et al. (2007) study from which Figure 1 is taken. Green et al. provided actual physicians with a set of "clinical vignettes" featuring hypothetical patients; they found that physicians who demonstrated higher levels of anti-Black bias on the IAT were more likely to prescribe thrombolysis (a treatment for dissolving blood clots) to hypothetical White but not Black patients. Another is a follow-up to Bertrand and Mullainathan's CV study, in which Bertrand, Chugh, & Mullainathan (2005) found that participants with stronger implicit associations between "Black" and "unintelligent" were less likely to select résumés with African-American names. Another well-known example of the behavioral implications of implicit bias is the First-Person Shooter Task developed by Correll et al. (2002), pictured in Figures 3 and 4. In this task, participants are presented with pictures of Black or White men holding either guns or visually similar objects and asked to decide as quickly and accurately as possible to shoot if it is a

gun and to not shoot if it is not. In Figure 3, the correct response would be *not* to shoot, since the man is holding a cell phone; in Figure 4 it would be to shoot, since the man is holding a gun.

Figure 3: Example of "Don't Shoot," from Corell et al. (2002), "The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals"

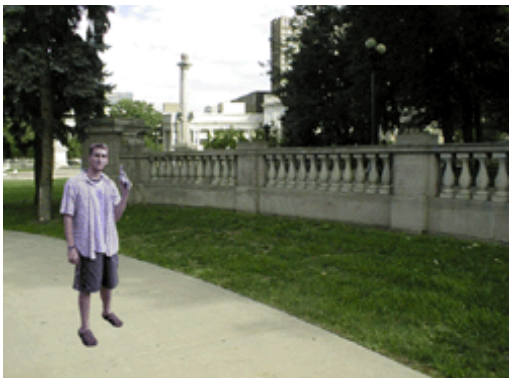

Figure 4: Example of "Shoot," from Correll et al. (2002), ibid.

Correll & colleagues' study of the First-Person Shooter Task among actual police officers found that, despite the lack of racial discrepancies in overall error rates (which were lower than those of civilian participants), officers were still faster at shooting at pictures of Black men than White men.

Measures of implicit bias have also been demonstrated to operate outside the laboratory in naturalistic contexts. For example, Rooth (2010) not only replicated Bertrand and Mullainathan's (2004) CV study, finding that native Swedish applicants were three times as likely as Arab applicants to receive callbacks, but also followed up by administering IATs to a subset (26%) of the actual managers who made those very decisions. He found that IAT scores significantly predicted the manager's likelihood of inviting an Arab applicant to interview, while their direct self-reports did not. In Italy, a study by Arcuri (2008) found that among self-identified "undecided" voters, IAT scores involving political candidates before the elections were highly predictive of actual voting behavior. The detection of implicit bias in real-life contexts is likely to be difficult, however. Dovidio and Gaertner (2000), Hodson, Dovidio, & Gaertner (2002), and Son Hing et al. (2008), for instance, found that in evaluation situations where candidates were *clearly* qualified or unqualified, participants displayed no patterns of racial bias; however, in ambiguous situations—for example, when a candidate was strong along one dimension but weak along another, racially biased patterns reemerged. In sum, a large literature suggests that

implicit biases do indeed exert influence on morally weighty judgments and behavior, in ways that people are not always consciously able to report or control—and, if they have egalitarian commitments to social equality, in ways that they would not reflectively endorse.

2.4 Prevalence

If the evidence about the influence of implicit biases is morally dismaying, the evidence about its widespread prevalence is more dismaying still. The IAT and other implicit measures do not measure only basic positively or negatively valenced associations; it also (as I have mentioned) measures associations between social categories and specific stereotypic traits. For example, the "Gender-Career" IAT measures the strength of associations between being a woman or man and having a career or being a caretaker. Moreover, implicit biases have been found across a wide range of social categories and traits. For example, Harvard's Project Implicit, an ongoing data collection project, currently hosts 14 different IATs. Their results indicate that participants exhibit implicit biases on the basis of sexual orientation, age, disability, skin tone, and religion; moreover, they associate Native and Asian-Americans with "foreign" while they associate White Americans with "American", they associate women with "family" and "liberal arts" while they associate men with "career" and "science," and they associate weapons with Black faces more than White faces ("Project Implicit"). Banaji and Greenwald (2013) estimate that approximately 75% of Americans harbor anti-Black implicit bias (p. 47).

Perhaps even more dishearteningly, members of stigmatized and stereotyped social groups *themselves* exhibit implicit biases against their own in-groups, although their negative associations may be weaker; Nosek, Banaji, & Greenwald (2002) found that even though Black participants on the IAT also demonstrated negative associations with Black faces, the associations were weaker than in White participants. Similarly, 70% of men and 80% of women associate women with the concept of "family" and men with the concept of "career" (Banaji and Greenwald 2013). This internalization of negative or stereotypic associations can have deleterious effects. Smyth, Greenwald, & Nosek (ms), for instance, found that the more strongly college women associated "science" with "male," the less likely they were to major in science, whereas the opposite was true for men. What these results reveal—a point I shall emphasize in Chapter 4—is that implicit bias is not a matter of "bad people with bad attitudes," but a reflection of current social realities. It must be understood as a *structural* rather than an individualistic

phenomenon, and solutions to the problem of implicit bias will be thus incomplete unless they contain a structural dimension. Although my analysis takes place at the level of individual psychology, implicit bias is *not* a problem of individual psychologies. Rather, all of us—of any race, any gender, and on all axes of privilege and disadvantage—are inextricably implicated in a system of unjust social structures. Keeping this structural dimension firmly in mind also lessens though it may not wholly eradicate the temptation to use individual's implicit biases as the basis for moral assessment of their quality as moral agents—a temptation against which I shall argue in detail in Chapter 2.

2.5 Learning, Unlearning, and Intervention

Moral philosophers have typically distinguished between three different targets of moral evaluation with regard to implicit bias: [1] merely having the bias, [2] manifesting the bias in action, and [3] responding (or failing to respond) to the bias. For the purposes of this dissertation, I will set aside [1] in favor of [2] and [3]. My focus in Chapter 2 will be on [2], specifically, on actions caused by implicit bias, while I shift to a consideration of [3] what we can be expected to do to weaken and eliminate our biases, both in ourselves and through transforming the structural conditions that produce them, in Chapters 3 and 4 respectively. I will refer to such de-biasing efforts to weaken and eliminate implicit biases in ourselves as "bias mitigation". In order to be responsible for mitigating implicit bias, however, it must be possible for us to carry out actions toward that end. In the next sections I briefly summarize empirical research that suggests there are indeed actions we can be expected to take in response to our implicit biases.

2.5.1 Learning Implicit Biases

From a moral perspective, it is implausible to think that we could be responsible for *acquiring* many of our implicit biases. Children develop in-group preferences for looking at and interacting with more familiar-looking faces in the first year of life, and express explicit in-group preferences at the age of 3 (Baron 2015, Hailey and Olson 2013). This is long before it is plausible to think that they have developed all the capacities required for full moral agency.[9]

[9] I assume here that moral agency is a graded, rather than binary property. In other words, I am not claiming that children are not moral agents at all, but only that they are not *full* moral agents in the sense required for standard practices of moral responsibility. Even at early ages, children can still be appropriate subjects of moral education, quasi-appraisal, and other practices that serve to "bootstrap" them into full moral agency.

However, these in-group preferences are not identical across groups: children from low-status groups (e.g. Hispanic and Black children) show weaker in-group preferences than those from high-status groups, and may not show any out-group bias against groups with higher status than their own (e.g. White children) (Baron 2015). The exact mechanisms by which implicit bias develop are not yet well understood, in part because questions concerning the interpretation of "implicit bias" have not yet been settled. Amodio and Ratner (2011), suggest that implicit biases may be divided into two types which are grounded in two different systems of memory. They propose that "implicit stereotypes" are primarily cognitive and based in semantic memory, and hence are learned slowly over time through repeated associative pairings of semantic concepts. By contrast, they claim, "implicit evaluations" are primarily affective and based in fear conditioning, which may established very quickly—even after a single instance—and renew quickly after having been unlearned. It has also been suggested that early childhood, a period of rapid neurological and cognitive growth, is a particularly sensitive period during which learned implicit associations are likely to remain especially robust in later life (Rudman, Phelan, & Heppen 2007). And a series of studies by Castelli et al. (2007), Castelli, De Dea, & Nesdale (2009), and Castelli, Zogmaister, & Tomelleri (2009) has found that children's racial attitudes and choices of playmates are significantly correlated with their mothers' (but not fathers') implicit racial biases, nonverbal behaviors, and children's *beliefs* about their mothers' attitudes—rather than the attitudes explicitly reported by the mothers themselves. In any case, by the age of 6, children display implicit racial bias on the IAT at levels comparable to those of adult participants, in addition to self-reported biases. However, by the age of 10, children adopt moral and social norms such that they no longer explicitly report racial biases, even though their levels of implicit racial bias remain constant (Baron and Banaji 2006).

What all the preceding evidence demonstrates is that childhood represents a crucial period of development during which children are highly sensitive to cultural and parental cues about social group differences and hierarchies, and to their own place within those hierarchies. And while it may be tempting to direct moral criticism toward parents whose nonverbal behaviors and implicit biases encourage the development of implicit bias in their children, it is overwhelmingly likely that those parents *themselves* acquired such biases in their early childhood as well. This, I claim, demonstrates again the need to pay attention to background social structures

whose persistence facilitates the intergenerational transmission of implicit biases, quite apart from and in ways that supersede the moral agency of particular individuals.

2.5.2 Unlearning Implicit Biases

Ideally, of course, the best way to respond to an implicit bias would be to *unlearn* it, i.e., to completely dissolve or retrain the association between that social category and the negative or stereotypic trait. Researchers have obtained a number of positive results from a various "de-biasing" strategies that attempt directly to change underlying associations (Lai, Hoffman, & Nosek 2013; Brownstein 2015). A series of studies by Kawakami et al. (2007), Kawakami, et al. (2008), and Phills (2011) found that training participants to pull a lever toward themselves when they saw Black faces and pushing it away when they saw White faces reduced levels of implicit racial bias; in effect, such exercises serve to associate Black people with oneself, toward which one has (presumably) positive attitudes, or perhaps to "assimilate" Black people into one's in-group. Similarly, Kawakami et al (2000) asked participants to practice affirming "Yes!" to counterstereotypical pairings of Black faces and positive words (e.g. "intelligent"), which also resulted in reduced levels of implicit bias. Notably, however, saying "No!" to stereotypical pairings does *not* appear to reduce implicit bias, serving instead to strengthen the association (Gawronski et al. 2008). In fact, merely being exposed to counterstereotypical pairings, e.g. Black faces and positive words, had the effect of reducing implicit racial bias, where that effect remained two days later in a follow-up exercise (Olson and Fazio 2006). Blair, Ma, & Lenton (2001) also found that participants asked to imagine counterstereotypical individuals (e.g. "what a strong woman is like, why she is considered strong, what she is capable of doing, and what kinds of hobbies and activities she enjoys") later showed weaker associations between women and stereotypically feminine traits. Dasgupta and Greenwald (2001) reported similar effects after participants were exposed to pictures of well-known admired Black and disliked White figures (e.g. actor Denzel Washington and serial killer Jeffrey Dahmer), which persisted in a follow-up exercise one day later. This suggests that greater familiarity with counterstereotypical individuals, e.g. through media portrayals but also through interaction with actual individuals, may be effective and morally valuable ways of mitigating implicit bias.

Lastly, and most importantly, implicit biases may be unlearned through extended intergroup contact. Indeed, intergroup contact has been the traditional focus of research on

stereotyping and prejudice since Gordon Allport (1954) proposed his "intergroup contact hypothesis" that prejudice is reduced when members of different social groups interact under conditions of equal status, in mutual cooperation in pursuit of shared goals, and with social and institutional support. A recent meta-analysis of 515 studies by Pettigrew and Tropp (2006), for instance, found in 94% of cases that higher intergroup contact was correlated with lower levels of prejudice. Pettigrew and Tropp's (2008) follow-up article on the underlying mechanisms of contact effects showed that affective mediators—decreased outgroup anxiety and increased empathy—were significantly stronger than the cognitive mediator of increased knowledge about outgroups. Turner, Hewstone, & Vico (2007), however, have argued that prejudice reduction may operate differently with respect to explicit and implicit attitudes. They found that decreases in explicit bias were correlated with explicit, high-quality interpersonal interactions; in particular, with the mutual self-disclosure of personal information that typically characterizes relationships between friends. But since contact also generates mere exposure effects, Turner and colleagues proposed that mere exposure might be sufficient for decreasing IAT scores, which is precisely what they found, even in contexts of intergroup tension. Similarly, Tam et al. (2006) found that quantity but not quality of time spent with older people was correlated with lower bias against the elderly on the IAT, while quality was correlated with more favorable explicit attitudes. This was also demonstrated in a natural field experiment performed by Shook and Fazio (2006), who looked at the relative levels of implicit bias among White first-year college students who had randomly been assigned either a White or African-American roommate. After their first term, students who with Black roommates demonstrated lower levels of implicit racial bias compared to the beginning of the term—an effect that did not occur for their counterparts assigned to White roommates, and in spite of the fact that they reported lower satisfaction and less involvement than their counterparts. These results, I believe, reiterate the need to transform background social structures, which in the contemporary U.S. consists of highly racially segregated neighborhoods, school systems, informal networks, and informal circles which do not facilitate intergroup contact (Anderson 2010). However, they also suggest that gender, racial, and other forms of underrepresentation in various domains are not only an *effect* of implicit bias, but also a contributor to its persistence; this again points to the moral importance of efforts to develop intergroup relationships and to address underrepresentation within educational, residential, organizational and other associations.

2.5.3 Interventions

It is also possible to block or weaken the influence of implicit bias on action even without completely eliminating the bias. For example, Bertrand, Chugh, & Mullainathan (2005) found that the effect of implicit bias was stronger when participants felt more rushed, suggesting that allowing more time—and less reliance on quick automatic judgments—may be effective for reducing the impact of implicit bias on important decisions. Other studies suggest that even if short-term exposure to counterstereotypical exemplars does not wholly dissolve the associations that constitute the bias, it may still function to weaken its operations in that short-lived setting (Lai, Hoffman, & Nosek 2013). One reason for this may be that counterstereotypical exemplars activate a distinct subset of the social category in question; for instance, the traits associated with Black academics may not be the same as those associated with Black men or Black women as a whole. Other contextual effects are generated by the presence of others: Lowery, Hardin & Sinclair (2001) found that participants who interacted with a Black experimenter subsequently showed lowered levels of implicit racial bias compared to those who interacted with a White experiment, and Castelli and Tomerelli (2008) found that the mere presence of other people (which presumably triggered the desire to uphold socially accepted norms) also reduced levels of implicit racial bias. More ominously, Richeson and Ambady (2001) found that men who anticipated an interaction with a high-status woman scored higher on a gender stereotype IAT compared to men who anticipated interactions with subordinate or equal-status women. In a later study, Richeson and Ambady (2003) found a similar pattern of results wherein White women who anticipated an interracial reaction with a Black person in a relatively superior social position showed more implicit racial bias than those who anticipated interacting with a Black person in a relatively subordinate position; this effect did not occur when participants were expecting a same-race interaction. Again, these results underscore the subtlety of some of the mechanisms that underwrite ongoing social inequalities: the implicit biases described in the two studies above likely contribute to the persistent underrepresentation of women, racial minorities, and other historically disadvantaged groups in high-status positions, but their selective manifestation only in highly specific, status quo threatening situations makes them difficult to detect and prevent. The importance of rectifying gender and other underrepresentation is supported by another natural experiment by Dasgupta and Argari (2004), who found that even

one year later, female undergraduate students with more female instructors had weaker implicit gender biases associating men with leadership than students with fewer female instructors. These results also point to the importance of being attentive to situational cues in local environments: the demographic make-up of any particular group, the presence of stereotypical or counterstereotypical exemplars, etc.

Besides situational and contextual interventions, there is also a growing literature on other ways to suppress the influence of implicit bias. Mendoza, Gollwitzer, & Amodio (2010), for instance, showed that teaching participants to practice "implementation intentions," that is, "if-then" plans tied to specific cues led to less racially biased task performance (e.g. thinking "If I see a Black face, I will think 'Safe!' on the First-Person Shooter Task). Moskowitz et al. (1999) found that the task performance of participants with "chronic egalitarian goals"[10] was less influenced by gender stereotypes than participants who did not have such goals. Moskowitz et al. argue that this "preconscious inhibition of stereotype content" occurs when the activation of implicit goals suppresses the activation of implicit bias. Monteith, Mark, & Ashburn-Nardo (2010) have also developed a model of self-regulation of prejudice on which the awareness of biased responses can lead to negative affect, reflection, and subsequently the establishment of "control cues" that spark reflection and behavioral inhibition when those control cues suggest the possible operation of bias. Monteith, Voils, & Ashburn-Nardo (2001) demonstrated that some participants show awareness of slower, racially biased responses on the IAT in ways that may initiate this process of self-regulation; however, this type of awareness may not be easy to achieve (Monteith, Mark, & Ashburn-Nardo 2010, p. 195).

However, Mann and Kawakami (2012) found that participants who were told that they were making progress toward egalitarian goals—compared to participants who received the opposite or no feedback at all—subsequently demonstrated higher levels of implicit racial prejudice in their IAT scores and in their choice of seating distance next to a Black partner, suggesting that the effect of implicit egalitarian goals are actually undermined when the goal is thought to have been achieved. Similarly, Monin and Miller (2001) found that participants who had opportunities to behave in egalitarian ways (e.g. opposing a sexist claim, choosing a racial minority applicant for non-racially stereotyped job) later felt licensed by their "moral credentials" to behave in more biased ways (e.g. choosing a male applicant for a stereotypically male job,

[10] These were participants who, when faced with a multiple choice question that included only gender-stereotypical answers, later compensated by giving especially counterstereotypical responses to a subsequent task.

rejecting a racial minority applicant for a job stereotypically filled by Whites) than a control group. These last results illustrate the complexities involved in the ways that implicit bias actually gets manifested in action, and the ways in which even "control-based" techniques of intervention may not function and are not phenomenologically experienced in the same way that ordinary voluntary actions do; while they do facilitate the achievements of our goals, they might do so only indirectly or at the non-conscious level. While we can voluntarily undertake such actions as engaging in de-biasing techniques or developing implementation intentions, we still do not exert conscious control over the course of these actions and we may only know of their success or failures indirectly through third-person observation.

Moreover, because most of the research in sections 2.5.2-2.5.3 is still in early stages, a number of basic questions remain unanswered: what the exact mechanisms of unlearning and intervention are, how durable their effects are, whether their effects can be achieved outside the laboratory, and whether they can be scaled into institutionally feasible practices (Lai, Hoffman, & Nosek 2013; Brownstein 2015). Even more caution is warranted when we consider the ways in which efforts may backfire. Given limited time and resources, then, I advocate that our moral focus remain on efforts to enact institutional policies and practices that will ultimately undermine the existence of pernicious implicit biases in the first place. And since implicit associations of some sort or other are an inevitable product of our bounded rationalities, the work we put into institutional design—e.g. the use of clearly-defined, specific criteria rather than holistic and impressionistic evaluations, or the use of blind evaluation and review whenever possible—for the purpose of blocking irrelevant implicit associations will not go to waste. Where de-biasing, environmental cues, and other such interventions are low-effort, low-cost, and low-risk to the best of our knowledge, however, it may be useful to implement them on an ongoing basis while the empirical study of such techniques continues to gain in sophistication and precision.[11]

2.6 Do Implicit Biases Belong to the Person?

A particularly noteworthy discussion in the social psychological literature for theorists of moral responsibility concerns the question of whether implicit biases as *personal*, that is, whether they may "be considered as part of a person's true self" (Gawronski, Peters, & Lebel 2008). Researchers have disagreed about whether implicit biases should be considered "personal," that

[11] However, see Haslanger (2015) for the view that too much attention to implicit bias (and hence de-biasing interventions) can obscure deeper structural causes of inequality, and see Madva (ms) for an opposing view.

is, as belonging to or reflective of a person's self and attitudes, or whether they should be considered "extrapersonal," that is, reflective of something outside a person's self, e.g. widespread cultural knowledge or knowledge of others' attitudes that are not incorporated into one's own attitudes. I will refer to this as the question of "ownership," that is, whether an implicit bias *belongs to* a person in the sense relevant for ascriptions of moral responsibility.

While Devine (1989) initially proposed that implicit biases were merely traces of knowledge of socially shared stereotypes (as a way of explaining their existence despite the fact they were not explicitly endorsed by participants), subsequent findings demonstrated that there was individual variation across a population that had all presumably been exposed to more or less the same cultural context (Devine and Plant 1998). This suggested that the relationship between implicit bias and cultural knowledge was more complex, and that even *if* there existed some monolithic set of cultural associations, individuals were differentially absorbing them into their own experience-based associationist networks. There have since been a number of studies investigating possible mechanisms and relationships between implicit bias, cultural knowledge, and knowledge of others' attitudes (Nosek and Hansen 2008, Olson, Fazio, & Han 2009). However, Gawronski, Peters, & Lebel (2008) and Gawronski (2009) have argued that the distinction between "personal" and "extrapersonal" is currently contested, and moreover is not wholly decidable on purely empirical considerations. In particular, Gawronski (2009) points out that two intuitively appealing pictures of what constitutes the "true self" are in direct contradiction with one another. On the first picture, the true self is what is revealed when a person's inhibitions are disabled, e.g. when a person is more willing to disclose her honest but unpopular opinions when she is drunk than sober. On the second picture, a person's true self is reflected in what she endorses or consciously intends, e.g. when a person's stepping on another's foot is wholly unintentional and does not reflect anything about her as an agent. Gawronski concludes that choosing between either is a non-empirical matter of pure subjective preference, while Gawronski, Peters, & Lebel (2008) state expressly: "The question of how to define personal and extra-personal associations involves a strong moral component, in that any answer to this question has the potential to influence judgments of personal responsibility" but that such considerations go "far beyond the scientific discourse on how to define a psychological construct". Indeed, as I will describe in Section 3.1.3, this has been precisely the topic of concern for philosophers working on moral responsibility for implicit bias.

I defend the contours of a view of the true self (the central question of what I will call "responsibility as attributability") which is broadly aligned with Gawronski's second picture in emphasizing the importance of reflective endorsement to moral agency in Chapter 2; however, what I argue for in this dissertation is intended to leave open multiple possibilities for how exactly we should understand the personal-extrapersonal distinction. In the context of this review, I will only note that *metaphor*, the use of which is not confined to the philosophical literature[12], *matters*. Whether we (perhaps implicitly) conceptualize implicit measures as analogous to alcohol-induced lowered inhibitions, or implicit bias as mere accident or bodily function[13], will make a difference to our moral intuitions. (As will be evident in Chapter 2, I have chosen to adopt the metaphors of miscalculation, which is consonant with current psychological understanding of implicit bias as a phenomenon within implicit social *cognition*, and of situationally-triggered trauma, in keeping with the understanding of implicit bias as grounded in memory and past experience.) However, the best choice of metaphor will ultimately depend on further precision in our understanding of the mechanisms underwriting implicit bias, and our theoretical and ontological interpretations of it—an understanding that awaits further empirical inquiry.

3. Moral Responsibility for Implicit Bias

Research on implicit bias has attracted considerable attention from moral and political philosophers who have recognized its potential to explain persistent social inequalities, and who have been alerted to the possibility of a previously unknown array of morally pernicious discriminatory behavior. Of course, those on the receiving end of implicit bias have long known that even "well-intentioned folks" act in biased and harmful ways. However, the establishment of a rich repository of empirical work has provided not only a powerful vindication but also an explanation of such cases. Philosophers working on moral responsibility have also been particularly interested in the phenomenon of implicit bias because it (along with other recent psychological work on moral reasoning[14]) requires the adoption of a new psychological model of moral action and decision-making—one on which conscious reflection is accompanied by many

[12] Banaji and Greenwald (2013), for instance, compare implicit biases to bodily functions like the flow of blood, cell regeneration, and the firing of neurons (61). Banaji, Nosek, & Greenwald (2004), moreover, in order to emphasize the "lack of introspective access and lack of conscious control" that characterizes implicit attitudes, write that "to speak of implicit attitudes as endorsed would be as nonsensical as speaking about a dog endorsing a bone" (280).
[13] See previous footnote.
[14] See, for example, Jonathan Haidt's (2001) "social intuitionist" model of moral reasoning.

forms of automatic, non-conscious processing. Accordingly, most moral responsibility theorists who have directly addressed the problem of implicit bias have been heavily engaged with the empirical literature, drawing out the ways in which current understandings of implicit bias map onto traditional categories such as "control" and "awareness." In the following sections I will summarize the emerging literature on moral responsibility for implicit bias, which seeks to make sense of it in terms of an agent's control, awareness, or ownership of the attitude. I would like to note up front, however, that my own project, is *not* (for the most part) to defend specific answers to these substantive, first-order questions of whether the operations of implicit bias support sufficiently robust kinds of control, awareness, or ownership[15] for moral responsibility. I will return to the aims of my own project in the final section.

3.1 Control

One of the most obvious reasons to think that we lack moral responsibility for implicit biases is that they do not seem to be under our control. Jennifer Saul (2013), in one of the earliest discussions of implicit bias in the philosophical literature, writes:

> A person should not be blamed for an implicit bias that they are completely unaware of, which results solely from the fact that they live in a sexist culture. Even once they become aware that they are likely to have implicit biases, they do not instantly become able to control their biases, and so they should not be blamed for them. (55)

I use these intuitions as one of my starting points in Chapter 2, to illustrate a tension between 1) thinking that lack of control and/or awareness over implicit bias undermines people's moral responsibility, while also 2) thinking that the harmful effects of implicit bias are such that people must be held responsible for them. My strategy will be to disentangle two different concepts of responsibility that underwrite these two different common intuitions.

Other philosophers, however, have proposed alternative accounts of the kind of control that is required for moral responsibility. Jules Holroyd (2012), for instance, argues that we need not have direct voluntary control over implicit biases' influence on our actions in order to be responsible for them. She contends that we need only indirect, long-range control over implicit bias through engaging in actions that can *mitigate* bias, of the sort I described in sections 2.5.2-2.5.3, in order to be. Suhler and Churchland (2009), moreover, have argued that there are forms

[15] I follow Brownstein (2015) in adopting this tripartite categorization and in much of the summaries below, though my own terminology and synthesis of the literature reflects my particular interests and purposes.

of non-conscious control that are sufficient for supporting ascriptions of moral responsibility. On their view, goal-maintenance and executive control, i.e. the ability to form and carry out plans in a way that is responsive to environmental contingencies and stimuli, is the hallmark of moral agency that is required for moral responsibility. Since the literature on automaticity has demonstrated that these capacities for goal-maintenance and executive control are carried out in large part through the operations of implicit (social) cognition, and since implicit bias is just a species of implicit social cognition, we are in principle[16] morally responsible on this view for the actions they cause. Finally, Angela Smith (2007) has defended an account of moral responsibility, which I discuss below in Section 3.3, on which voluntary control is not even a necessary condition.

3.2 Awareness

Just as in the case of control, the importance of awareness for moral responsibility depends on what sense of "awareness" is in play. Gawronski, Hofmann, & Wilbur (2006), have distinguished between three different types of awareness of attitudes: content awareness (e.g. whether the attitude toward some object is positive or negative), source awareness (e.g. whether the origin of the attitude is cultural, innate, etc.), and impact awareness, i.e. how the attitude impacts other psychological processing. They suggest that people may have content awareness of their implicit biases, as suggested by the fact that certain manipulations and measures (e.g. a "bogus pipeline" manipulation in which participants are made to believe that false reports will be detected, or measures that ask for "gut feelings" and other affective rather than cognitive responses) of implicit bias generate higher correspondence with direct self-report. With regard to source awareness, they point out, people often do not have source awareness of their explicitly reportable attitudes either, so the lack of source awareness cannot be a reason to treat implicit biases and explicit attitudes differently when it comes to ascribing responsibility for them. However, Gawronski and colleagues think that implicit biases can be distinguished from explicit attitudes in that people lack impact awareness of the ways in their implicit biases interact with and influence their subsequent judgment and decision-making. For example, a person might have content awareness of an implicit dislike for some social group, and may even know that the source of it lies in early childhood exposure to parental or cultural disdain, but might not know

[16] However, Suhler and Churchland (2009) do not explicitly discuss how their theory applies to the phenomenon of implicit biases.

that such an implicit bias is affecting her judgments and treatment of members of that social group; indeed, she may reject the *influence* of that dislike even if she endorses its content, e.g. if her aim is to be a wholly impartial judge unswayed by personal preference.

This claim by Gawronski and colleagues, however, has been challenged by Jules Holroyd (2014), who cites evidence from situational priming and other areas of social psychology to show that we may also lack impact awareness of our explicit attitudes, e.g. we may not know that our dislike of lawn-mowing sounds is having an effect on our judgments of job applicants' ability. For Holroyd, the relevant type of awareness for moral responsibility is "observational awareness," that is, occurrent awareness of the morally relevant features of actions, e.g. slower responses to pairings of Black faces and pleasant words on the IAT, as was demonstrated by Monteith, Voils, & Ashburn-Nardo (2001). Another type of awareness that may determine culpability, proposed by Natalia Washington and Daniel Kelly (forthcoming) is sociohistorical context: the more that general knowledge of implicit bias spreads, and the more relevant it is to a person's professional role (e.g. whether a person has the authority to assess and evaluate candidates, job performance, etc.), the more blameworthy a person becomes if she lacks impact awareness of her own implicit bias. On the other hand, Neil Levy (2012) has argued that conscious awareness is a necessary condition for moral responsibility because it is in turn a necessary condition for some attitude's *belonging* to an agent. I turn to this last category of ownership views below.

3.3 Ownership

As I have already mentioned in section 2.6.1, the question of whether implicit biases "belong" to a person or reflect her "true self" is a contested philosophical question, This is in part because the more general question of which parts of a person's psychology are "internal" or "external" to her, or, to put it another way, whether some behavior is genuine *activity* on the part of an agent or something with respect to which she is merely a "passive bystander," has been heavily debated since Harry Frankfurt's (1988) work on first- and second-order desires. According to Frankfurt (1988), a first-order desire that is not endorsed by a second-order desire—in other words, when an agent does not *want* to want something—then that desire is an "external force" acting the agent. This discussion of Frankfurt's can be interpreted as making the following claim: if the unendorsed desire is strong enough to move an agent to action, then the agent has not *acted*, but was instead *acted on* by the desire. (Frankfurt's example is that of an

unwilling addict, a person who does not want to be addicted but whose addiction is so powerful that he ends up taking more drugs anyway.) Theorists of moral responsibility have since taken up this criterion of agency, whereby some actions are attributable to a person's exercising her rational and moral agency while other "actions"[17] are not, to delineate when agents are and are not morally responsible for actions. Actions for which we are thought to be morally responsible in this way are often called "autonomous" or "authentic" actions. They are autonomous because they are caused by an exercise of our own agency, and not by external forces acting on us; and they are authentic because they are thus truly expressive of who we are as moral agents.

Neil Levy (2012), for instance, argues within this framework that people are not morally responsible for actions caused by implicit biases. Levy relies on a neo-Lockean notion of moral agency as requiring diachronic decision-making, planning, and execution: the idea that a single *agent* is unified across time by the plans that she selects, formulates, and implements. Levy also subscribes to a "global workspace" view of consciousness, on which an attitude's being conscious means that it is being broadcast to multiple processing modules within the mind that would otherwise not be able to communicate to one another. For Levy, it is only when actions emerge from the *integration* of such disparate modules that they are truly expressive of the *person*; when they do not, e.g. when they issue from an implicit bias, they express only some "subpersonal" subset of mechanisms that contribute to moral agency.[18] And since these actions are not the product of a person's exercising her moral agency, he concludes, she is not morally responsible for them.[19]

Against Levy, Angela Smith (2012, ms) has argued that moral agency does not require conscious endorsement or voluntary control. She supports this claim with examples of more ordinary phenomena, such as the fact that we would judge a person responsible for forgetting her best friend's birthday even if she did not endorse or have control over it, or being caused to act in a sullen manner toward a partner without being consciously aware of an underlying resentment.

---

[17] In the philosophy of action literature, full-blooded "actions" are contrasted with mere "behavior," the latter of which lacks the distinctive features of rational and moral agency available to normally-functioning human adults. However, my own use of the term "action," as in "action caused by implicit bias," is *not* intended to beg any questions about whether they are properly attributable to the agent.

[18] See also Korsgaard (2009), for another view that agency requires the integration of disparate psychological parts. Velleman (1992) has a slightly different version of this view, according to which an agent is represented by "the desire to act according to reasons," which "throws its weight" behind *one* of the potentially multiple motivating desires and in doing so renders that particular desire expressive of the agent.

[19] Levy (2014), however, has adopted a more restrained conclusion. In interpreting implicit biases as "patchy endorsements" that do not fit into familiar categories of mere associations or beliefs, Levy claims as a result that we should wait for more nuanced forms of moral assessment that can accommodate such new phenomena.

For Smith, exercises of moral agency are represented by "judgment-dependent states," that is, states that "rest upon and embody rational evaluations or appraisals and…both reflect our agency and are the sorts of states for which it makes sense to ask for justification" (MS 2). Since it is intelligible to demand justification for a person's forgetting something or behaving in a sullen manner, this demonstrates that these actions or omissions were reflective of the agent's priorities or judgments of her partner's previous behavior, even if they did not involve her conscious endorsement or control. Because implicit biases are broadly cognitive and evaluative states that reflect appraisals of the way things are and whether that is experienced as good or bad, Smith claims that they *are* reflective of moral agency; she also draws on evidence by Mandelbaum (2015) that implicit biases are belief-like in their propositional structure and hence genuinely involve rational evaluative activity. She points out, furthermore, that we do not have *a priori* reason to assume that agents are so integrated as to have one single point of view, and that conflicting attitudes—such as an implicit bias and a higher-order rejection of it—may both be expressive of the agent.

Finally, another view recently proposed by Michael Brownstein (ms) is that we are responsible for implicit biases because they express our *cares*, rather than our evaluative judgments. Following Agnieszka Jaworska (2007) and Chandra Sripada (ms), Brownstein characterizes cares as things that *matter* to a person, where mattering manifests itself in a distinct emotional and motivational profile, e.g. to feel positively when things go well for the cared-for and negatively when they don't, to act in ways that would make things go well for the cared-for and prevent things that would not, etc. Moreover, it is possible to care about things that one does not evaluatively judge to be valuable: there are some things a person can't *help* caring about, even if she does not identify with that care or "recognize it as her own" and does not judge it good to care about that thing (14). We are morally responsible for what we care about because our cares express what really matters to us, and hence who we are–rather than who we ideally would like to be—as moral agents. The way we determine what matters to a person is by looking at "multitrack" patterns of attention, perception, thought, feeling, and action over time. Thus, Brownstein argues, we can understand implicit biases as manifesting an agent's cares because they serve to shape attention and perception, generating thoughts and feelings that subsequently produce action. Thus, even though an agent might not *endorse* her implicit bias against some social group, the fact that it influences her judgment and behavior (as well as affective and

emotional responses) in certain ways is evidence that she *cares* about the purported traits associated with that social group.[20] Hence, Brownstein concludes, she is morally responsible for actions caused by her implicit biases.

4. Conclusion

My own project is not to adjudicate amongst the competing views I have described in sections 3.1-3.3. Instead, my aim is to re-examine our concept—or, as I shall claim, our two concept*s*—of moral responsibility in light of the challenges raised by phenomena such as implicit bias. Precisely because implicit biases *cannot* be understood in terms of control, awareness, and ownership in clear and uncontested terms, they serve to highlight the theoretical limitations and epistemic difficulties involved with thinking in terms of those categories. However, I do not believe that we should, need, or can afford to sit back and wait for empirical scientists to carry out all the investigation that would needed to settle questions of theoretical interpretation. For one thing, it may be very long before a scientific and philosophical consensus is reached, if at all. For another, as I discussed in Sections 2.6.1-2.6.2 and again in Sections 3.1-3.3, many of the crucial questions are not answerable on the basis of empirical investigation alone: these are questions concerning whether implicit biases are personal ("belong to us") or not, and whether the kinds of awareness and control we have over them are sufficient for moral responsibility. And finally, on my view, we should not overlook the deeper issues involved in understanding the different purposes for which we need concepts of moral responsibility at all in the first place. While it is certainly important to refine our existing theories of moral responsibility in light of new understandings of the psychology of moral reasoning, it is also important to develop new theory and concepts that can guide us in deciding *what to do* in real-world cases where implicit biases manifest themselves, where "what to do" is grounded in our moral and political commitments to improving our social world. My claim is that, if we take seriously the sociopolitical exigencies of our current historical moment and what is required to improve it, certain questions determinative of moral responsibility—those that concern accountability rather than attributability—loom larger than others.

[20] While I do not have the space to fully address this issue here, it is unclear whether this formulation of care-based view as stands is really a distinct view. Since presumably *anyone* would care about traits such as competence, dangerousness, etc. *if* others genuinely possessed them, the crucial question would seem to be whether an agent "truly" believes or judges that they do possess those traits—which would, however, collapses back into the question debated by Levy and Smith.

References

Allport, Gordon. The Nature of Prejudice. Cambridge, MA: Addison-Wesley, 1954.

American Association of University Women. (2012). "The Simple Truth About the Gender Pay Gap." Retrieved from <http://www.aauw.org/learn/research/upload/simpletruthaboutpaygap1.pdf>.

Amodio, David M. (2014). "The Neuroscience of Prejudice and Stereotyping." *Nature Reviews Neuroscience* 15: 670–682.

Amodio, D.M., & Ratner, K.G. (2011). "A Memory Systems Model of Implicit Social Cognition." *Current Directions in Psychological Science* 20(3): 143-148.

Anderson, Elizabeth. The Imperative of Integration. Princeton, NJ: Princeton University Press, 2010.

Anwar, S., Bayer, P., & Hjalmarsson, R. (2012). "The Impact of Jury Race in Criminal Trials." *Quarterly Journal of Economics* 127(2): 1017-1055.

Arcuri, A., Castelli, L., Galdi, S., Zogmaister, C. & Amadori, A. (2008). "Predicting the Vote: Implicit Attitudes as Predictors of the Future Behavior of Decided and Undecided Voters." *Political Psychology* 29(3): 369-387.

Arkes, H.R. & Tetlock, P.E. (2004). "Attributions of Implicit Prejudice, or 'Would Jesse Jackson 'Fail' the Implicit Association Test?'" *Psychological Inquiry* 15(4): 257-278.

Banaji, Mazahrin R., and Anthony Greenwald. Blindspot: Hidden Biases of Good People. New York: Delacorte Press, 2013.

Banaji, M.R., Nosek, B.A., & Greenwald, A.G. (2004). "No Place for Nostalgia in Science: A Response to Arkes and Tetlock." *Psychological Inquiry* 15(4): 279-310.

Baron, Andrew S. (2015). "Constraints on the Development of Implicit Intergroup Attitudes." *Child Development Perspectives* 9(1): 50–54.

Baron, A.S., & Banaji, M.R. (2006). "The Development of Implicit Attitudes: Evidence of Race Evaluations From Ages 6 and 10 and Adulthood." *Psychological Science* 17(1): 53-58.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). "Implicit Discrimination." *The American Economic Review* 95(2): 94-98.

Bertrand, M. & Mullainathan, S. (2004). "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94(4):991-1013.

Blair, I.V., Ma, J.E., & Lenton, A.P. (2001). *Journal of Personality and Social Psychology* 81(5): 828-841.

Blank, L.M., Dabady, M., & Citro, C.F. Measuring Racial Discrimination. Washington, D.C.: National Academies Press, 2004.

Brownstein, Michael. (2015). "Implicit Bias." in Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy, URL = <http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/>.

Brownstein, Michael. "Attributionism and Moral Responsibility for Implicit Bias." (Unpublished manuscript).

Castelli, L., Zogmaister, C., & Tomelleri, S. (2009). "The Transmission of Racial Attitudes Within the Family." *Developmental Psychology* 45: 586–591.

Castelli, L., & Tomelleri, S. (2008). "Contextual Effects on Prejudiced Attitudes: When the Presence of Others Leads to More Egalitarian Responses." *Journal of Experimental Social Psychology* 44: 679–686.

Castelli, L., De Dea, C., & Nesdale, D. (2008). "Learning Social Attitudes: Children's Sensitivity to the Nonverbal Behaviors of Adult Models During Interracial Interactions." *Personality and Social Psychology Bulletin* 34: 1504–1513.

Castelli, L., Carraro, L., Tomelleri, S., & Amari, A. (2007). "White Children's Alignment to the Perceived Racial Attitudes of the Parents: Closer to the Mother than the Father." *British Journal of Developmental Psychology* 25: 353–357.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). "The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals." *Journal of Personality and Social Psychology* 83: 1314-1329.

Dasgupta, N. & Asgari, S. (2004). "Seeing is Believing: Exposure to Counterstereotypic Women Leaders and its Effect on the Malleability of Automatic Gender Stereotyping." *Journal of Experimental Social Psychology* 40: 642–658.

Dasgupta, N. and Greenwald, A.S. (2001). "On the Malleability of Automatic Biases: Combating Automatic Prejudice with Images of Admired and Disliked Individuals." *Journal of Personality and Social Psychology* 81(5): 800-814.

De Houwer, Jan. (2014). "A Propositional Model of Implicit Evaluation." *Social and Personality Compass* 8(7): 342-353.

Desilver, Drew. "On MLK Day, A Look at Black and White America." 19 Jan. 2015. *Fact Tank: News in the Numbers*. Pew Research Center. Retrieved from: <http://web.archive.org/web/*/http://www.pewresearch.org/fact-tank/2015/01/19/on-mlk-day-a-look-at-black-and-white-america/>.

Dovidio, J. F., & Gaertner, S. L. (2000). "Aversive Racism and Selection Decisions: 1989 and 1999." *Psychological Science* 11: 319–323.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). "On the Automatic Activation of Attitudes." *Journal of Personality and Social Psychology*, 50: 229-238.

Frankfurt, Harry. The Importance of What We Care about: Philosophical Essays. Cambridge, MA: Cambridge University Press, 1988.

Gaddis, S. Michael. (2015). "Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market." *Social Forces* 93(4): 1451-1479.

Gawronski
, B. & Bodenhausen, G.V. (2011). "The Associative-Propositional Evaluation Model: Theory, Evidence, and Open Questions." *Advances in Experimental Social Psychology* 44: 59–127.

Gawronski, B. & Payne, B.K. (2010). "A History of Implicit Social Cognition: Where Is It Coming From? Where Is It Now? Where Is It Going?" In B. Gawronski & B.K. Payne (eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications (pp. 1-15), New York: Guilford Press.

Gawronski, B. Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). "When "Just Say No" is Not Enough: Affirmation versus Negation Training and the Reduction of Automatic Stereotype Activation." *Journal of Experimental Social Psychology* 44: 370–377.

Gawronski, Bertram. (2009). "Ten Frequently Asked Questions About Implicit Measures and Their Frequently Supposed, But Not Entirely Correct Answers." *Canadian Psychology* 50(3):141-150.

Gawronski, B., Peters, K. R., & LeBel, E. P. (2008). What Makes Mental Associations Personal or Extra-personal? Conceptual Issues in the Methodological Debate About Implicit Attitude Measures. *Social and Personality Psychology Compass* 2: 1002-1023.

Gawronski, B., Hofmann, W., Wilbur, C.J. (2006). "Are 'Implicit Attitudes' Unconscious?" *Consciousness and Cognition* 15: 485-499.

Gendler, Tamar. (2011). "On the Epistemic Costs of Implicit Bias." *Philosophical Studies* 156: 33–63.
Goldin, C. & Rouse, C. (2000). *The American Economic Review* 90(4): 715-741.

Green A.R., Carney D.R., Pallin D.J., Ngo L.H., Raymond K.L., Iezzoni L.I., Banaji M.R. (2007). *Journal of General Internal Medicine* 22(9): 1231-1238.

Greenwald, A. T., Poehlman, T. A., Uhlmann, E., & Banaji, M. (2009). "Understanding and Using the Implicit Association Test III: Meta-Analysis of Predictive Validity." *Journal of Personality and Social Psychology* 97(1): 17-41.

Greenwald, A.G., Nosek, B.A., & Banaji, M.R. (2003). "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm." *Journal of Personality and Social Psychology* 85(2): 197-216.

Greenwald, A.G., McGhee, D.E., & Schwartz, J.L. (1998). "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74(6): 1464-1480.

Griffin, D., & Ross, L. (1991). "Subjective Construal, Social Inference, and Human Misunderstanding." In M. P. Zanna (ed.), Advances in Experimental Social Psychology (pp. 319–359). San Diego, CA: Academic Press.

Haidt, Jonathan. (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108(4): 814-834.

Hailey, S.E., & Olson, K.R. (2013). "A Social Psychologist's Guide to the Development of Racial Attitudes." *Social and Personality Psychology Compass* 7(7): 457-469.

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). "Processes in Racial Discrimination: Differential Weighting of Conflicting Information." *Personality and Social Psychology Bulletin* 28: 460– 471.

Holroyd, Jules. (2015). "Implicit Bias, Awareness, and Imperfect Cognitions." *Consciousness and Cognition* 33: 511-523.

Jaworska, Agnieszka. (2007). "Caring and Full Moral Standing." *Ethics* 117: 460-497.

Jost, J.T., Rudman, L.A., Blair, I.V., Carney, D.R., Dasgupta, N., Glaser, J., & Hardin, C.D. (2009). "The Existence of Implicit Bias is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies that No Manager Should Ignore." *Research in Organizational Behavior* 29: 39-69.

Kawakami, K., Steele, J. R., Cifa, C., Phills, C. E., & Dovidio, J. F. (2008). "Approaching Math Increases Math = Me, Math = Pleasant." *Journal of Experimental Social Psychology* 44: 818–825.

Kawakami, K. Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). "(Close) Distance Makes the Heart Grow Fonder: Improving Implicit Racial Attitudes and Interracial Interactions Through Approach Behaviors." *Journal of Personality and Social Psychology* 92: 957–971.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). "Just Say No (To Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation." *Journal of Personality and Social Psychology* 78: 871–888.

Kay, A.C., Wheeler, C., Bargh, J.A., & Ross, L. (2004). "Material Priming: The Influence of Mundane Physical Objects on Situational Construal and Competitive Behavioral Choice." *Organizational Behavior and Human Decision Processes* 95: 83–96.

Korsgaard, Christine. Self-Constitution: Agency, Identity, and Integrity. USA: Oxford University Press, 2009.

Lai, C.K., Hoffman, K.M., & Nosek, B.A. (2013). "Reducing Implicit Prejudice." *Social and Personality Psychology Compass* 7(5): 315-330.

Legault, Lisa, Gutsell, Jennifer N., & Michael Inzlicht. (2011). Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (but Also Increase) Prejudice. Psychological Science 22(12): 1472–1477.

Levin, P.F., & Isen, A.M. (1975). "Further Studies on the Effect of Feeling Good on Helping." *Sociometry* 38(1): 141-147.

Levin, P.F., & Isen, A.M. (1972). "Effect of Feeling Good on Helping: Cookies and Kindness." *Journal of Personality and Social Psychology* 21(3): 384-388.

Levy, Neil. "Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements." *Noûs* (forthcoming).

Levy, Neil. (2012). "Consciousness, Implicit Attitudes, and Moral Responsibility." *Noûs* 48: 21–40.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). "Social Influences on Automatic Racial Prejudice. *Journal of Personality and Social Psychology*, 81, 842–855.

Madva, Alexander. "Biased Against De-Biasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle Against Prejudice." (Unpublished manuscript).

Mandelbaum, Eric. "Attitude, Association, and Inference: On the Propositional Structure of Implicit Bias." *Noûs* (forthcoming).

Mann, N. H., & Kawakami, K. (2012). "The Long, Steep Path to Equality: Progressing on Egalitarian Goals." *Journal of Experimental Psychology* 141: 187–197.

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

Monin, B., & Miller, D. T. (2001). "Moral Credentials and the Expression of Prejudice." *Journal of Personality and Social Psychology* 81: 33–43.

Monteith, M.J., Woodcock, A., & Lybarger, J.E. (2013). "Automaticity and Control in Stereotyping and Prejudice: The Revolutionary Role of Social Cognition Across Three Decades." In D. Carlston (ed.), Oxford Handbook of Social Cognition (pp. 74-94). New York: Oxford University Press.

Monteith, M.J., Mark, A.Y., & Ashburn-Nardo, L. (2010). "The Self-Regulation of Prejudice: Toward Understanding Its Lived Character." *Group Processes and Intergroup Relations* 13(2): 183-200.

Monteith, M.J., Voils, C.I., & Ashburn-Nardo, L. (2001). "Taking a Look Underground: Detecting, Interpereting, and Reacting to Implicit Racial Biases." *Social Cognition* 19(4): 395-417.

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). "Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals." *Journal of Personality and Social Psychology* 77: 167–184.

Mullainathan, Sendhil. "Racial Bias, Even When We Have Good Intentions." *The New York Times*. 3 Jan. 2015. Web. Accessed 8 Jun 2015.

Nosek, B. A., & Hansen, J. J. (2008). "The Associations in our Heads Belong to Us: Searching for Attitudes and Knowledge in Implicit Evaluation." *Cognition and Emotion* 22: 553-594.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition* 19(6): 161-176.
National Center for Education Studies. (2011). "Achievement Gaps: How Hispanic and White Students in Public Schools Perform on the National Assessment of

Educational Progress." National Assessment of Educational Progress. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/studies/2011485.pdf>.

National Center for Education Studies. (2009). "Achievement Gaps: How Black and White Students in Public Schools Perform on the National Assessment of Educational Progress." National Assessment of Educational Progress. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/studies/2009495.pdf>.

Nosek, B.A., Banaji, M.R., & Greenwald, A.G. (2002). "Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site." *Group Dynamics: Theory, Research, and Practice* 6(1): 101-115.

Olson, M. A., & Fazio, R. H. (2006). "Reducing Automatically-Activated Racial Prejudice Through Implicit Evaluative Conditioning. *Personality and Social Psychology Bulletin* 32: 421–433.

Olson, M.A., Fazio, R.H., & Han, H.A. (2009). "Conceptualizing Personal and Extrapersonal Associations." *Social and Personality Psychology Compass* 3(2): 152-170.

Payne, B.K., Cheng, C.M., Govorun, O., & Stewart, B.D. (2005). "An Inkblot for Attitudes: Affect Misattribution as Implicit Measurement." *Journal of Personality and Social Psychology* 89(3): 277-93.

Pettrigrew, T.F. & Tropp, L.R. (2008). "How Does Intergroup Contact Reduce Prejudice? Meta-Analytic Tests of Three Mediators." *European Journal of Social Psychology* 38(6): 922-934.

Pettrigrew, T.F. & Tropp, L.R. (2006). "A Meta-Analytic Test of Intergroup Contact Theory." *Journal of Personality and Social Psychology* 90(5): 751-783.

Phills, C. E., Kawakami, K., Tabi, E., Nadolny, D., & Inzlicht, M. (2011). "Mind the Gap: Increasing the Associations Between the Self and Blacks with Approach Behaviors." *Journal of Personality and Social Psychology* 100: 197–210.

"Project Implicit." (2011). Retrieved from: <http://www.projectimplicit.net/>.
Radzik, Linda. Making Amends: Atonement in Morality, Law, and Politics. USA: Oxford University Press, 2011.

Richeson, J. A., & Ambady, N. (2001). "Who's in Charge? Effects of Situational Roles on Automatic Gender Bias." *Sex Roles* 44: 493–512.

Richeson, J. A. & Ambady, N. (2003). "Effects of Situational Power on Automatic Racial Prejudice. *Journal of Experimental Social Psychology* 39: 177–183.

Rooth, Dan-Olof. (2010). "Automatic Associations and Discrimination in Hiring: Real World Evidence." *Labour Economics* 17(3): 523-534.

Ross, L., & Ward, A. (1996). "Naive Realism in Everyday Life: Implications for Social Conflict and Misunderstanding." In T. Brown, E. S. Reed & E. Turiel (eds.), Values and Knowledge (pp. 103–135). Hillsdale, NJ: Erlbaum.

Rudman, L.A., Phelan, J.E., & Heppen, J.B. (2007). "Developmental Sources of Implicit Attitudes." Personality and Social Psychology Bulletin 33(12): 1700-1713.

Saul, Jennifer. "Implicit Bias, Stereotype Threat, and Women in Philosophy." In Katrina Hutchison & Fiona Jenkins (eds.), Women in Philosophy: What Needs to Change? (pp. 39-60). New York: Oxford University Press, 2013.

Schnall, S., Haidt, J., Clore, G.L., & Jordan, A.H. (2008). "Disgust as Embodied Moral Judgment." *Personality and Social Psychology Bulletin* 34(8): 1096–1109.

Shelby County, Alabama v. Holder, Attorney General, et al. No. 12-96. (U.S. Supreme Court 2013).
Sherman, S,J., Sherman, J.W., Percy, E,J., Soderberg, C.K. (2013). "Stereotype Development and Formation." In D. Carlston (ed.), The Oxford Handbook of Social Cognition. New York: Oxford University Press.

Shook, N,J., & Fazio, R.H. (2008). "Interracial Roommate Relationships: An Experimental Field Test of the Contact Hypothesis." *Psychological Science* 19(7): 717-723.

Smith, Angela. (2007). "On Being Responsible and Holding Responsible." *Journal of Ethics* 11(4): 465-484.

Smith, Angela. "Implicit Biases, Moral Agency, and Moral Responsibility." (Unpublished manuscript).

Smyth, F.L., Greenwald, A.G., & Nosek, B.A. (2009). "Implicit Gender-Science Stereotype Outperforms Math Scholastic Aptitude in Identifying Science Majors." (Unpublished manuscript).

Springer, Elise. Communicating Moral Concern. Cambridge, MA: MIT Press, 2013.

Sripada, Chandra. "Self-Expression: A Deep Self Theory of Moral Responsibility." (Unpublished manuscript).

Son Hing, L. S., Chung-Yan, G., Hamilton, L., & Zanna, M. (2008). "A Two-dimensional Model that Employs Explicit and Implicit Attitudes to Characterize Prejudice." *Journal of Personality and Social Psychology* 94: 971–987.

Suhler, C.L., & Churchland, P.S. (2009). "Control: Conscious and Otherwise." *Trends in Cognitive Science* 13(8): 341-347.

Tam, T., Hewstone, M., Harwood, J., Voci, A., & Kenworthy, J. (2006). "Intergroup Contact and Grandparent–Grandchild Communication: The Effects of Self-Disclosure on Implicit and Explicit Biases Against Older People." *Group Processes and Intergroup Relations* 9(3): 413-429.

Taylor, P., Fry, R., & Kochhar, R. (2011). "Wealth Gaps Rise to Record Highs Between Whites, Blacks, Hispanics." Pew Research Center. Retrieved from <http://www.pewsocialtrends.org/2011/07/26/wealth-gaps-rise-to-record-highs-between-whites-blacks-hispanics/>.

Turner, R.N., Hewstone, M., & Voci, A. (2007). "Reducing Explicit and Implicit Outgroup Prejudice via Direct and Extended Contact: The Mediating Role of Self-disclosure and Intergroup Anxiety." *Journal of Personality and Social Psychology* 93(3): 369-388.

Tversky, A. & Kahneman, D. (1974). "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185(4157): 1124-1131.

Velleman, J. David. (1992). "What Happens When Someone Acts?" *Mind* 101(403): 461-481.

Walker, Margaret Urban. Moral Repair: Reconstructing Moral Relations After Wrongdoing. New York: Cambridge University Press, 2006.

Washington, N. & Kelly, D. "Who's Responsible for This? Implicit Bias and the Knowledge Condition." In M. Brownstein & J. Saul (eds.), Implicit Bias and Philosophy: Moral Responsibility, Structural Injustice, and Ethics, Oxford: Oxford University Press (forthcoming).

Williams, L.E., & Bargh, J.A. (2008). "Experiencing Physical Warmth Promotes Interpersonal Warmth." *Science* 322(5901): 606-607.

Young, Iris Marion. Responsibility for Justice. USA: Oxford University Press, 2011.

CHAPTER 2

Attributability, Accountability, and Implicit Bias

1. Introduction

When assessing moral responsibility for implicit bias, most of us are likely to find ourselves pulled in opposite directions. On the one hand: we aren't aware of our implicit biases, we can't control their influence on our actions and judgment, and they often undermine our considered beliefs and judgments. So they don't look like the sort of thing we can be responsible for. On the other hand, we live at a time when many people do sincerely affirm commitments to social equality, where laws already exist to prohibit discrimination, and thus where implicit bias is likely contributes to the persistence of deep social inequalities. (Recall from Section 2.2.5 the evidence furnished by "CV studies," which show that identical CVs labelled with typical names from different social categories—African-American or White American, male or female, and so on—will receive significantly different response rates and evaluations.) Indeed, implicit biases are at once the *products* and the *perpetrators* of social inequality: unjust social structures breed implicit biases, and implicit biases impede structural change. And so we cannot simply let people off the hook. In this paper I will show how we can resolve this tension. The key is a distinction between two different concepts of moral responsibility, which I lay out and explain in Section 2. Responsibility as "attributability" depends on the notion that actions are expressions of our agency. We are morally responsible for our actions in this sense only when they reflect the practical identities that define us as moral agents—that is, when they are properly *attributable* to us as manifestations of our ends, commitments, and values. Responsibility as "accountability," however, depends on the social and institutional practices that govern the distribution of duties and burdens across different roles and positions within a society. We are morally responsible for our actions in this second sense when it is appropriate for others to enforce certain expectations

and demands on those actions—in other words, when it is appropriate for others to hold us *accountable* for them. In Section 3 I consider the question of whether we are attributively responsible for actions caused by implicit bias, and in Section 4, whether we are accountable for them. I thus reconcile our conflicting intuitions by showing that we can lack attributability for implicit bias[21], but at the same time still be *accountable* for them. What this amounts to, I further argue and defend in Section 5, is that we should refrain from what I call "appraisal-based responses" in favor of "non-appraising responses." Leaving aside ascriptions of attributability and focusing on accountability will not only lead to more effective practices for mitigating and eliminating the harms of implicit biases, but will also do more justice to our moral experience and agency.

2. Two Routes to Responsibility

A number of philosophers have explored the distinction (or very closely related distinctions) between attributability and accountability. T.M. Scanlon (1998), for instance, distinguishes between what he calls "responsibility as attributability" and "substantive responsibility". Gary Watson (2004) proposes a distinction between the "aretaic" or "attributability" face of responsibility on the one hand, and "accountability" on the other.[22] And a number of political philosophers have proposed distinctions between backward-looking versus forward-looking, and "metaphysical" as opposed to "practical" or "political" models of responsibility.[23] My own interpretation, which is more Scanlonian than Watsonian, shares similarities with these other distinctions. However, I motivate this interpretation by grounding it in a conceptual genealogy that illustrates how these two different concepts of moral responsibility arise from two fundamentally distinct sources of philosophical concern.

The first route to responsibility—attributability—begins in metaphysics and action theory. It emerges out of the "problem of action," that is, the problem of understanding ourselves

[21] For reasons of pure convenience, I will use "responsibility for implicit bias" in this chapter to stand in for "responsibility for actions caused by implicit bias." Any references to responsibility for other aspects of implicit bias will be explicitly marked. I will also focus on actions rather than judgments or beliefs. To be sure, implicit bias does operate—perhaps primarily so—on the formation of beliefs. There is undoubtedly some sense in which we can be morally responsible for our doxastic states, but to avoid such thorny theoretical issues (see Watson (2004) for an illuminating discussion) I will restrict my discussion to actions, some of which may essentially involve beliefs caused by implicit bias, e.g. the action of hiring one job candidate over another.

[22] See also Fischer and Tognazzini (2011), who offer an extensive treatment of the conceptual distinctions between, and within, attributability and accountability. Unlike Fischer and Tognazzini or Watson, however, I do not take attributability of an action to be a necessary condition of accountability for that action.

[23] See Goodin (1995), Kelly (2002), Young (2011), and Baier (1972).

as agents in a naturalistic picture of the world where we, like everything else, are just segments of a tightly-linked causal chain. The question, in other words, is this: how do we make room for the idea that we are the *origins* and *authors* of actions, and not just the arenas in which sequences of mental events occur one after the other? One prominent tradition of solutions to the problem of action suggests that we can justifiably conceive of ourselves as agents when our actions "belong" to us in the appropriate way or when we act "autonomously" rather than being passively imposed upon from the outside. On a Kantian picture, this is because rational (human) action is distinctively subject to self-reflective awareness, and thus deliberative choice or endorsement on the basis of reasons. The idea is that we can, when moved to act in some way, become aware of this motive within us; we can then deliberate about the reasons for or against that action, and ultimately choose the grounds on which we will perform or refrain from it. It is an agent's endorsement of the principles or motives on which she acts that makes it *her* action, and what makes her into a full-blooded agent rather than a mere cause.[24] Non-Kantians reject the emphasis on reflective deliberation and endorsement, but maintain that we are morally responsible for actions that express our characters and selves. Because our agency is constituted by our ends, values, and commitments—by our "practical identities", in short[25]—our actions provide the grounds for a kind of appraisal that is not available to non-rational and non-moral agents like children and animals. When and only when an action is thus *attributable* to a person is she properly subject to such assessment. While agents are responsible for all kinds of actions—clumsy or graceful, intelligent or not—many of those ways of being and acting are morally relevant. So this kind of "deep" appraisal,[26] deep because it attaches to the person *qua* agent, furnishes us with the first concept of moral responsibility when it takes in morally relevant features of an agent. *We are morally responsible for actions in this attributability sense when that action is expressive of our morally evaluable character, ends, values, or commitments.* And the responses that make up such moral appraisal, like the paradigmatic (non-consequentialist) praise and blame[27] and the reactive attitudes, are what I will call "appraisal-based responses," since they reflect assessments of a person's quality of character as a moral agent.

---

[24] See for example Frankfurt (1988), Velleman (1992), and Korsgaard (2009).
[25] The term is Christine Korsgaard's (2009).
[26] Cf. Wolf (1990).
[27] I do not have the space to offer a full account of blame here. However, I take it to essentially involve an assessment of the agent on the basis of an action; the action "sticks" to the agent, as it were, in a way that renders judgment of the action also a judgment of the agent. See also Ch. 3, Section 2 for further discussion.

Attributability is typically understood to depend primarily on metaphysical or psychological facts about the relation between an agent and her action: about the chain of (mental) events leading up to or constituting the action, or else about the attitudes and other mental states directed toward the action. Theories of (attributive) moral responsibility thus propose that we are only morally responsible for actions if we identify with the desires that moved us to them, if they express our deep selves, if they constitute the agent as an integrated whole, or if we could have done otherwise[28]. By the same token, a number of "excusing conditions" are generally accepted as specifying when actions cannot be attributed to agents; these include behaving unknowingly, unintentionally, accidentally, under coercion, or in an altered state of mind.[29] These are conditions under which a person's behavior does not flow from her practical identity—in other words, when she is not acting fully as an agent. Some beings, like young children and non-human animals, are also subject to "exemption conditions," which indicate a global lack of well-formed characters or the capacities required to reflectively deliberate and choose ends.

Contrast this with the second route to responsibility—accountability—which has its origins in moral and political philosophy.[30] It is practical through and through, and "responsibilities" show up in the first stage as the solution to a problem concerning the moral division of labor.[31] People are assigned various responsibilities (often called "prospective" or "role" responsibilities) for carrying out particular duties and tasks that serve our social goals (which are generally also moral goals). This notion of substantive role responsibilities thus gives rise to a second, distinct concept of moral responsibility. *We are morally responsible for an action in this accountability sense when it is appropriate for others to hold us to certain expectations and demands with regard to our duties and tasks—and to sanction us when we fail to carry them out.* When a person's action brings about some negative consequences for others, this generates a social problem: those costs must be picked up somehow and by someone. These are problems that simply cannot go unaddressed,

[28] Here I am alluding to views espoused by the likes of Frankfurt (1988), Wolf (1990), Korsgaard (2009) and Levy (2014), and in the free will debate.

[29] I will be somewhat lax in my use of the terms "action and behavior," so as not to commit myself to any particular action of theory (many of which sharply distinguish the former as exemplifying certain, perhaps distinctively human, features that the latter does not). But this in itself reflects something about the difference between attributability and accountability; for only "full-blooded" actions will in general be attributable to us, while we can be accountable even for "mere" behavior.

[30] Cf. also Oshana (1997), who distinguishes between determining the conditions of free or responsible agency, on the one hand, and requiring an agent to give an account of her action, e.g. when she fails to carry out some task, on the other.

[31] I am indebted to Elizabeth Anderson for this way of putting the point.

even if there are no bad intentions or any fault on the part of the persons involved, because there are victims who deserve redress. This means that under a fair and equal system of distributing burdens, it will often be appropriate for the person who performed the action to bear a large share of the costs: she can be asked to compensate for damages, make reparations, or to change her practices to prevent future failures. But notice that none of this requires an assessment of character, intentions, ends or values—she would be required to bear at least some of the costs of her behavior whether she had performed them out of malice, negligence, or sheer (non-culpable) ignorance or accident. In a traffic accident, for instance, it is appropriate for the driver (or her insurance company!) to pick up the costs of the damage, whether she was driving recklessly or perfectly responsibly; she ought to do so either way, and it is a further, separate question whether there should be sanctions on top of that. In other words, a person's behavior need not reflect anything about her at all in order for her to be appropriately required to deal with their consequences. These ways of holding agents accountable for their actions are thus what I call "non-appraising responses". Of course there *is* still an important moral difference between the merely unlucky and the negligent driver. Thus for social and moral reasons, we might want to further blame or punish someone for her failure to fulfill her responsibilities, e.g. if it would deter future violations. This could be justified if it were expedient (according to consequentialists), or only if it were truly warranted (according to non-consequentialists), or perhaps both.[32]

[32] My own view is any plausible theory which requires appraisal responses to be normatively justifiable will need to revert back to questions of attributability: condemnatory appraisal-based responses are only deserved if the action is genuinely expressive of or reflects on the agent herself. Strawsonian proponents of appraisal-based responses still have recourse to notions of attributability: see for instance appeals to the "quality of will", for Strawson (1962), or desert, for R.J. Wallace (2004, p. 107). But attributability here still plays a secondary role, relative to the primary goal of maintaining social relationships and institutions. It is only because human relationships so often depend on knowing the true intentions and characters of others that these become relevant.

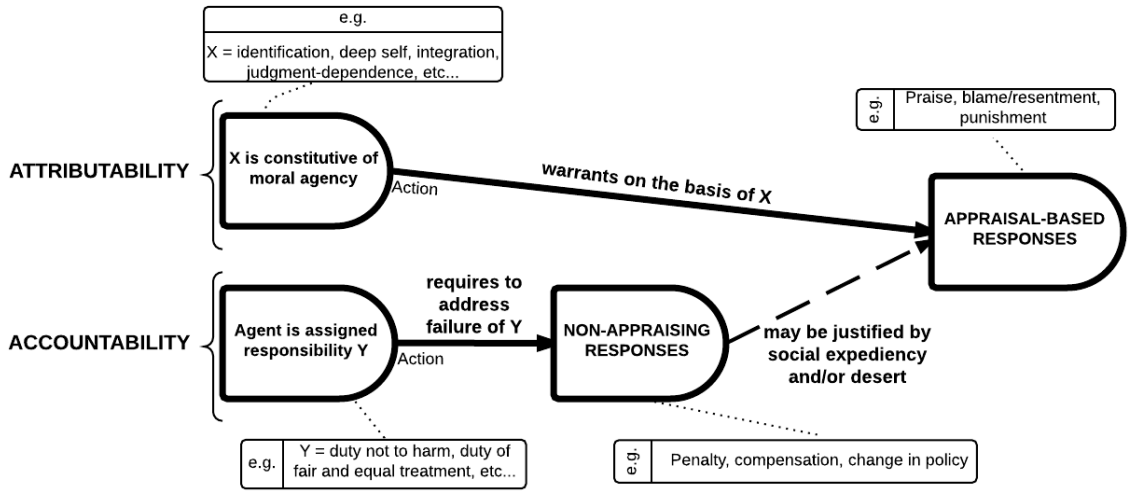

Figure 5: Two Routes to Responsibility. Note that this framework is intended to be as ecumenical as possible, allowing for many conceptions of attributability, and for both consequentialist and non-consequentialist, Strawsonian and non-Strawsonian theories of appraisal-based response.

Note that we arrive at appraisal-based responses for both kinds of responsibility, but along different paths. (To better orient myself in relation to the literature: I construe both consequentialists and Strawsonians as taking the accountability route, where moral responsibility is defined in terms of the aptness of particular social relationships and practices. Non-Strawsonians, by contrast, take the attributability route: they are fundamentally and directly concerned with whether metaphysical relations hold that would make it the case that people deserve praise, blame, and punishment.[33]) Accountability is thus primarily a matter of interpersonal, not metaphysical, relations: not what it takes to be an agent, but what it takes to be a member of a community of agents. Of course, being a member of a community of agents requires first satisfying the minimal conditions of being an agent (whatever those are), but the sets of questions that are most salient for ascribing attributability and accountability differ in their focus.

As a last way to illuminate the contrast between attributability and accountability, consider another commonly invoked distinction between being responsible and holding responsible. In my sense, roughly, attributability is about being responsible and accountability is about being held responsible. I say "roughly," however, because the difference is more a matter of priority relations, not definition. We can *hold a person attributively responsible* by praising or blaming her, but this is only licensed when she is responsible in the right way—that is, when her

[33] I am grateful to Neil Levy for pushing me to clarify this point.

action flows from her practical identity. Conversely, a person *is accountable* when she is being held to an appropriate network of expectations, demands, and responses by others in her community. In the former case, holding responsible depends on first being attributively responsible; in the latter, being responsible depends on first being appropriately held accountable (beyond the minimal conditions of agency).

My aim in this section has been to clarify the two distinct families of questions that each *concept* of responsibility raises, without yet providing substantive or comprehensive *conceptions* of attributability or accountability that would answer those questions.[34] Similarly, in what follows I will sketch out the contours of a view about responsibility for implicit bias, without grounding it in any particular first-order theory of attributability or accountability.

3. Attributability of Implicit Bias

Recall that attributive responsibility is thought to require some sort of special metaphysical or psychological relation between an agent and her actions. It is this concept of responsibility, I suggest, that we have in mind when we are inclined to let people off the hook for implicit bias. To show this, I will reconstruct what I call "The Simple Argument" against ascribing moral responsibility for actions influenced by implicit bias. I will not be arguing that the argument is sound, but that it is instructive to examine the motivations behind it.

The strength of the argument, as I see it, rests largely on two standard intuitions about moral responsibility. The first is this:

> **The Distinctiveness Intuition** We are responsible for our actions and attitudes in a way that young children, non-human animals, and other such beings are not.

I take it that this is uncontroversial: though the difference may be one of degree rather than kind, and though it is difficult to pinpoint exactly what it consists in, it seems clear that non-human animals and children are not fully agents—or at least, not fully moral agents—in the way that normal human adults are. The second intuition is as follows:

> **The Endorsement Intuition** There is some important moral difference between a person who would endorse the influence of an implicit bias on her judgment and behavior if she were reflectively aware of it, and some other person who would reject such an influence.

[34] I follow Rawls (1999) in distinguishing between a "concept," which is the general specification of some notion and the function it performs, and a "conception," which is a particular substantive theory of what realizes or instantiates that concept.

Again, this seems absolutely correct—indeed, vitally important—because to deny that there is any difference whatsoever between the two cases is to lose sight of fundamental features that make us moral agents at all, or that make it possible (at least in part[35]) to have fully-developed moral characters: our capacity for reflecting on and choosing our ends. These two intuitions lead naturally to the following argument:

> **The Simple Argument**
> 1. We are attributively responsible for our actions when and only when we stand in some special relation to them.
> 2. We do not stand in such a relation to our actions when they are caused by implicit biases that we would not endorse.
> Therefore,
> 3. We are not attributively responsible for actions caused by implicit bias.

The Simple Argument, simple though it may be, is actually rather compelling. Premise 1 is supported by the Distinctiveness Intuition that, since young children and non-human animals are not morally responsible for their behavior in the way we are, there must be something special about normal adult human action. The Distinctiveness Intuition also drives Premise 2: since implicit biases are one of the older automatic processes that we share with other species, and *not* the more recently-evolved deliberative processes distinctive to normal adult humans, we have reason to think that actions caused by implicit bias lack that special, distinctive feature. The Endorsement Intuition further buttresses this line of thought by directing us toward what the special feature might be: it must be somehow related to our distinctive capacity for choice and reflective endorsement of (influences on) our actions. This schema can then be adapted to a variety of candidates for the special responsibility-conferring relation.

Jennifer Saul (2013), for instance, has put forth versions of the Simple Argument in which she contends that people are not blameworthy for having (and, presumably, acting on) implicit sexist biases under the following conditions: when they are completely unaware of them, when those implicit biases result solely from living in a sexist culture, and when mere awareness does not enable people to immediately control them. Each of these can be understood as specifying some way in which the special relation fails to hold, in line with the Endorsement Intuition: we

[35] Note, for example, that even Smith (2004), who explicitly rejects the view that moral responsibility requires conscious endorsement, still restricts moral responsibility to attitudes that are dependent on our judgment, that is, which rest on our rational evaluation of reasons and for which it makes sense in principle to demand justification.

do not endorse our implicit biases if they result solely from external cultural forces, we cannot endorse them if we are not aware of them, and our inability to control them means that they operate without our endorsement. Lawrence Blum also suggests that we may not be responsible for implicit stereotypes, since "the entirely automatic and cognitively uninvested character of stereotypic associations" allows for "at best an extremely minimal epistemic, and moral, responsibility" (Blum 2004, p. 270). Again, we should understand deliberativeness and cognitive investment as candidates for the special relation that must hold between us and our actions if we are to be attributively responsible for them. And Neil Levy (2012) has defended the view that consciousness is required for moral responsibility because it performs the function of integrating disparate cognitive processes in a way that is required to constitute a genuine moral agent that can reason, plan, and execute projects over time; in other words, integration into the rest of a (mostly) unified agent is necessary in order for there to be a moral agent at all, with a perspective from which she evaluates and endorses reasons to act. Levy argues that, since implicit associations cannot be used in rule-based reasoning or to plan coherent projects over time, the actions they cause cannot be attributed to a unified agent but only some narrow subpart of her, and she is thus not morally responsible for them. To be sure, other philosophers have argued persuasively against just these sorts of proposals. But this points to another virtue of the Simple Argument: the fact that it admits of many different particular ways of filling it out. If awareness, control, non-automaticity, cognitive investment, and consciousness do not do the job alone, it is easy to remain steadfast that the elusive special feature or features—whatever it is, or in whatever combination is required—is also missing from actions caused by implicit bias.

I won't stake out any claims about whether the Simple Argument can ultimately be rendered sound, though I think there are reasons to doubt that it can. For one thing, as more and more candidates for the special feature or relation get eliminated, the argument may begin to lose some plausibility.[36] Moreover, as Nomy Arpaly (2004, p. 145) has argued, it may turn out that the difference between normal human adults and children and animals can be explained much more prosaically, without any appeal to complex notions like autonomy or authenticity as a condition of moral responsibility. Children and animals may lack the intelligence and communication skills required to engage in moral life at all—that is, they fail to satisfy a *precondition* for moral responsibility, not just a condition of it. From the fact that children and

[36] See Holroyd (2012) for an excellent catalogue of just such arguments.

animals do not reflectively endorse their behavior (or participate in whatever special relation is proposed), we should not be too quick to draw the conclusion that *we* are also not morally responsible for our actions when they lack reflective endorsement (Arpaly 2004, p. 147).

Instead, I will argue for a weaker claim that still accounts for our standard intuitions. I claim only that we are not *always* attributively responsible for actions caused by implicit bias. It is worth noting at this point that I have been referring to actions "caused" by implicit bias; this is to be distinguished from actions that have merely been somehow influenced by implicit bias. To say that an action is "caused by implicit bias," in my usage, is just to say that it *would have turned out differently* were it not for the influence of the implicit bias.[37] Given the pervasiveness of implicit associations in our cognitive economies it is likely that many or most of our actions are subject to the influence of unendorsed implicit biases, but this does not mean we lack moral responsibility for most of our actions. So I restrict my attention to cases where the implicit bias is actually difference-making. It is in these cases that implicit biases may most closely resemble what Frankfurt (1988, p. 61) calls "external" psychological forces: forces which act on us and with which we do not identify. Since we are only "passive bystanders" to actions caused in this way, they cannot be properly attributed to us. Contrast this with a case in which an implicit bias is merely *influenced* by an implicit bias, that is, where the person would have behaved in the same biased manner even without the presence of that implicit bias. I take it that our intuitions in this case would be that the biased behavior can still be properly attributed to the person. (This resembles the lesson from Frankfurt-style free will cases: we think a person attributively responsible for an action even if she could not do otherwise if it is also the case that she would have acted the same way even if she *could* have done otherwise.) Actions caused by implicit bias, on the other hand, more closely resemble hypnosis or irresistible drug addiction (if these exist), or sheer inadvertent accident, where we intuitively do not find people attributively responsible. While it will often be difficult to detect whether a particular action on a given occasion was genuinely caused by implicit bias, I want to emphasize that we have good reason to think that cases of difference-making bias do occur—and if they do, that they matter very much. As I described in the introduction, research has shown that people usually do not manifest racial bias in judgments where the evidence is clear and uncontroversial, and where failure to make the

[37] Per Fn. 21, an action caused by implicit bias (e.g. hiring one job candidate over another) could be one that would not have been performed if a judgment essential to it (e.g. believing that she was more qualified) would have been different were it not for the influence of an implicit bias.

proper judgment could easily be chalked up to racial attitudes. But when evidence is ambiguous enough that a non-racial explanation is possible, people *do* manifest implicit racial bias in their judgments (Pearson, Dovidio, and Gaertner 2009).

Without much further argument[38], then, I propose that we are not attributively responsible when both of the following conditions hold. (Note that I am not trying to offer a full theory of attributability for implicit bias here; I am only proposing the existence of *some* cases where we lack attributability, which I believe must satisfy these conditions.)

Conditions for Excuse

1. The agent would not upon reflection endorse the influence of the outcome-shifting bias, and
2. The agent has done what she can reasonably be expected to do with respect to avoiding and responding to the implicit bias.

The first condition captures the standard intuition that reflective endorsement plays some important role in moral responsibility. If an agent *would* endorse the influence of an implicit bias on her action, then that is an indication that the action *is* an expression of her practical identity.[39] Since it is properly attributable to her, she is morally responsible for her action, whether or not she can control or be aware of it.

The second condition is intended to screen out what Fischer and Ravizza (1998) call "tracing cases," in which an agent is indirectly responsible for an action because it can be traced back to other actions or omissions for which she is directly responsible. For example, a drunk driver is still attributively responsible for causing a traffic accident, but a driver who hydroplanes in pouring rain is not. We can view the first driver's causing the accident as a downstream effect of her agentic values, ends, and choices to drink so heavily, and we could justifiably go on to conclude that she was reckless, imprudent, or inconsiderate of the safety of others. On the other

[38] It bears note, however, that my proposal fits well with contemporary social psychological theories of how people actually do process information in forming judgments of blameworthiness. Malle, Guglielmo, & Monroe (2014), for instance, propose a bifurcated process by which people attempt to discern 1) if the event was brought about intentionally, what were the agent's reasons for so acting, or 2) if the event was brought about unintentionally, whether the agent should have (i.e. obligation) and could have (i.e. capacity) prevented the event. My first condition effectively establishes that the event was not brought about intentionally since it involved an unendorsed bias. My second condition invokes both an agent's obligation and capacity to prevent the influence of bias, since it is her obligation that grounds our expectation that she so do, but that expectation must be reasonable in light of her limited capacity to do so. Similar remarks apply to Alicke et al (2008).

[39] I say that endorsement *indicates* that the action is an expression of the agent's practical identity; it does not necessarily *constitute* that expression. Here again I am trying to be ecumenical with respect to different theories that could ground these conditions, by allowing that endorsement might be an evidential rather than a criterial condition on responsibility. I am thankful to Chandra Sripada for discussion on this point.

hand, we cannot draw such conclusions about the second driver because the car accident does not tell us anything about her choices, values, or ends at all, and because she otherwise behaved as could reasonably be expected. (Note, as I mentioned earlier, that both drivers can be expected to help pick up the costs of the accident. This is a matter of accountability, not attributability.) Analogously, then, a person who has made reasonable efforts to respond to or avoid implicit biases is not attributively responsible for such bias when it influences her action despite these efforts (though she will still be accountable).

Can I say more about how to determine what a person can "reasonably be expected to do" with respect to acquiring and responding to the implicit bias? Yes, I think, but I can only hazard a few cautious claims here. While it may be possible to completely eliminate or avoid acquiring any morally objectionable implicit biases, in many cases this would require an unreasonable amount of time, effort, and resources. Children as young as six years old show evidence of racial implicit biases (Baron and Banaji 2006), and cannot be expected to seek the social environments required to avoid them. It is only slightly less unreasonable to expect parents to provide these; minimizing children's exposure to gender stereotypes, for example, is notoriously difficult. The costs of, say, refusing a kindly relative's birthday present or switching out one's circle of friends might be high. On the other hand, we can be expected to stay vigilant when deciding to engage in activities or enter environments that are especially likely to foster certain implicit biases, such as choosing to watch genres of entertainment that rely on stereotypical stock characters. And once we are made aware of our own implicit biases, we can certainly be expected to take measures that reduce their impact, like implementing anonymous evaluation procedures or exposing ourselves to counterstereotypical conditioning.[40] (Again, though, the costs of completely eliminating their impact—say, by forgoing all face-to-face talks, conferences, and interviews—might be too high.) Preventing and responding to implicit bias is thus an imperfect duty: we cannot never try to do it, but we are not obligated to spend all our time and resources on it. As with any imperfect duty, gauging precisely when it is fulfilled is an extremely difficult matter that I cannot hope to settle it here.[41] It is my view, however, that we would do well to focus our

[40] See for example Dasgupta and Greenwald (2001) and Olson and Fazio (2006).

[41] There is also a deeper kind of *moral* indeterminacy here, which I discuss further in Ch. 3, Section 5. If, as Marcia Baron (1987) argues, there simply is no clear demarcation (outside of clear cases at the extremes) of when our imperfect duties are or are not fulfilled, and if it is in the nature of morality itself that there be such indeterminacy, then there may simply be no fact of the matter whether a particular agent is attributively responsible for some particular act caused by implicit bias. This would be different from cases where a person is blameworthy (i.e. when

energies on elucidating this criterion of reasonable expectation for general duties of preventing and eliminating bias, rather than getting preoccupied over whether people are attributively responsible for particular instances of manifesting implicit bias. This criterion of reasonable expectation—with its reference to duties and expectations—forms the bridge which conjoins responsibility as attributability to responsibility as accountability, to which I turn in the next section.

3. Accountability for Implicit Bias

I have argued that there are some limited conditions under which we are not attributively responsible for implicit bias. Now I claim that, even under those conditions, we can still be *accountable* for them. We are accountable because it is appropriate for us to clean up after our own actions when a mess has been made. Spilled milk has got to be wiped, though we needn't impugn a person's character just for having spilled it! This need to make amends is particularly urgent when the "spilled milk" is the harm suffered by the victims of implicit bias. After all, from the victim's perspective the damage is done whether anyone is attributively responsible for it or not: harm is harm, and she is owed compensation, apology, and redress.[42] Still, we can often make such demands without invoking appraisal-based responses.

Take Scanlon's discussion of substantive responsibility, which roughly corresponds to my concept of accountability (1998, p. 266). Scanlon uses an example of strict liability to show that a person can be substantively responsible for some action without being attributively responsible. In his example, a milk vendor accidentally sells some contaminated milk, through no fault or negligence on her part. But since there are laws against selling contaminated milk, it is still appropriate for her to bear the costs of having done so: she can be required to recall her product, pay a fine, compensate those who were harmed, change her inspection practices in the future, and so on. This is so even though she had no bad intentions or indeed did not act deficiently in any way. On the strict liability model, she remains accountable because she had the opportunity to avoid incurring such burdens: she could have refrained from entering the milk business. Entering the milk business, in effect, is an example of entering what Scanlon calls an "affected area": those who do so lay down their rights not to be penalized for violating the regulations that

---

she genuinely does not satisfy the conditions for excuse), but where we are not justified in blaming her because we lack the evidence.

[42] I am grateful for discussion with Kristie Dotson on this point.

govern the affected area. It is worth noting that because these penalties are given regardless of fault, strict liability requires strong justification; the social goals (e.g. public safety) it promotes must be sufficiently important to outweigh the risks imposed on people who enter the affected area.

Strict liability provides an easy example for understanding how we can have accountability without attributability. (Note, however, that I am not advocating a strict liability model for all cases of implicit bias or even to justify it as a practice; I am merely using it to demonstrate that the idea of accountability without attributability is not such an unfamiliar one, and is compatible with existing practice.) And it seems clear that many of the most pressing situations that involve implicit bias can be read as calling for strict liability. Where people enter into positions that call for them to make decisions regarding others' merits and ability—hiring, promotions, evaluations, admissions, and the like—they accept responsibility (as accountability) for that position and lay down the right not to be penalized for failing to uphold institutional standards of decision-making. This line of argument is manifested in "disparate impact" clauses of anti-discrimination employment law, which operate when it can be shown that a policy differentially affects two groups that are entitled to equal consideration, and that there is some another policy that could have been implemented without leading to such differences. No intent to discriminate need be shown; the policy does not have to reflect prejudicial motives or anything at all about the employer. Indeed, as I suggested earlier, these may very well be cases satisfying the conditions under which agents are not attributively responsible: they may have been caused by biases which the employer does not endorse. What matters is simply that people have been harmed in a way prohibited by law, and this demands redress. Strict liability here is justified by the great importance of the goals of social equality and fairness promoted by such anti-discrimination laws.

Strict liability when it comes to implicit bias in hiring and admissions is thus a straightforward instance of holding people accountable for implicit bias. But what about harder cases? What about more ordinary, everyday instances of harmful actions brought about by the influence of implicit biases? I have in mind here what Rowe (2008) and Sue (2010) call micro-inequities or microaggressions, prototypical examples of which include clutching purses, crossing the street, avoiding eye contact or otherwise awkward interactions, and jokes that are not maliciously intended but still alienating. Such quotidian situations, far from being "affected areas"

that one could choose to enter or not, are ubiquitous and unavoidable, and the duty to avoid harming others is not a burden one can refuse. Many of these behaviors may be subtle and unconscious, but they cumulatively have a real and large effect on the lives of historically disadvantaged groups. Thus, people on the receiving end of such treatment—and I hasten to add that the same person can be on both receiving and giving ends—are undoubtedly harmed. Yet it seems clear that these cases should not be dealt with using legal sanctions; such legislation would be ineffective at best and draconian at worst. On the other hand, if I am right that we often cannot be assured that people are attributively responsible for their micro-aggressions, then resentment, blame, and other appraisal-based moral responses are also not be justified. What this points to, I think, is a paucity of more nuanced forms of moral communication. Given what we now know about subconscious, unendorsed influences on ordinary reasoning and behavior, we should expand our moral repertoire accordingly to capture the shades and varieties of these more complicated processes. What we need, in effect, are more ways of holding people *accountable* for their biases without *attributing* those biases to them—to engage in moral criticism that does not amount to accusations of racism, sexism, or condemnations of bad character. (This will be the subject of the next chapter.[43]) In any case, the justifiability of non-appraising forms of holding people accountable for implicit bias is what explains our intuition that people should not get off the hook for actions that contribute to social inequality, even when those actions are unknowingly influenced by unendorsed biases.

4. Accountability Without Attributability

My arguments in the last section reveal a way to escape the tension I outlined at the beginning of the paper. To sum up, I have distinguished between two different concepts of moral responsibility, which account for our conflicting intuitions with regard to moral responsibility for implicit bias. I argued that there is a set of conditions for excuse under which we are not attributively responsible for actions influenced by implicit bias, but that even then we can still be held accountable for them. I will now argue for the following claim: *we ought to hold people morally accountable, but not attributively responsible, for actions caused by implicit bias.* In doing so we can resolve

[43] A number of philosophers have already issued a call for more nuanced moral responses; see for example Smiley (1992, p. 254) on "a form of accountability that falls somewhere between legal punishment, on the one hand, and intrinsic moral guilt, on the other" and Moody-Adams (1994, p. 303) on the stance of the "forgiving moralist". Others have started developing such possibilities, like Fricker's (2007, p. 104) "resentment of disappointment" and Springer's (2013) model of criticism as communicating moral concern.

the tension between our conflicting intuitions that we both are and are not responsible for implicit bias. We can also dissolve a dilemma that Cheshire Calhoun has posed between what she calls the "justification" and "point" of moral responsibility.[44] Calhoun (1989) is concerned about "abnormal moral contexts" where certain kinds of wrongdoing are socially accepted and normalized, and moral knowledge is only available to a select enlightened subgroup—as we surely are now with respect to knowledge of implicit bias. Under these conditions, she thinks, a person's ignorance excuses them from responsibility, but failure to sanction them will serve to perpetuate the problematic social norms. But we can see from what I've argued that this is a false dilemma: normalized ignorance excuses from *attributability*, but not accountability. We can still engage in non-appraising responses that educate and exhort people to behave in ways that would combat structural injustice and implicit bias.

It may be objected at this point that, even if everything I have said so far is correct, I have not shown why we should refrain from ascriptions of attributive responsibility for implicit bias *in addition* to accountability, especially since I have been at pains to emphasize that excuses may exist only in a limited range of cases. This complaint may be reinforced by the thought that it would be overly sanguine to think that there do not still exist large segments of the population who *would* endorse their implicit biases, or who have not otherwise behaved as could reasonably be expected with respect to avoiding and responding to them.[45] Evidence from a 1998 study by Patricia Devine and E. Ashby Plant, for instance, showed that individuals committed to egalitarianism for its own sake ("internally motivated egalitarians") showed lower levels of implicit bias than those committed to egalitarianism out of concern for others' approval ("externally motivated egalitarians"). Furthermore, Gawronski, Hofman & Wilbur (2006) argue that there is a sense in which many individuals *can* be aware of their implicit attitudes. This "content awareness" is shown by the variability in correlation between explicitly reported attitudes and implicitly measured attitudes; for example, individuals subject to a "bogus pipeline" manipulation where they are told that experimenters can detect false statements subsequently report attitudes that more closely match their implicitly measured attitudes. This again suggests that some people really would in their heart of hearts endorse their implicit biases. Given this evidence, why should we not try to sort out, for the purposes of appraisal-based responses like blame and the reactive attitudes, the people whose implicit biases really are attributable to them?

[44] Cf. also Langton's (2001) distinction between the "accuracy" and "usefulness" of resentment.
[45] I am grateful to Lawrence Blum for discussion of this point.

First of all, I want to point out that on my account, externally motivated egalitarians really are more likely to turn out as attributively responsible for their implicit biases. To the extent that their implicit biases are linked up with these other attitudes, it is less likely that their actions were caused by implicit bias, since these other attitudes could have produced the action even without the additional presence of the implicit bias in question. This evidence also suggests that externally motivated egalitarians might actually endorse rather than reject the influence of their implicit biases, say under a different description ("non-Whites are less likely to be competent" rather than just "discriminate against non-Whites"). But this would violate my second condition as well, rendering those biases properly attributable to them. Finally, it is worth noting that even those with content awareness of their biases might still satisfy my excusing conditions. The same Gawronski, Hofman, and Wilbur (2009) paper presented evidence that people still lack "impact awareness," which is the awareness of how an attitude influences other psychological processes. Thus a person who knows about her biases might still reject the influence of such a bias on some piece of reasoning, for instance if her intent is to be an impartial judge. It might be said that the action still reflects an agent's flaws if the bias ends up influencing her after all, but this seems beside the point. It would be unfair to blame a person for treading on your foot by total accident, even if she only intended to avoid harming you out of a self-interested fear of punishment rather than concern for your well-being. The point, in both cases, is that the lack of intention makes it less of a full-blooded *action* genuinely attributable to the agent. (In reality, however, it seems very unlikely that anyone who endorsed the content of their bias would not fail to satisfy the second condition for excuse: that of having done what could reasonably be expected to avoid developing a bias. So for the most part, people who endorse the contents of their biases are likely to be attributively responsible.) In the rest of this section I will give two positive arguments in defense of the claim that we ought to refrain from ascriptions of attributive responsibility. One is moral, the other pragmatic, and they work in tandem. On this view, we should refrain from appraisal-based responses for *moral* reasons if people are not attributively responsible (as is likely to be the case if they are internally motivated) but we should also refrain from appraisal-based responses for *pragmatic* reasons if people really are attributively responsible (if they are externally motivated). In either case, since we should always refrain from appraisal-based responses, we should not bother with attempts to determine the attributability of particular instances of implicit bias.

The moral argument is that we should err on the side of caution when it comes to engaging in appraisal-based responses. On the conditions for excuse that I have proposed, it will be nearly impossible, epistemically, to determine whether they hold in *particular* cases. The counterfactuals involved, which must account for the myriad influences that culminate in action, will be extremely complex, and it is unlikely that even our best neuroscience will be able to achieve the fine-grained assessments needed to determine conclusively whether a particular action on a given occasion was really caused by a particular implicit bias. Of course, when we have good evidence on *other grounds* that implicit biases are attributable to a person—for example, if she demonstrates repeated offenses without repentance—then we are justified in blaming or resenting her. This is why I noted that we ought to spend more of our efforts on understanding and education about general duties of awareness, prevention, and response to implicit bias, because these can be invoked without getting entangled in complicated psychological investigation.[46] But otherwise, it is a serious thing to engage in appraisal-based responses, especially when it involves socially sensitive domains of race and gender, or when these judgments could be hypocritical.[47] For those who are sincerely and internally motivated, ascribing attributive responsibility for unwanted implicit biases is too harsh. It would be like accusing a cashier of trying to cheat you on the basis of an innocent arithmetical error, or derogating a devoted but inept violinist ("Sounds like you never practiced a day in your life!"). It would be inappropriate in these cases, I take it, because such remarks get close to something like violating the person's dignity, and constitute a kind of disrespect. Or perhaps more aptly, it would be like blaming a person for a behavior that they acquired as the result of some trauma, which gets triggered under certain circumstances; while such a disposition is something to be managed by her and others, it is not something for which she deserves blame or deep moral criticism.[48] (Indeed, it is likely that we are all unavoidably scarred by moral flaws: this is a tragic reflection of a deeply unjust society.) We all fall short of our ideals—and should be informed

[46] It may be difficult to determine whether the agent is guilty of, say, complicity or willful ignorance. I think the answers here will in part depend on the criterion of reasonable expectation, so for example, is their ignorance of their biases something that we could reasonably have expected them to overcome? I am grateful to Nathaniel Coleman for discussion of these cases.

[47] Appraisal-based responses might be hypocritical if we lack the moral standing required for them—if we ourselves (as is likely) are prey to the same failings. See Hieronymi (2004) and Smith (2007) for the view that justifiable praise and blame require a certain moral standing.

[48] Here, with respect to this particular trait, we adopt the Strawsonian "objective" attitude, the attitude we take towards non-human animals, young children, and beings that are not fully moral agents. I am indebted to Sarah Buss for this example.

when we do—but being informed so need not amount to being condemned. Holding people accountable by demanding amends or revised behavior is enough. In fact, people who do hold genuine egalitarian commitments ought to embrace the latter. If such responses to discrimination were totally detached from accusations of racism, sexism, and the like, it is likely that many more people would be ready to accept them and strive to do better. I believe this would be especially true of the everyday harms I discussed earlier; it remains frustratingly difficult to convey that jokes, language choice, and many other aspects of unreflective daily life can be morally objectionable even if they are not maliciously intended.

This last observation puts us in range of the pragmatic argument for refraining from appraisal-based responses. As Saul (2013) puts it: "If acknowledging that one is biased means declaring oneself to be one of those bad racist or sexist people, we cannot realistically expect the widespread acknowledgement that is requires. Instead, we'll get defensiveness and hostility" (55). Appraisal-based responses thus carry the cost of having to overcome resistance to changing practices that such defensiveness and hostility can engender. Indeed, Czopp, Monteith & Mark (2006) found that high-threat confrontations (where the subject was accused of sounding "like some kind of racist") induced higher levels of denial and resistance, negative affect, and negative evaluations of their confronter, as compared to low-threat confrontations (where the subject was pressed to be a "a little more fair"). Although both groups of subjects subsequently showed less stereotypical responses after the confrontation, other studies suggest that we should not conclude that confrontations will always be effective, particularly in the long term. Legault, Gutsell, and & Inzlicht's (2011) study on the ironic effects of anti-prejudice messages showed that participants presented with a pamphlet detailing external reasons for controlling bias showed higher levels of both explicit and implicit bias than participants who received no intervention at all. Plant and Devine's (2001) study on internally and externally motivated egalitarians showed that, among the latter, compliance with pressure to treat Blacks favorably actually provoked a backlash: in addition to experiencing greater feelings of anger and threat, externally motivated egalitarians were more likely to show greater resistance and were less likely to comply in the future compliance after the removal of direct pressure. These last results are supported by research on the conditions necessary for norms to become internally motivated. These conditions are autonomy (the need "to self-organize experience and behavior and to have activity be concordant with one's integrated sense of self"), competence (the need "to have an effect on

[one's] environment as well as to attain valued outcomes within it"), and relatedness (the need "to feel connected to others") (Deci and Ryan 2000, p. 231). As Katherine Bartlett (2009) summarizes in the *Virginia Law Review*:

> [T]hreat and confrontation about race and gender bias, which people do not want to possess or exhibit, may inadvertently provoke shame, guilt, and resentment, which lead to avoidance and resistance, and ultimately to more stereotyping. In other words, pressure and threat will often deepen bias rather than correct it. Positive strategies that affirm people's good intentions, in contrast, engage people constructively in defining their better, nondiscriminatory selves and aligning their conduct accordingly. (1901)

It is clear how knowledge of implicit biases over which they have no awareness or control—and the prospect of being negatively judged on the basis of it—can decrease people's experiences of autonomy and competence.[49] Feeling that one may be blamed or viewed as racist or sexist, and being unable to trust that one can proceed on others' good faith and the benefit of the doubt, undoubtedly prevents feelings of relatedness. All of this hampers the *internalization* of egalitarian norms, which, as Devine and Plant (1998) showed, would help reduce implicit bias. In order to increase encourage egalitarian norms to be internally motivated, people need to be in environments where they can feel autonomous, competent, and related to others. This is far more easily achieved in a climate where appraisal-based responses are avoided, that is, where blaming, shaming, and resentment are avoided. Non-appraising responses, by contrast, serve to enhance feelings of autonomy and competence by pointing to ways in which agents can actively make amends for their mistakes and improve their efforts to attain egalitarian ideals.

An environment where appraisal-based responses are common can also lead to more subtle and unexpected harms. For example, it appears that widespread fear or aversion to being perceived (or perceiving oneself) as racist can cause teachers to be less effective. Harber, Stafford & Kennedy (2010), for instance, demonstrate the existence of a "positive feedback bias", in which teachers whose commitments to egalitarianism are challenged subsequently grade the work of minority students less harshly than that of White students. (The effect does not occur when their egalitarianism is affirmed.) Such skewed grading, however, has negative consequences. When perceived by minority students, overly positive feedback can serve to erode their trust in the

[49] This does not mean that we should conceal the existence of implicit biases, but that education and training should always include, for example, constructive strategies for mitigating their effects. Such education would also do well that emphasize that *everyone* harbors some such set of biases and hence faces the same problem, so to be informed of one's bias is not to be singled out as a "bad person".

mainstream educational system, alienate or under-challenge them, and deprive them of useful feedback. But Croft and Schmader (2012) found that the feedback bias was found primarily among externally motivated egalitarian instructors who were "concerned about appearing racist," and Harber et al's (2012) most recent study found that, interestingly enough, positive feedback bias toward Black students decreased when teachers were in a more socially supportive environment, where measures of social support included the degree to which they felt their colleagues were friendly and supportive. Moreover, Norton et al's (2010) research on "racial paralysis" found that Whites' concerns about appearing racially biased led them to make significant efforts to avoid making cross-race judgments, while Plant and Devine's (2003) and Plant and Butz's (2006) studies on interracial anxiety showed that the fear of appearing biased led White people to avoid interracial interactions—all of which can only serve to strengthen implicit bias. Feeling socially supported and engaging in intergroup relationships requires a trusting climate in which one's motives and character are not under suspicion, and this can almost certainly be better achieved by eschewing appraisal-based responses in favor of non-appraising ones.

This is where I part company with Jules Holroyd (2012), who argues against such pragmatic reasons by claiming that taking ourselves to be liable to blame for being influenced by implicit biases may have beneficial effects even if no one is in an epistemic position to actually engage in blaming. Holroyd writes:

> Classifying certain actions as prohibited, for which individuals are liable to blame, can have numerous important effects, including: strengthening norms against so acting; encouraging individuals to self-monitor; and leading us to change our expectations of the steps others might take in monitoring their own behavior…[N]ote that [these effects] do not depend upon us being able to in fact engage in blaming, although some of them might encourage us to challenge others' decisions and provide careful justification for them. (2012, p. 300)

The whole problem with implicit biases, however, is that *being influenced by an implicit bias* is not the sort of "action" or "behavior" that it is useful to prohibit. Discriminatory behavior itself—the thing that can be called an action or a way of acting—has long been prohibited. Strengthened norms against being unconsciously influenced by implicit bias, if it even makes sense for agents to follow norms that by definition they cannot consciously try to do, is more likely to have the effect of threatening people's conceptions of themselves as committed to egalitarianism. Self-monitoring and monitoring of others, on the other hand, can be required and expected as part of

holding people accountable, without inducing the further requirement that they feel blameworthy. Again, feeling liable to these appraisal responses is probably counterproductive.

Before concluding, I should emphasize that mine is not a purely consequentialist account of moral responsibility, on which appraisal-based responses are justified or unjustified in virtue of good or bad consequences. Recall my two-pronged argument that internally motivated egalitarians often do not, for moral reasons, deserve appraisal-based responses and externally motivated egalitarians should not, for pragmatic reasons, be confronted with them. Desert still determines the aptness of appraisal-based responses, but the formidable epistemic barriers to establishing it means that appraisal-based responses will often be morally unjustified.[50] (Thus desert is a necessary but not sufficient criterion on my view. A person must be attributively responsible, and we must be able to confidently know that, in order for appraisal-based responses to be justified.) In other cases, appraisal-based responses may be truly warranted, but the moral reasons for them are outweighed by pragmatic considerations. I should also emphasize that accountability goes beyond merely procuring good consequences: the shape of a given society's social arrangements is also a representation of its members' relations to one another, and is thus capable of expressing (or failing to express) mutual respect and equal concern. Systems of accountability do not themselves rest on purely consequentialist grounds of justification.

I should also caution that my recommended solution should be understood as a general policy, one that can justifiably be overridden. The appropriateness of various practices of critical moral response is always sensitive to context, especially the contexts of particular relationships, and the nature of these relationships may license deviations from my main thesis.[51] For example, in the case of *self*-relation, it might be appropriate for me to feel guilt or self-blame. For one thing, my epistemic situation with regard to myself is different from that with regard to others, and I may well be in a position to know that I have not done all that could be reasonably expected of

[50] To mention just one more such barrier, it may be difficult to discern the difference between, say, "unconscious racism" and implicit racial bias. I take it that, if such a distinction can be made tenable, the former would ground attributability even while the latter might not. The difference would lie in the robustness and sophistication of the subpersonal machinery: unlike unconscious racism, racial implicit bias as measured by the Implicit Association Test, for instance, might be a much thinner cognitive association (as, say, in the case of an association between "salt" and "pepper") that is not hooked up to other elements of a person's psychological economy. I am grateful to Janine Jones, Megan Mitchell, and Chandra Sripada for discussions on this point.

[51] For this reason it might be more apt to conceive of the aptness of critical moral responses as relative, rather than absolute: for example, a person P is blameworthy by some set of others S in situation C (or in the context of relationship R or with social identity I or in background conditions B), not blameworthy *tout court*. I do not have the space to defend such a view here, but see Springer (2013) and Calhoun (1989).

me to combat bias.[52] Similarly, a close friend with whom I have had a long history might be in a position to chastise me for slipping up yet again. In this case, however, it may be that I have violated the second condition of behaving otherwise responsibly by not taking due care with an implicit bias I know myself to have, and my friend has good enough epistemic grounds on which to thus attribute the biased behavior to me. Or, it may be that her chastisement—while it has the appearance of an appraisal-based response—does not express the genuine negative judgments that are characteristic of blame; here it would function more as a mechanism of accountability, a way of indicating where I have fallen short and reminding me to be more cautious in the future. All of these possibilities are consistent with my view.

Finally, where there are only pragmatic reasons to refrain from appraisal-based responses, as in the case of externally-motivated egalitarians, these may be trumped by other, stronger reasons for engaging in them. For example, if there are cases in which it is critical for a person's being able to heal from the trauma of being subjected to biased behavior that she name it for what it is, then blame may be appropriate. Even though confrontation is generally counterproductive and should be avoided—which would be easier if there were well-established, effective structures of accountability!—we cannot demand of the victims of bias that they should, in addition to suffering the harms of biased behavior, bear the further burden of absorbing such damage to their self-respect.[53] It is also possible, though the calculation of risks is an empirical matter that will be extremely difficult to ascertain in particular cases, that there are potential long-term effects that could outweigh negative effects in the short term. One hopes, for instance, that a person who gets called out on her (unconscious) racism might eventually come to understand her partly in virtue of having undergone that experience. As I mentioned, however, this is risky and it seems on balance better pragmatically to avoid it as a strategy. But the existence of this possibility means that I certainly do not deny the need for there to be a wide range of approaches "out there" in the world, including both appraisal-based and non-appraising

[52] Smith (2004, p. 347) has made the pragmatic argument that viewing unendorsed implicit attitudes as non-attributable to oneself may turn out to be a self-fulfilling prophecy, in that it can blind, de-motivate or otherwise prevent a person from trying to change those parts of herself. To the extent that this is true, it would be a reason to allow appraisal-based responses toward oneself. However, Young (2011) argues just the opposite, that feeling blameworthy represents a "self-indulgence" that is "unproductive" because it focuses attention on oneself rather than what needs to change in the world. Thus the appropriateness of blaming oneself will probably need to be assessed on a case-by-case basis. It is also worth noting that we might be able to avoid the paralyzing effect Smith is concerned about by holding ourselves *accountable* for implicit biases, where accountability consists precisely in undertaking efforts to change. We could undertake these efforts without feeling the self-blame licensed by viewing our implicit biases as genuinely attributable to ourselves.

[53] I am indebted to Kristie Dotson for this point.

responses, in order to address all levels of readiness. A person who is initially best served by a non-appraising, largely educational response may develop to a point where she may be receptive to certain kinds of appraisal-based response, but this will usually require supportive, low-threat conditions.

5. Conclusion

For pragmatic reasons alone, then, I might have been able to argue for the view that we ought to hold people accountable but not attributively responsible for implicit bias. But I don't want to lose sight of the moral reasons for avoiding appraisal-based responses. In my view, pragmatic reasons are more often than not tied up with moral reasons; psychological recalcitrance tends to indicate failures to recognize important aspects of what it is like to be and to conceive of oneself as a fully-developed moral agent worthy of respect, or the lack of relationships of moral community and mutual respect. Autonomy, competence, and relatedness are important for internalizing and realizing social norms because these are precisely the conditions that respect people's experiences of themselves as efficacious moral agents responding to reasons. As we know, implicit biases act on us in ways that undermine the exercise of our rational capacities for self-reflective awareness and deliberative choice and endorsement—the things that make us it possible for us to be morally responsible beings at all. They can be utterly invisible to us. Thus, in respecting people as morally responsible agents, we should pay attention to what it is like for an agent trying to act rightly within the limits of what is visible from her practical point of view. But, as Angela Smith (2008) has argued, being held responsible is not only a burden, but a privilege: an expression of respect for their status as moral agents. This is why it remains vitally important that we hold people accountable for implicit bias: that we view them as agents whose actions express their social relationships to us. After all, what makes it necessary for us to hold each other responsible in the first place are the *social* needs based on relationships within the moral community and what is required to fashion and uphold acceptable forms of those relationships. The trick is to figure out how to be sensitive to the complexities of human action and the many ways human agents fall short—all the while still holding them accountable when they do.

References

Alicke, M.D., Buckingham, J., Zell, E., & Davis, T. (2008). "Culpable Control and Counterfactual Reasoning in the Psychology of Blame." *Personality and Social Psychology Bulletin* 34: 1371-1381.

Arpaly, Nomy. Unprincipled Virtue: An Inquiry Into Moral Agency. New York: Oxford University Press, 2004.

Baier, Kurt (1987). "Moral and Legal Responsibility." In M. Siegler, S. Toulmin, F. Zimring, & K. Schaffner, Medical Innovation and Bad Outcomes: Legal, Social, and Ethical Outcomes (pp. 101-130). Ann Arbor, MI: Health Administration Press.

Baron, A.S. and Banaji, M.R. (2006). "The Development of Implicit Biases: Evidence of Race Evaluations From Ages 6 and 10 and Adulthood." *Psychological Science* 17(1):53-58.

Baron, Marcia. (1987). "Kantian Ethics and Supererogation." *The Journal of Philosophy* 84(5):237-262.

Bartlett, Katharine T. (2009). "Making Good on Good Intentions: The Critical Role of Motivation in Reducing Implicit Workplace Discrimination." *Virginia Law Review* 95: 1893-1972.

Blum, Laurence. (2004). "Stereotypes and Stereotyping: A Moral Analysis." *Philosophical Papers* 33(3): 251-289.

Calhoun, Cheshire. (1989). "Responsibility and Reproach." *Ethics* 99(2): 389-406.

Croft, A., & Schmader, T. (2012). "The Feedback Withholding Bias: Minority Students Do Not Receive Critical Feedback from Evaluators Concerned About Appearing Racist." *Journal of Experimental Social Psychology* 48(5): 1139-1144.

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

Dasgupta, N. and Greenwald, A.S. (2001). "On the Malleability of Automatic Biases: Combating Automatic Prejudice with Images of Admired and Disliked Individuals." *Journal of Personality and Social Psychology* 81(5): 800-814.

Deci, E. L., & Ryan, R. M. (2000). "The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-determination of Behavior." *Psychological Inquiry* 11: 227-268.

Devine, P.G. and Plant, E.A. (1998). "Internal and External Motivation to Respond Without Prejudice." *Journal of Personality and Social Psychology* 75(3): 811-832.

Frankfurt, Harry. The Importance of What We Care about: Philosophical Essays. Cambridge, MA: Cambridge University Press, 1988.

Fischer, J.M. & Tognazzini, N.A. (2011). "The Physiognomy of Responsibility." *Philosophy and Phenomenological Research* 82(2): 381-417.

Fricker, Miranda. Epistemic Injustice. Oxford: Oxford University Press, 2007.

Gawronski, B., Hofmann, W., and Wilbur, C.J. (2006). "Are 'Implicit' Attitudes Unconscious?" *Consciousness and Cognition* 15: 485-499.

Goodin, Robert. Utilitarianism as a Public Philosophy. Cambridge: Cambridge University Press, 1995.

Greenwald, A. T., Poehlman, T. A., Uhlmann, E., & Banaji, M. (2009). "Understanding and Using the Implicit Association Test III: Meta-Analysis of Predictive Validity." *Journal of Personality and Social Psychology* 97(1): 17-41.

Harber, K.D., Stafford, R., and Kennedy, K.A. (2010) "The Positive Feedback Bias as a Response to Self-image Threat." *British Journal of Social Psychology* 49: 207-218.

Harber, K.D., Gorman, J.L., Gengaro, F.P., Butisingh, S., Tsang, W., & Ouellette, R. (2012). "Students' Race and Teachers' Social Support Affect the Positive Feedback Bias in Public Schools." *Journal of Educational Psychology* 104(4): 1149-1161.

Hieronymi, Pamela. (2004). "The Force and Fairness of Blame." *Philosophical Perspectives* 18(1): 115-148.

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

Kelly, Erin. (2002). "Doing Without Desert." *Pacific Philosophical Quarterly* 83(2): 180-205.

Korsgaard, Christine. Self-Constitution: Agency, Identity, and Integrity. New York: Oxford University Press, 2009.

Legault, L., Gutsell, J.N., & Inzlicht, M. (2011). "Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (but Also Increase) Prejudice." *Psychological Science* 22(12): 1472–1477.

Malle, B.F., Guglielmo, S. & Monroe, A.E. (2014). "A Theory of Blame." *Psychological Inquiry* 25(2): 147-186.

Norton, M.I., Mason, M.F., Vandello, J.A., Biga, A., & Dyer, R. (2013). "An fMRI Investigation of Racial Paralysis." *Social Cognitive and Affective Neuroscience* 8: 387-393.

Olson, M.A. & Fazio, R.H. (2006). "Reducing Automatically Activated Racial Prejudice Through Implicit Evaluative Conditioning." *Personality and Social Psychology Bulletin* 32(4): 421-433.

Oshana, Marina. (1997). "Ascriptions of Responsibility." *American Philosophical Quarterly* 34(1): 71-83.

Pearson, A.R., Dovidio, J.F., & Gaertner, S.L. (2009). "The Nature of Contemporary Prejudice: Insights from Aversive Racism." *Social and Personality Psychology Compass* 3: 1-25.

Scanlon, T.M. What We Owe Each Other. Cambridge, MA: Belknap Press, 1998.

Plant, E. A. & Devine, P.G. (2001). "Responses to Other-Imposed Pro-Black Pressure: Acceptance or Backlash?" *Journal of Experimental Social Psychology* 37: 486-501.

Rowe, Mary. "Micro-affirmations and Micro-inequities." (2008). *Journal of the International Ombudsman Association* 1(1): 45-48.

Saul, Jennifer. "Implicit Bias, Stereotype Threat, and Women in Philosophy," in Women in Philosophy: What Needs to Change?, ed. Fiona Jenkins and Katrina Hutchison. Oxford: Oxford University Press, 2013.

Shoemaker, David. (2011). "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121(3): 602-632.

Smiley, Marion. Moral Responsibility and the Boundaries of Community. Chicago: University of Chicago Press, 1992.

Smith, Angela. (2012). "Attributability, Answerability, and Accountability: In Defense of a Unified Account." *Ethics* 122(3): 575-589.

Smith, Angela. (2007). "On Being Responsible and Holding Responsible." *Journal of Ethics* 11(4): 465-484.

Smith, Angela. (2004). "Conflicting Attitudes, Moral Agency, and Conceptions of the Self," *Philosophical Topics* 32(1/2): 331-352.

Springer, Elise. Communicating Moral Concern. Cambridge, MA: MIT Press, 2013.

Strawson, P.F. (1962). "Freedom and Resentment." *Proceedings of the British Academy* 48:1-25.

Velleman, David. "What Happens When Someone Acts?" Mind 101 (1992): 461-481.

Wallace, R. Jay. Responsibility and the Moral Sentiments. Cambridge, MA: Harvard University Press, 2004.

Watson, Gary. Agency and Answerability: Selected Essays. New York: Oxford University Press, 2004.

Young, Iris Marion. Responsibility for Justice. New York: Oxford University Press, 2011.

CHAPTER 3

Expanding the Moral Repertoire: Oughts, Ideals, and Appraisals

1. Introduction

In the last chapter, I argued that we should take seriously the distinction between attributability and accountability in ascribing responsibility for actions caused by implicit bias. I proposed two conditions for excuse that would need to be satisfied in order for an agent to lack attributability for such actions: first, the agent would not endorse the influence of the bias if she were aware of it, and second, the agent must have done what she could reasonably be expected to do with respect to avoiding and responding to the implicit bias. Under these two conditions, actions caused by implicit bias are not properly *attributable* to the agent as expressions of her agency, and she thus does not deserve appraisal-based responses like blame and punishment. However, she remains *accountable* for those actions, which entails that she is appropriately subject to non-appraising responses like penalty and contributing to compensation for victims. In this chapter, I take a closer look at the distinction between appraisal-based and non-appraising responses. My main task will be to justify the existence of such a distinction, which I defend by appeal to both moral-theoretical and psychological reasons. However, I also offer a preliminary characterization of the kind of interpersonal non-appraising responses I suggest we will need in order to properly respond to actions caused by implicit bias, as well as eliminating implicit bias in our selves and our social structures. In effect, I am addressing two different questions and killing them with one stone: What can we reasonably be expected to do to mitigate bias? and, What sorts of non-appraising responses are appropriate for particular failures to do so? I develop answers to both of these questions using the concept of imperfect duties, which I interpret in terms of role-based ideals.

What this involves is an alternative "derivation" of the distinction between appraisal-based and non-appraising responses. The moral-theoretical story I tell involves two distinct domains of morality, which I call the "Ought" realm and the "Ideal" realm. Philosophical inquiry has focused on the appraisal-based responses that are warranted by violations of "Ought" standards, while overlooking the question of what types of responses that are warranted by particular occasions of failing to pursue "Ideal" ends—a gap that, I propose, is filled by the category of non-appraising responses. The psychological story involves two distinct systems of moral regulation, which can most basically be characterized in terms of "avoidance" and "approach". Moral criticism, I argue, should thus come in two different varieties—appraisal-based and non-appraising—for engaging these two different systems. By locating and characterizing non-appraising responses, I claim, we can deploy them in a variety of moral situations relevant for dealing with the problem of implicit bias. This analysis dovetails with my explication of the criterion of reasonable expectation I mentioned earlier. What we can reasonably be expected to do, on my view, is what is required by our imperfect duties, and those imperfect duties are given content by role-based "Ideal" ends.

2. Beyond Blame and Resentment

Philosophical preoccupation with blame and resentment belies the fact that our ordinary vocabulary is packed full of words for describing different forms of moral criticism. In addition to blaming a person, we can reproach, reprimand, reprove, and so on. And in addition to feeling resentment toward a person who has done wrong, we might feel indignation, opprobrium, disappointment, etc. Given the similarity and relative interchangeability of use among many of these terms, it is worth taking a brief look at their etymological origins[54]. Note that what I am interested in here is *not* the terms themselves or their actual definitions, but their etymological *origins*.

[54] This represents a modified list of the "commonly used verbs denoting moral criticism" used by Voiklis, Cusimano, & Malle (2014). I excluded idiomatic phrases ("tell off," "lash out at," "let X have it", "point the finger at," and "chew out") as well as those verbs that seemed to go beyond ordinary moral criticism ("slander," "revile," and "vilify"). I also added two presumably less common synonyms ("reprehend" and "remonstrate." Voiklis, Cusimano, & Malle (2014) generated their list by consulting synonyms corrobated by Roget's Thesaurus and the WordNet database, while filtering out words with too many unrelated meanings or outside current and common usage through the Oxford English Dictionary and the Corpus of Contemporary American English. All etymologies are taken from the Online Etymology Dictionary, which is compiled from the list of sources available at http://www.etymonline.com/sources.php.

Accuse	"to give as a cause or motive"	Admonish	"to put into mind, remind, advise"
Berate	"to thoroughly count over, reckon, consider"	Attack	"to join (battle)"
Blame	"to blaspheme," "to speak evil of"	Chide	"to contend, quarrel, complain"
Castigate/chastise	"to purify"	Object	"to put or throw before or against"
Censure	"to assess, appraise, judge"	Reproach	"to bring back close" (cf. to get up in someone's face)
Condemn	"to sentence, to find guilty, to doom	Reprimand	"to repress, push back"
Criticize	"to pass judgment on"	Rebuke	"to beat back"
Denounce	"to proclaim (downward)"	Reprehend	"to pull or hold back"
Disapprove/ reprove	"to not find worthy"	Remonstrate	"to make plain/show reasons against"
Fault	"to find a lack, flaw, deficiency"	Scold	"to be quarrelsome"
Resentment	"to feel deeply," "to feel aggrieved by"	Disappointment	"to deprive of an appointed office"
Indignation	"to regard as unworthy"	Dissatisfaction	"to not do enough"
Disapprobation	"to not find worthy"	Opprobrium	"to be against a disgraceful act"

Table 1: Etymologies of Words Denoting Moral Criticism

Although all of these terms clearly express negative moral criticism, there are, broadly speaking, two discernible groups of meanings. Among the blame-related words, there are some ("to proclaim," "to assess, appraise, judge," "to sentence, to find guilty, to doom," "to not find worthy") which centrally involve some form of *evaluation* or *judgment* of worth. By contrast, others ("to bring back," "to repress, push back," "to beat back," "to pull or hold back," "to make plain or show the reasons against," "to remind, advise") involve various forms of *communication* and *action* intended to prevent the (re-)occurrence of some (wrongful) act. In parallel fashion, some of the "resentment"-related words ("to regard as unworthy") also involve measures of worth, while others ("to deprive of an appointed office", "to be against a disgraceful act) target more specific *acts* or failures to meet some (role-based) goal.

This supports the familiar notion that blame essentially involves some *appraisal of the agent on the basis of her action*. As Coates and Tognazzini (2012) write in the introduction of their anthology on Blame: Its Nature and Norms:

> There is a rather wide array of judgments [constitutive of blame] one could appeal to here, but we can borrow a phrase from Gary Watson and categorize them all as judgments about "the quality of the other's moral self as exemplified in action and attitude" (2004, p. 226). Jonathan Glover, Ishtiyaque Haji, and Michael Zimmerman, for example, all seem to view blame as though it is a type of "moral accounting" (Glover 1970, p. 64). When we blame someone, we judge "that there is a 'discredit' or 'debit' in his ledger, a 'negative mark' in his 'report card,' or a 'blemish' or 'stain' on his 'record'; that his 'record' has been 'tarnished'; that his 'moral standing' has been 'diminished'" (Zimmerman 1988, p. 38)." (8)

The centrality of appraisal to blame is also apparent in contemporary social psychological work on blame. Malle, Guglielmo, & Monroe (2014), for instance, draw a distinction between three types of moral judgments: Type 1 judgments are concerned with "setting and affirming *norms*," Type 2 judgments with "evaluating *events* (i.e. outcomes, behaviors) in light of those norms," and Type 3 judgments with "evaluating *agents* for their involvement in such norm-relevant events" (148). Blame, according to them, is "the paradigmatic Type 3 judgment" (148). They go on to distinguish blame from judgments of wrongness ("whereas blame judgments target an agent, wrongness judgments target a behavior"), as well as from anger ("unlike blame, anger can be directed at or caused by impersonal events") (149-150). Their Path Model of Blame proposes a bifurcated path of information processing according to which people evaluate, in the following order: 1) whether an event violated a norm, 2) whether that event was caused by an agent, 3) whether the agent did it intentionally, and then either 4a) what the agent's reasons were for so acting, if she did intentionally or 4b) whether the agent had the obligation and capacity to prevent the event, if she did it unintentionally. Intentionality plays a crucial role here because it is a sufficient, though not necessary, marker of agency. Once agency has been established, the model incorporates an *evaluation* of quality of the agent at the stage of considering the agent's reasons for action. The authors write:

> Considering an agent's reasons is an intrinsic part of the moral perception of intentional actions because these reasons determine the *meaning* of the action (Binder, 2000; Scanlon, 2008)—*what the action reveals about the agent's motives, beliefs, and attitudes* (Malle, 2004; Stueber, 2009). (154) (second emphasis mine)

Motives, beliefs, and attitudes, of course, are precisely what make up an agent's practical identity and hence make her a proper object of moral appraisal. The Path Model of Blame thus provides an empirically-supported view that blame centrally involves the appraisal of agents.

Other philosophers, however have pointed out that blame involves far more than a mere (cognitive) judgment. I do not intend to deny that blame also has communicative and expressive functions, that it can function to induce remorse and affirm moral norms, that it has the effect of modifying relationships, or that it functions as a negative sanction, as has been variously proposed.[55] Indeed, I follow Cheshire Calhoun (1989) in thinking that practices of moral responsibility can have multiple functions. Calhoun (1989) writes:

> Assigning responsibility licenses reproachful or approving responses: anger, admiration, chastisement, praise, seeking out, and snubbing. Moral reproach reminds or perhaps teaches us what actions are morally unacceptable. So the first point is educational. The second point is motivational. Moral reproach motivates us to change the way we act. The third is conceptual. Reproachful labels—for example, "oppressor," "exploiter," "sexist,"—confirm our identities as moral agents. (405)

I agree that blame can do all these things. But I suspect that it derives its peculiar "force," as Pamela Hieronymi (2008) calls it, from its distinctive appraisal function. Hieronymi argues that blame has its characteristic force not only because it is used as a negative social sanction, but because, in and of itself, it is a judgment that "gains force from the importance of its content and the importance of the opinions of others on that topic" (122). It is, she says, "a particularly deep form of grading or a particularly important sort of record-keeping" (125). What Hieronymi is highlighting is that it *matters* to deeply social and moral creatures such as we are when we are judged to have ill will, and when we thus feel that we do not stand in "relations of recognition of mutual regard" (145). Notice, however, that this would appear to depend on the degree to which a person cares about her relationships with others, and the degree to which she genuinely does have good intentions. A person might *not* care very much about standing in relations of recognition of mutual regard with another if she simply does not hold the other in high regard at all, or if she really does intend to demean or harm them. (Indeed, the problem with bias and prejudice is precisely that people do *not* take some group of others to be worthy of mutual regard.) But, I claim, people rarely fail to care about evaluations of themselves, or about being wrong! In my view, even in these cases where mutual regard or good will is missing, being blamed may still

[55] See, respectively, Holroyd (2007), Bennett (2013), Fricker (2013), Scanlon (2008), and Smart (1961). For a review, see Tognazzini and Coates (2014).

have a "sting" to it insofar as it represents a negative evaluation. The person being blamed may still feel indignation, defensiveness, or hostility insofar as she feels that the negative evaluation is unwarranted—if, for example, she feels that her victims deserve the demeaning treatment she gave them (so that it is wrong to negatively evaluate her for doing so). So it is the *appraisal-based* nature of blame, I suggest, that distinguishes it from more ordinary forms of disapproval that can more easily be shrugged off. Blame, resentment, and punishment are thus what I call "appraisal-based responses," because they reflect assessments of a person's quality as an agent.

However, as the second group of meanings suggest, it is possible to set aside the appraisal function of blame and resentment while retaining their potential for communication, i.e. for reminding, warning, and making reasons plain, as well as for exhortation, i.e. attempts to push, pull, and block potential wrongdoing. I will use the term "non-appraising responses" to designate such responses that seek to persuade and engage, rather than appraise. I thus embrace Elise Springer's (2013) theory moral criticism as having the primary goal of *communicating a moral concern*:

> Reaching a hearer's attention with a concern means pressing one's agency just far enough in the other's direction to spark her own engagement with moral concerns without overriding the distinct perceptual and perspectival qualities introduced by her own agency. (156-157)

In short, I am seeking forms of moral criticism that carry out Calhoun's first two purposes—education and motivation—while leaving aside the conceptual labeling that denominates an agent's moral worth. While recognizing the importance of appraisal-based responses in our moral lives, we can still search for other responses that enhance our moral repertoire, preparing us for a wider range of moral situations. Where should we go to find these non-appraising responses?

3. Two Moral Domains: "Ought" and "Ideal"

Let me begin by taking an aerial view of the moral landscape in order to point at a divide between two different domains of morality, which I call the "Ought" realm and the "Ideal" realm. This divide has been noticed by a number of different philosophers. Peter Railton (2010), for instance, has distinguished between two types of normative concepts: regulatives and evaluatives. Regulatives ("right" vs. "wrong," "correct" vs. "incorrect," "obligatory," "permissible," "impermissible") are binary concepts that invoke a rule or norm—in other words, a standard—by which something either meets or does not meet that standard. Indeed, this is

precisely their function: "What distinguishes regulatives in the first instance is that they involve laying down or appealing to a rule or standard relative to which we can make an assessment of an act, thought, practice, etc., as fitting or failing to fit, often (typically, perhaps) in a formal sense" (Railton, ms). According to Railton, this function leads to three other distinctive features: failure to conform to the standard demands revision in the thing that fails to conform, and not the standard itself, conforming to the standard must be voluntarily attainable (i.e. "Ought implies can"), and failure to conform must be reliably detectable. Without these features these concepts would not be able to effectively perform their regulative function. It would be useless for a person to try to regulate her behavior using a standard if there was no way for her to know whether she was meeting that standard or not, or if meeting the standard was up to factors completely beyond her control. By contrast, evaluatives ("good" vs."bad," "virtuous" vs. "vicious") admit of degrees, and they can guide us even though they may not be voluntarily attainable or clearly articulated. Their function is not regulative, but evaluative: "Evaluatives go into the formulation and characterization of goals or aims, so that when we are deliberating about which goals to pursue, we ask what is good, or bad, or fine, or credible" (Railton, ms). Schroeder (2009) makes what appears to be the distinction between "deontic" and "evaluative" concepts, which he also refers to as the "Right" and the "Good". Similarly, Bernard Gert's (2007) theory of common morality appeals to both "moral rules" and "moral ideals." Moral rules are requirements and prohibitions on conduct that set *constraints* on the pursuit of all ends, violations of which are liable to punishment. Moral ideals, by contrast, function to applaud and deplore various forms of conduct by prescribing *ends*, but failure to perform them does not make a person liable to punishment (Gert 2002). These considerations bring us to the more familiar Kantian notion of perfect and imperfect duties: duties that require us to act according to certain ends, but that do not specify with precision when and what acts are required. I will say more about the concept of imperfect duties in Section 4, but for now let us merely note that perfect duties are standardly thought to include requirements such as not lying, not stealing, not killing, and so on, while imperfect duties include ends such as helping others and improving your talents.

It is clear that violations of perfect duties warrant blame, resentment, and punishment (if these are ever warranted). But what about failures to take up a *particular* opportunity for carrying out an *imperfect* duty? One of the defining features of imperfect duties is that a person does not violate the duty merely by failing on one occasion to perform some act, but only across a whole

series or pattern of actions. Since we are creatures limited by finite time and resources, we are allowed (some) discretion to choose when and where to carry out our imperfect duties. It is thus generally believed that people are not blameworthy for failing on any one particular occasion to donate money to charity, to develop their talents, or whatever the case may be. According to Kant (emphasis mine): "Fulfillment of [imperfect duties] is merit, but failure to fulfill them is not in itself culpability but rather mere deficiency in moral worth, *unless the subject should make it his principle not to comply* with such duties" (390). Of course, a long-standing pattern of failures *does* constitute a violation of imperfect duty: it is grounds for attributing a blameworthy lack of concern, or failure to adopt a morally required end. I will refer to such patterns of actions and omissions as "violations of imperfect duty." Conversely, when a person maintains a sufficiently robust overall pattern of bias-mitigating behavior, we are not justified in attributing any particular failure to perform a bias-mitigating action on a particular occasion to a lack of commitment to mitigating bias. I will refer to such *particular* actions and omissions—where we may not know its relation to larger patterns of action—as "failures of imperfect duty," which are to be distinguished from violations of imperfect duty.

The crucial question here is this: *what is the appropriate critical moral response to failures of imperfect duty?* This is an important gap that has been largely overlooked in discussions in blame, resentment, and responsibility. Once we set aside appraisal function of critical moral response in favor of education and motivation—that is, once we shift our focus from whether people are good or bad agents to what they can reasonably be expected to do to reshape our moral world for the better[56]—we can see that such critical moral responses are always warranted, for there is always more work to be done. Even more importantly for my purposes, it also explains why there should also be a divide between two families of moral criticism. Violations of "Ought" standards are expressions of agency because these standards are (for the most part[57]) voluntarily attainable. It is comparatively easy not to kill or harm people, not to steal from, lie to, or cheat them, so an agent's doing so constitutes her actively flouting a specific constraint that supports a judgment of blameworthiness. But from a single token of a failure to carry out an imperfect duty, we cannot similarly infer that the agent is lacking the appropriate ideal or end, because she has some

[56] I have tried to express this in a way that demonstrates how these questions are distinct, but related: they are linked through the criterion of reasonable expectation.

[57] This is true, at least, if we consider simple cases. If we include structural harms produced in a globalized, causally complex world, it may not be so easy to avoid violations of perfect duty. For now, however, I bracket this issue.

discretion in how and on what occasions she works toward that end. Ideals are *ideal* precisely because they are never wholly attainable, so failure to "conform" to an ideal does not in itself reflect much about the agent. Relatedly, it is not possible to alter, by sheer force of will in the moment, one's historical patterns or habits of action. "Ought implies can" in the "Ought" realm thus becomes "Ought implies can-strive-toward" in the "Ideal" realm; and this, I claim, provides us deep *moral-theoretical* grounds for thinking that we need an additional category of non-appraising responses. This is where non-appraising responses are naturally found naturally "in the wild," as it were.

4. Two Motivational Systems: Approach and Avoidance

Interestingly enough, it appears that there also are psychological reasons for drawing a distinction between appraisal-based and non-appraising responses. A line of research in attribution theory has shown that violations of perfect duties lead people to make stronger trait attributions than violations of imperfect duties (Trafimow & Trafimow 1999, Trafimow & Ishikawa 2012). For example, people needed *more* instances of uncharitable behavior to infer that a person was uncharitable, than they needed in order to infer that a person was dishonest; moreover, people adjusted their judgment of how charitable a person is by taking into account situational features, but they did not take these into account for judgments of dishonesty. This is preliminary evidence that, as a matter of fact, people's moral responses to violations of perfect duties are more likely to involve appraisals of the agent than failures of imperfect duties. In a different context, the difference between appraisals of person and action is reflected in research by Kamins and Dweck (1999), whose studies on children's responses to failure has shown that "person or trait-oriented feedback" (global evaluations of ability and worthiness) subsequently led to decreased expectations and poorer performance in the face of failure than did "process-oriented feedback" (specific evaluations of strategies and effort). In yet another context, the prevailing empirically-based theory of the difference between shame and guilt is that "shame involves a negative evaluation of the global self; guilt involves a negative evaluation of a specific behavior" (Tangney, Stuewig & Mashek 2007, p.349). As Tangney, Stuewig, & Mashek (2007) put it: "[E]mpirical research supports that this differential emphasis on self (I did that horrible thing") versus behavior ("I did that horrible thing") sets the stage for differentially adaptive patterns of motivations and subsequent behavior" (349). This convergence of similar results from

a variety of different social psychological phenomena suggests that appraisal-based and non-appraising responses tap into very distinct patterns of cognition, affect, and motivation that lead to different behavioral consequences.

These last observations lead us into the psychological story that I propose to tell behind the existence of two different families of moral criticism. Consider a recent set of experiments by Does, Derks, & Ellemers (2011). Participants were given a bogus newspaper article about discrimination against non-native Dutch, with only a slight change in wording: one condition described "ideals concerning fairness and equal treatment" while the other described "obligations concerning fairness and equal treatment". Participants in the first condition reported higher agreement with items like "Cultural diversity is an asset to my organization," "I would treat nonnative Dutch fairly even if it means I personally have to take a step back," and "I think it's more important that my organization treats nonnative Dutch employees fairly and justly than that it performs well financially." Importantly, subjects in the moral ideals condition also scored lower on measures of social identity threat, as measured by items such as "I regret that I am a member of the group of native Dutch". In light of the two domains of morality I outlined earlier, this is a striking result: it suggests that appeals to the obligations and standards of "Ought" realm may be psychologically different consequences compared to appeals to the ends and goals of the "Ideal" realm.

Does & colleagues situate their work within a larger literature on motivation and self-regulation. Beginning in the early 20th century, a central theme in self-regulation research has been a difference between two motivational orientations: pursuing positive outcomes (e.g. ideals), and avoiding negative outcomes (e.g. violations of obligation). E. Tory Higgins's "self-discrepancy theory," first introduced in 1987, proposes that people assess their own achievements and failures relative to standards or goals known as "self-guides," and that they are motivated to eliminate the discrepancy between their actual selves and their self-guides. What is most distinctive about Higgins' theory is that the postulation of two distinct *kinds* of self-guides—the "Ought Self" and the "Ideal Self"—which give rise to two distinct patterns of emotional vulnerabilities and motivational dispositions.[58] The Ought Self, according to Higgins, is "your

[58] More precisely, there are actually two different "Ought selves" and two different "Ideal selves," depending on perspective. One's ought/own self is what a person herself believes she ought to be, while her ought/other self is what she believes others believe she ought to be. Similarly, one's ideal/own self is what a person herself wishes to be, while her ideal/other self is what she believes others wish her to be.

representation of the attributes that someone (yourself or another) believes you should or ought to possess (i.e. a representation of someone's sense of your duty, obligations, or responsibilities)," while the Ideal Self is "your representation of the attributes that someone (yourself or another) would like you, ideally, to possess (i.e. a representation of someone's hopes, aspirations, or wishes for you)" (320-321). Corresponding to these two "domains of the self" are the "two basic kinds of negative psychological situations": on the one hand, "the *absence of positive outcomes* (actual or expected), which is associated with dejection-related emotions (e.g. dissatisfaction, disappointment, sadness)" and on the other hand, "the *presence of negative outcomes* (actual or expected), which is associated with agitation-related emotions (e.g. fear, threat, edginess)" (322). Sure enough, Higgins (1987) found that subjects whose actual attributes did not match up to their Ought attributes reported feelings of anxiety, fear (including fear of punishment), irritation, and resentment, while those whose actual attributes did not match up to their Ideal attributes reported feelings of disappointment, dissatisfaction, embarrassment, and frustration. Importantly, self-discrepancy theory countenances both individual and situational variance in which kind of self-guide is most active: while people may be chronically more predisposed to Ought or Ideal Selves, they can also be primed to make an Ought or Ideal self more accessible in particular situations.

While certain elements of self-discrepancy have been challenged, with mixed results[59], it continues to be widely used to explain a variety of psychological phenomena including depression, anxiety, eating disorders, and procrastination. Moreover, its main tenets have since evolved into Higgins's (1997) "regulatory focus" theory, which has been taken up and developed across the past two and a half decades. A promotion focus leads to an *eager strategy*: attempts to maximize "hits" even if they lead to many "misses," whereas a prevention focus leads to a *vigilant strategy* of minimizing misses even if that means foregoing potential hits. Like self-guides, prevention and promotion foci can be situationally primed. Most recently, this has been applied to moral domain by Ronnie Janoff-Bulman and colleagues, who propose a dual-systems model of

[59] What remains under some contention is whether Ought self-discrepancies and Ideal self-discrepancies uniquely predict distinct patterns of emotion. Tangney, Niedenthal, Covert, & Barlow (1998) found that different self-discrepancies were not correlated with differences in negative emotions, as did Ozgul, Heubeck, War, & Wilkinson (2003). However, Boldero, Moretti, Bell, & Francis (2005) criticized these studies on methodological and interpretive grounds, and Hardin and Lakin (2009) found that different self-discrepancies *did* uniquely predict agitation- and depression-related emotions. Phillips and Silvia (2010) found that self-discrepancy theory was "broadly but incompletely supported," in that Ought and Ideal discrepancies were shown to be distinct constructs, and Ought discrepancies uniquely predicted anxiety-related emotion (but both kinds of discrepancies predicted depression-related emotion).

"moral regulation": a proscriptive system that tells us what we should *not* do, and a prescriptive system that tells us what we should do. Janoff-Bulman, Sheikh, & Hepp's (2009) evidence derives from developmental work on moral learning, priming effects, and linguistic coding, along with explicit judgments about morality. They cite work by Kochanska and colleagues showing that children throughout early childhood are more likely to comply with "don'ts" rather than "do's," and that fearfulness was associated with successfully learning "don'ts," but not "do's." Janoff-Bulman, Sheikh, & Hepp' (2009) priming experiment found that participants non-linguistically primed by solving a maze to escape from an owl (prevention focus) rather to get a piece of cheese (promotion focus) subsequently listed more proscriptions than prescriptions when asked to generate sentences describing what people should or should not do to be moral. In another study, they coded participants' answers to questions asking them to complete sentences that asked either "To be moral or not be immoral I should…" or "To be moral or not be immoral I should not…" They found that moral proscriptions were couched in more concrete language using specific action verbs, while moral prescriptions involved the use of more abstract language using state verbs and adjectives. This evidence thus points to the existence of distinct moral regulation systems. Furthermore, they found the following difference in participant's judgments about prescriptions and proscriptions:

> Participants felt it was equally important, overall, to engage in each—they indicated people should do "good things" and should not do "bad things" to the same extent—and yet they nevertheless accorded greater freedom to prescriptive behaviors. In this sense the latter appear to be a combination of oughts and ideals in Higgins's (1997, 1998) regulatory focus theory: *oughts in the sense of responsibilities regarded as shoulds, but nevertheless somewhat akin to ideals in the sense of behaviors to strive for rather than deemed mandatory* (see also, in the domain of moral philosophy, Gert, 2004, on moral ideals, Pincoffs, 1986, on nonmandatory virtues, and Heyd, 1982, e.g., on supererogation). (529) (emphasis mine)

These results, I claim, suggest that the proscriptive moral regulation system is primarily associated with the perfect duties of the "Ought" realm, while the prescriptive moral regulation system is primarily associated with the imperfect duties—"regarded as shoulds, but nevertheless somewhat akin to ideals"—of the Ideal realm.[60] I should note that this research is still in early stages, and that these studies only indicate rather than establish the existence of two moral regulations. Moreover, it is certainly possible that even though one system has been primarily

[60] While the authors provide another study to purporting to demonstrate that these differences are not due to the difference in perfect and imperfect duties, their usage of "imperfect duty" on that occasion does not perfectly correspond to the traditional philosophical sense of the term.

developed for one domain of morality, there may still be conditions under which it can be used for the other domain; it is also likely that the two systems often work in conjunction and are not cleanly separable. Nevertheless, if it is true that we indeed have two different motivational systems for moral self-regulation, then this furnishes additional *psychological* reasons for two different families of moral criticism that engage different motivational systems.

5. Implicit Bias and Imperfect Duties

Before I give my proposed characterization of non-appraising responses, let us turn back to the problem of implicit bias. We are now in a better position to take up the question of what we can reasonably expect people to do to avoid and respond to implicit bias. This marks a shift away from *bias manifested in actions* to *actions undertaken to mitigate bias*, that is, a shift from assessments of individual actions at a time to assessments of *patterns* of action over time (and in the next chapter, from individual actions to joint, coordinated actions of collectives). The central problem at hand is this: what can we reasonably be expected to do with respect to avoiding and responding to implicit bias? This question can be divided into two parts: First, what can we reasonably be expected to do about *our own personal biases* (and perhaps, those under our care or influence), where this could consist in preventing the acquisition of implicit bias, blocking the influence of such bias, or undertaking the modification or elimination of them? Second, what can we reasonably be expected to do to prevent the *general* production, maintenance, propagation, or reinforcement of implicit bias? This second question is what I will address in the next chapter. Here, my concern here is to explore the deeper issue underlying both parts of the question: How much of our time and resources should we spend on what is required by our imperfect duties? And how can we *know* just how much is required?

Ubiquitous though this question is throughout applied ethics, there is hardly a universally accepted answer. I will not be attempting to propose one either; however, I will say a few words *why* not, using the concept of imperfect duties—drawn from the Ideal realm—to conceptualize what we can reasonably be expected to do.[61] I propose that we have an "imperfect duty to mitigate bias," which I will use as a stand-in for all activities which we may be encouraged to

[61] An alternative way to proceed would be through virtue ethics, which has a long tradition of emphasizing longstanding dispositional traits that issue in stable patterns of action. I do not doubt that one could develop a cogent account of the moral and epistemic virtues involved in mitigating bias: compassion, sympathy, wisdom, (testimonial) justice, and humility, to name just a few. Since I take "virtues" to be another inhabitant of the "Ideal" realm, I believe that similar results could be obtained.

perform in efforts to avoid developing pernicious implicit associations, to weaken those associations, and to put mechanisms in place that block their operations.

It is difficult to pin down an exact definition of imperfect duties, and my aim is not to provide one. Instead, I will proceed by considering what the concept of imperfect duties does, *functionally*, for a normative moral theory, which in turn depends on the functions of moral theory itself. Let me distinguish two of them. The first way that imperfect duties figure into a moral theory is in answer to the question of what is morally obligatory, permissible, and impermissible. I will call this the Demarcation Question, because the goal is to determine a criterion for classifying some acts as having the status of "morally right" (or "required," "obligatory," "what we ought to do," etc.) and others not. Having an answer to the Demarcation Question is a prerequisite for answering questions about moral responsibility—at least, for responsibility as *attributability*. It is only once we know the moral status of your action that we are able to judge you as an agent positively or negatively in light of it. Contrast this with responsibility as accountability: you could be accountable for things even where there was no action (right or wrong) or agent (good or bad) at all, say, in the aftermath of a natural disaster. In this case, it is natural to think, members of the community would be morally obligated (even if not legally) to share some of the burdens of caring for victims, repair, etc. even though that damage and suffering in no way reflects on them as agents because it was not caused by an agent at all. These imperfect moral requirements are simply part of the demands and expectations we are subject to as members of a moral community. So the Demarcation Question is primarily relevant for ascriptions of attributability and not accountability.[62]

This is why the criterion of reasonable expectation figured into my excusing conditions for attributability for implicit bias: if an agent has not lived up to what is morally required in *general* to block implicit biases, in the rest of her life, then *that is good evidence* that she is not fully committed to treating people fairly and equitably as individuals. And that gives us grounds to blame her, to attribute a moral failing to her. We can think of these general duties as imperfect duties, since they do not pick out particular actions at a particular time but govern patterns of actions across time, e.g. making efforts to gain exposure to counterstereotypical exemplars, widening one's social circles to increase intergroup contact, and so on. (This is important for

[62] Some minimal knowledge is needed to assess what we are accountable for, e.g. if we are accountable for making amends to those who have been wronged, we will need to know what it is for someone to be wronged—but we do not need to know with great precision exactly why and how.

implicit bias, because implicit associations are learned and unlearned gradually over time.) Widening the scope of assessment to include patterns of action opens up the possibility that what a person does and has done *outside* of a situation has a bearing on whether and how what she does *in* that situation is properly attributable to her.[63] A pattern of repeated offenses without attempts at self-improvement betrays a lack of commitment to mitigating bias, as does a pattern of never or very rarely contributing to collective efforts to reform social structures that generate implicit bias. If a person fails to do *enough* of this, then we can blame her for actions caused by implicit bias, because we can infer that her biases reflect her insufficient commitment to treating people fairly and equitably. These, again, are violations of perfect duty, and in these cases we are *attributing* such patterns of omission to the agent as reflecting her lack of the right values, ends, commitments, or character traits; we are justified in engaging in appraisal-based responses because we are able to make evaluations of the agent's practical identity. But where exactly is the cut-off that would tell us that an agent lacks the appropriate end?

The short answer is: there isn't one. What we can reasonably be expected to do is highly variable and context-sensitive. Iris Marion Young (2011) identifies a number of parameters (which I will say more about in the next chapter) along which individuals differ in ways that affect their ability and hence responsibility for improvement: these include how much power they have to causally affect outcomes, how privileged they are to have time and resources they could devote. These do not vary linearly or systematically. The fact that there are multiple and irreducible axes of power and privilege—race, gender and gender identity, class, sexual orientation, and disability, to name just a few—means that different individuals will be differently placed and better or worse-equipped to work on some issues rather than others. Even if there were some way to find a general combinatorial principle that could correctly weight all these factors, it is difficult to see how we could easily apply and implement it. This means that we again have epistemic reasons not to engage in appraisal-based responses, because the difficulty of the moral casuistry involved will often be beyond us. Moreover, we often will not know the private and detailed circumstances of people's lives that change what we can reasonably expect them to

[63] In the last chapter I gave a straightforward and intuitive example of this: "tracing cases," in which an agent is indirectly responsible for an action because it can be traced back to other actions or omissions for which she is directly responsible. I claimed that a person is still attributively responsible for an action influenced by implicit bias if she previously did not take reasonable precautions to prevent the acquisition of such a bias, or if she did not respond to the knowledge that she had such a bias, because these prior omissions indicate a blameworthy lack of commitment to treating people fairly and equitably.

do with respect to mitigating implicit bias.[64] We might only be able to know this in the limiting cases, when a person really has done nothing or hardly anything at all, where again this means we now have good enough epistemic evidence to justifiably blame them. Never or rarely acting toward some goal is evidence of *not* having adopted that end—which can be attributed to a person as part of her practical identity, and hence *is* blameworthy.

Yet there is a further and deeper moral reason why it is important and *good* that there should be no well-defined cut-off. Warner Wick writes that imperfect duties are "without assignable limits," such that: "We can never say, 'There, I have at last done all that I ought to do for other people'" (li). As Marcia Baron puts it, imperfect duties do not permit "drawing a line around one's 'own life' or 'own projects' and seeing morality the way one might see mowing the lawn (or as a child might see doing his homework or performing his boyscout deeds): as something to get out of the way" (250). Tempting though it may be to some, I think we can all recognize that these are deeply odd ways of conceiving of moral duties. Baron defends this claim by appealing to Kant's theory of imperfect duty[65], but I think her point extends beyond Kantian ethics. Utilitarianism, for example, is famously subject to such demandingness objections, since it follows from our being morally required to maximize overall utility of all the consequences of our actions that the vast majority of our actions will be morally mandated, even those involving the most mundane choices (e.g. which cereal to have for breakfast, far from being morally neutral, might depend on which company will do the most good—or least evil—with our consumer support). Virtue ethics, moreover, would surely never countenance such contentment with meeting minimal moral thresholds: such an attitude seems utterly at odds with any conception of the virtuous person or moral exemplar.

It is thus in the nature of morality itself that we should not be able to reply to the question *How much?* with: "*This* much is enough." To do is to invite complacence, bad faith, and rule-fetishism. For a recent and notorious example from a slightly different context, consider the following passage from a site visit report conducted by Valerie Gray Hardcastle, Peggy

---

[64] A similar observation has been made by in Bell (2013); Bell uses it to argue against the view that blame requires a certain kind of moral standing.

[65] She writes: "This is all quite alien to Kant's ethics. There is no clear line of demarcation between what I must do, morally, and what is nice but morally optional. Nor does Kant attempt to trace such a line of demarcation. To do so he would have to give up a very central, even foundational, duty: the imperfect duty we each have to strive to perfect ourselves morally. This duty, which we cannot escape by being "good enough," underlies our other imperfect duties. (249)

DesAustels, Carla Fehr on behalf of the American Philosophical Association Committee on the Status of Women, addressing issues of sexual harassment and an otherwise hostile climate that threatened the flourishing of women in the department (and, by extension, the discipline):

> The Department uses pseudo-philosophical analyses to avoid directly addressing the situation. Their faculty discussions revolve around the letter rather than the spirit of proposed regulations and standards. They spent too much time articulating (or trying to articulate) the line between acceptable and unacceptable behavior instead of instilling higher expectations for professional behavior. They spend significant time debating footnotes and "what if" scenarios instead of discussing what they want their department to look and feel like. In other words, they spend time figuring out how to get around regulations rather than focusing on how to make the department supportive of women and family-friendly. (7)

Here the problem of demarcation is about what constitutes acceptable and unacceptable behavior. But as this example suggests, trying to find an overly precise answer to the Demarcation Question lends itself to a trivialization of and perfunctory engagement with the problem at hand, rather than sustained and careful reflection and re-reflection. Such deep engagement is particularly important—as with cases of implicit bias, privilege, and discrimination—when the problems are complicated and ill-defined. There is thus good reason, *even if* easy answers existed, to refrain from handing them out like candy: for these are the sorts of problems that require deep and personal engagement by each individual agent. As we have seen before, externally-imposed standards that fail to get internalized are likely to be perceived as intrusive and oppressive, and thus are likely to be ineffective. Issues surrounding implicit bias are particularly likely to "hit too close to home" because they indict a person's own deep-seated habits and patterns of thought, and further because they implicate the environments and communities of which people take themselves to be loyal members.[66] The kinds of intensive critical reflection on oneself and one's community that are thus required to ferret out unwarranted privilege, moral blind spots, and implicit biases are no easy accomplishment. And the difficulties are not merely psychological: as implicit bias perhaps illustrates best, the

[66] Howard McGary (2008) gives this description of a failure of accountability: "[E]ven members of faculties who have spoken out against bad practices in their departments are reluctant to make matters public when the department fails to change those practices. To do so is often seen as a betrayal of their colleagues. People's lives and their sense of self-worth are often closely connected with being a part of a congenial workplace. So blowing the whistle in this context is very different from exposing wrongdoers with whom you don't share a personal relationship." (305) On my view, this reluctance is likely grounded in an assumption that whistle-blowing amounts to denunciation and that it is the only option. Sometimes it might be, but more likely there will be other ways of reframing the problem in terms of professional ideals, engaging with recalcitrant colleagues on their own terms, and what Patricia Hill Collins (1990) calls "working the cracks" through "insider resistance" (see Ch.4, Section 5.2).

difficulties are also *epistemic*, since introspection (by definition) cannot guarantee the detection of bias. The value of taking ourselves to have imperfect duties, then, lies not so much in their providing us a demarcation between right and wrong acts, but in occasions to reflect deeply on pressing and complex moral problems in which we are personally and inextricably implicated. But it takes serious, sustained hard work and critical self-reflection to come to grips with privilege, an understanding of which is necessary for social change. A study by Powell, Branscombe, & Schmitt (2008), for example, found that framing racial inequality in terms of Black disadvantage or White privilege had an effect on the racial attitudes of White participants. Framing inequalities in terms of White privilege made them score higher on measures of collective guilt, which mediated greater willingness to recognize the continued existence of anti-Black discrimination and disadvantage, as well as empathy for efforts to promote demands for racial equality.[67] This is difficult to achieve, especially given what Robin DiAngelo (2011 calls "white fragility," a lack of "racial stamina" caused by the fact that an "insulated environment of racial privilege builds white expectations for racial comfort while at the same time lowering the ability to tolerate racial stress" (55-56). We can also read the "pseudo-philosophical" antics of the Colorado case as an example of evasive maneuvering undertaken to shirk the "intolerable" work of identifying and taking responsibility for the workings of gender privilege. What I want to stress here is that no amount of philosophical fussing about the requirements of imperfect duties, however good the arguments, is likely to bring people to undergo the hard work of confronting implicit bias, privilege, and discrimination in their own persons and institutions. (By contrast, I will argue, appeals to people's genuine commitment to moral ideals—in other words, when they are internally motivated by standards with which they themselves identify—may fare better.)

All this means that merely appealing to the concept of imperfect duties will not help us much with the Demarcation Question of determining when someone has acted wrongly or rightly, or done what she ought or ought not to have done. But we need not know the answer to the Demarcation Question in order to know some of the things that we can do to fulfill our imperfect duties—indeed, we are morally disingenuous or confused if we fixate on the former rather than the latter. So rather than try to determine whether a person has reached some threshold of what she is required to contribute toward mitigating her implicit biases, on the basis

[67] However, appealing to guilt alone—without assuaging the social identity threat that likely accompanies it—is also risky, and people are likely to avoid situations where they anticipate being made to feel guilty. (I present Does, Derks, & Ellemer's (2011) discussion of this particular result in the next section.)

of which we can blame or punish her, we should instead be focusing on efforts on how we can respond critically to her—how we can educate and motivate—when she misses out on opportunities to do so. There is a difference between wanting to perform morally good actions because one wants to avoid being morally deficient, and wanting to perform morally good actions because one wants to bring about morally good changes in the world. It is this latter project that provides purpose to practices of moral responsibility, not the former.

It is clear that what is needed here is not some well-defined cut-off in answer to the question of exactly how much we must do to fulfill our imperfect duties, but something more like a standing disposition to strive toward doing as much as we can, where the "can" must be understood in a highly (but not overly) flexible and contextualized way. This disposition to strive is a sort of fundamental moral orientation, which includes a commitment to moral self-improvement and a willingness to engage in hard critical reflection on the imperfections of oneself and one's community, but beyond that a sort of intimate *devotion* to and personal *investment* in righting injustice and doing good. This is what moral exemplars display in such abundance, and what allows them to light the same flames of striving within us.[68] I think that most of us have something of this disposition to strive in some more form or another: just consider the devotion with which people strive to be good parents, friends, neighbors, citizens, teachers, and so on.

Although imperfect duties do not do demarcate with precision what we ought to do, they do prescribe what we ought to do in a more general way. Let us consider, then, a different function of imperfect duties. Consider Susan Hale's characterization of what I will call the *end-* or *ideal-setting* function of imperfect duties (emphasis mine), and which highlights that imperfect duties belong to the Ideal realm:

> Perfect duties, then, are duties prescribing particular actions, in contrast to imperfect duties prescribing *adoption of particular ends or adoption of particular moral ideals as guides for our actions*. Imperfect duties do not provide strict guides for our actions since they prescribe no particular actions. Rather, *they prescribe only that agents ought to strive toward attainment of particular ideals*, leaving each agent free to choose by which actions she will so strive and toward which of these ideals she will strive most devotedly. (275)

This gets us closer to the fundamental moral orientation I described in the last section. What I want to suggest is that imperfect duties represent the relationship in which we stand relative to

[68] Again, talk of "virtue" comes naturally to mind here; this disposition to strive strikes me as a sort of pure moral virtue, not identical to the practical wisdom of the truly virtuous person, which may be beyond the reach of most of us, but is common to the ordinary decent person.

our moral ideals. When we take ourselves to be under an imperfect duty, we *orient* ourselves toward an ideal. As I mentioned earlier, this orientation encompasses a range of cognitive, affective, motivational, and behavioral components. We can see this in Elizabeth Anderson's (1995) characterization of ideals:

> The core of an ideal consists in a conception of qualities of character, or characteristics of the community, which the holders regard as excellent and as central to their identities. Associated with this core is a conception of admirable conduct or worthy practices and projects that demand the cultivation, exercise, and expression of these qualities. An ideal is constitutive of a person's identity if it governs her self-assessment and her responses to achievement and failure and if she uses it to discipline her desires and frame her choices. (8)

Note that what Anderson is describing here is precisely Higgins's (1987) self-regulatory process of reducing discrepancies from one's Ideal self-guide. For example, a person who holds the ideal of being a good teacher will be committed to improving herself by a teacher, where this might involve reading articles and attending workshops on pedagogical theory on implicit bias, experimenting with different teaching techniques, and seeking out feedback and other ways to measure her students' learning outcomes. She will engage in hard, sustained scrutiny of her own methods and that of her department, institution, and wider field of higher education, by at once consulting and being spurred on by student feedback, discussions with her peers and colleagues, and regional, national, or even international trends and patterns. Such scrutiny has the potential to be highly threatening to her "sense of self-worth" or to be a "betrayal of [her] colleagues,"[69] insofar as it might reveal her own mistakes and inadequacies, lead her to encounter hostility and resistance in her community, force her to confront deep-seated injustices as well as privileges and benefits that she enjoys at the cost of other educators and students differently positioned in the economy of higher education, and so on. But a person who genuinely holds the ideal of being a good teacher does so willingly because it is important to her—because, as Anderson puts it, it forms part of her "conceptions of what kind of person she ought to be, what kinds of character, attitudes, concerns, and commitments she should have" (6).

This example also demonstrates the way in which imperfect duties acquire a specific shape and *content* when they are grounded in particular ideals. The imperfect duty of being a good teacher, can be seen as a species of the paradigmatic imperfect duty of beneficence: one

[69] See Fn. 66.

helps others *by* or *through* being a good teacher (or a good firefighter, parent, neighbor, citizen, etc). Conceiving of imperfect duties as end-setting makes it somewhat easier to gauge how well a person is fulfilling her imperfect duties—though as I argued in the last section, it is part of the nature of imperfect duties that one does not treat the duty as something merely to be gotten out of the way. It is easy to see how a teacher who consistently ignores negative student feedback and never modifies any parts of her course fails to live up to (one instance of) her imperfect duty of beneficence—moreover, a person like that obviously just isn't a *good teacher*. Now imagine that this teacher's courses have long had the effect of discouraging women and minorities from continuing their studies in that subject, in part due to her implicit biases about which authors and topics are worth covering, her style of lecturing and moderating discussion, her choices and manner of assessment, etc. In this case I think we *could* attribute such actions to the teacher as reflective of her practical identity, even if she openly avows that she does not intend to discriminate against women and minorities and even if she would not endorse her implicit biases, because of her failure to do what could reasonably be expected of her to monitor and improve her course.

One could aver that this simply shows that she never had the ideal of being a good teacher to begin with, and thus that she doesn't have an imperfect duty at all. But I think this would be too hasty. While it is certainly true that individuals vary in the extent to which they embrace different ideals, it is not the case that ideals are wholly a matter of individual choice: after all, the Ideal realm is still a domain of morality, and is thus morally *required*. In the first place, some ideals are unacceptable because they are immoral or unjust: an ideal of being a good murderer, for instance, or good slave-owner. In the second place, ideals are not entirely under an individual agent's discretion because (as my examples have already shown) ideals are often *role-based*. Being assigned the job of being a teacher—being expected to *do the job well*—already means that one has been assigned the ideal of *being a good teacher*. The same goes for being a good parent, neighbor, or citizen. Ideals are thus part of what is distributed along with duties and burdens in the moral division of labor, i.e. the system of demands and expectations that ground moral accountability. No matter what we do, being a member of a moral community entails being assigned certain moral ideals that we are expected to embrace—or at least, to strive toward (where embracing the ideal is certainly a way of becoming motivated to strive). This notion of distribution also captures the sense in which people are subject to different ideals, and thus different ways of carrying out imperfect duties. As Anderson notes:

> Different ideals may require the cultivation of incompatible virtues of the pursuit of some projects that necessarily preclude the pursuit of others. Individuals with different talents, temperaments, interests, opportunities, and relations to other rationally adopt or uphold different ideals (7).

This is because the exigencies of communal life require the assignment of different jobs and hence ideals to different individuals. Moreover, I claim, this provides yet another reason to for thinking that the "Ideal" realm corresponds to the prescriptive system in Janoff-Bulman, Sheikh, & Hepp's (2009) dual-systems model of moral regulation. They make the following conjecture:

> The mandatory, concrete, restraint-based nature of proscriptive morality readily lends itself to a system of laws that is focused on what we should not do and that can apply to all in a society…[By contrast,] although caring for others is a positive value, societies typically recognize the more discretionary nature of caring…In a given society, then, prescriptive but not proscriptive morality is likely to be role contingent...Given that prescriptive morality is more discretionary, societies tie our positive obligations to social roles in an attempt to proactively regulate and promote these beneficial acts. (534-535)

Let me briefly summarize, then, my interpretation of the distinction between perfect and imperfect duties. Perfect duties take *particular* actions at a time as their object. Imperfect duties, by contrast, govern *patterns* of actions over time in accordance with striving toward a specific moral ideal. Perfect duties provide answers to the Demarcation Question by specifying whether some act is morally required, permissible, or impermissible to perform them; imperfect duties do not always generate clear answers to the Demarcation question, but they provide guidance in virtue of their role-specific content. To have an imperfect duty is to be oriented toward a role-specific moral ideal, and to be motivated to reducing discrepancies with the ideal. The content of perfect duties is not up to the discretion of the individual agent; they act as universal constraints on action. The content of imperfect duties, since they are grounded in specific role-based moral ideals over (some of) which the agent has freedom of choice and prioritization, is to this extent up to the agent's discretion. But since imperfect duties are also *assigned* to an agent in virtue of the social roles she occupies, they are in this sense *not* up to her discretion. And while the Demarcation Question of whether or not a person has lived up to her imperfect duty to mitigate bias is relevant for determining whether an action caused by implicit bias is genuinely attributable to her, I have argued that we should leave that question aside in favor of the non-appraising responses that are appropriate whenever we have occasion to strive harder in pursuit of our moral ideals.

6. Toward an Account of Non-Appraising Responses

I will now offer a tentative proposal for how we can recognize and develop the non-appraising responses I have been advocating. I propose that effective non-appraising responses will have the following features:

1. Focus on the action, and not the actor
2. Appeal to (shared) role-based moral ideals of the relevant community
3. Offer directions for future improvement

For lack of a better term, let me just stipulate that I will refer to this kind of response as "moral admonition" or "moral admonishment," since the word "admonish" does have something of the range and tone that I am aiming for. (Its dictionary definitions include "to indicate duties or obligations to," "to express warning or disapproval to especially in a gentle, earnest, or solicitous manner," and "to give friendly earnest advice or encouragement to,"[70] as well as "to caution, advise, or counsel against something," "to reprove or scold, especially in a mild and good-willed manner," and "to urge to a duty; remind"[71].) To my ear, "admonition" carries more of the meaning of urging to a duty, while "admonishment" strikes me as closer to disapproval and reproof; I shall hence use them accordingly.

Moral admonition has the first of these features because that is simply what it is to be a non-appraising rather than an appraisal-based response. It may thus incorporate an "inoculation" making clear at the beginning that the target is for the action and not the actor. Moral admonition invokes shared ideals (not obligations) of the relevant community, where I mean "community" here in the broadest possible sense of the multiple relations an agent stands in to others as a teacher, colleague, parent, friend, citizen, neighbor, and so on. Our enmeshment in these relationships means that role-based ideals—which can be appealed to in different ways, by different people, in different contexts—can flexibly be brought to bear on whatever people intrinsically care about, in order to engage their prescriptive, promotion-focused moral motivations. The existence of actual relationships—as any good organizer knows—is crucial for combating various forms of psychological recalcitrance and threat (which I will describe shortly), meaning that there is important work to be done by each individual that cannot be performed by others who lack those relationships. These relationships need not be intimate, since promising

[70] Merriam-Webster Dictionary.
[71] Random House Dictionary.

ideals are likely to be close at hand depending on whether one is in the workplace, at home, at a party, in the street, at a public forum, etc. Moreover, since moral criticism is not a one-shot deal, the ongoing conversation in which moral admonishment takes place allows the possibility of trying out different ideals until one "gets a bite," as it were. To be sure, the more common ground there is, and the more there is an underlying relationship of trust and respect, the more likely it is that a person will manage to pull off successful moral admonition, but this is surely true of blame and the other reactive attitudes as well (if we agree with Springer (2013) that their aptness depends on how well they achieve the aim of communicating moral concern). Finally, moral admonishment can incorporate directions for future improvement in more or less explicit ways. While specific suggestions may be available and appropriate in some circumstances, at other times merely feeling a possible discrepancy with an ideal is already sufficiently informative, and may even already be pushing the limits of knowability. As Springer (2013) writes:

> A concern calls for a response whose nature is partly to help sort out what response is called for. Hence, in attempting to answer "What's wrong?" the stance of concern focuses more on bringing the hearer's attention to the same difficulty than on spelling a claim out explicitly. (138)

Moral admonition, then, may be no more than a "gesture," as Springer puts it, in the direction of the discrepancy. In short, moral admonition serves to *remind* people of the moral ideals that make up their imperfect duties, by *drawing attention to* ways of falling short of those ideals, sometimes by *making clear* the reasons (the ideals) against so acting, in order to *prevent* such failures in the future.

I would also like to emphasize that these three features accord with the tenets of self-determination theory: moral admonition promotes autonomy by aiming at helping an agent better approximate her own intrinsically-motivating ideals, offers constructive feedback that increases her competency for doing so, and maintains relatedness by avoiding the accusatory and divisive effects of blame. Better yet, moral admonition's appeal to *shared* moral ideals reinforces the relational bonds between the criticizer and the criticized, reaffirming that they are "on the same team." Rather than disrupting, modifying, or withdrawing from relationships in the manner of suggested by popular accounts of blame[72], such responses *preserve* these relationships, enlisting them in service of communal efforts to improve. Non-appraising responses that support

[72] Cf. Hieronymi (2008), Scanlon (2008), and Bennett (2002).

these conditions thus enable the re-internalization and re-affirmation of norms, moral ideals, and self-guides that commit people to social equality.

A recipe for an effective non-appraising response, then, might look something like the following: "I'm sure you didn't mean it this way, but that thing X you just said might have the effect Y on [e.g. some of our students]. Since we're all trying to [e.g. improve the climate, increase enrollments, etc.], why don't you try saying Z instead?" (Of course, it need not be quite so flat-footed as this, since these elements can often be implicitly assumed or made clear from context.) Another possibility might be to look to the way that we respond to children and young adults. In these cases we are often not interested in *blaming* or assessing their quality of character as agents, since that is precisely what is still under formation, but we admonish them in order to evoke and promote certain ideals that we want them to grow into adopting.

In this connection, Cheshire Calhoun (1989) raises a potential objection to my suggested way of proceeding. She considers and rejects the option of saying "I know you didn't know any better and I'm not blaming you. Just remember in the future that you shouldn't…" on the grounds that this only works in interactions of unequal moral status, as between parents and children, and that between equals it is likely to come off as insulting or arrogant (1989, p. 401). I think, however, that the invocation of shared ideals—from which we *all* remain at some distance, as our vulnerability to implicit bias clearly attests—can preserve equal standing while simultaneously encouraging cooperation, shared vigilance, and joint efforts to work toward them. Moreover, in the less clearly-defined cases of moral admonition I mentioned above, an open-ended invitation for the other to help bring clarity to this inchoate concern may be crucial to successfully engaging her attention and motivation. Part of what explains the amorphous nature of these concerns is precisely that they can be sharpened in different ways according to the different roles and obligations occupied by those sharing the concern. Springer (2013) illustrates this using a medical case in which a patient's death was found to have been the result of 48 different factors. Nurses, doctors, pharmacists, hospital administrators, equipment manufacturers, and industry regulators were "informed by a different practical vantage point, a different set of relationships, and hence a somewhat different angle of concern"; nevertheless, Springer points out, "what is most striking is how the *continuity* of alarm and concern carried the same provocative event through multiple social vantage points" (156-157). This multiplicity of roles and vantage points means that a given moral admonishment often will not contain the

whole story, and we who understand the problem from one perspective cannot claim any privileged authority over those we admonish who may exhibit a different kind of uptake. In morally admonishing others, remembering this can help us avoid insult and arrogance.

A more general objection to be raised at this point is: why should we think that moral admonition will be any more effective than blame and resentment at bringing about positive change? I think that there are reasons to be optimistic, based on positive findings such as the Kamins and Dweck's (1999) work on person- vs. process-oriented feedback I described in Section 4, along with research revealing what *not* to do. Recall the Does, Derks, & Ellemers (2011) study on the effect of appealing to moral ideals (as I have advocated through the use of non-appraising responses) rather than moral obligations. They conclude their study as follows:

> The current findings suggest an alternative way for these group members to be confronted with the ingroup implications of inequality, without the cost of a lowered collective esteem. In this respect, framing the moral implications of inequality in terms of yet-to-be-attained ideals appears to be an effective way to confront Whites with their group's advantaged position without raising social identity threat, and to increase their support for a more equal society. (569)

It is important to note here that the "cost of a lowered collective esteem" should not be construed merely in terms of preventing discomfort to privileged groups. Rather, the costs of threatening self-esteem accrue much more severely to members of disadvantaged groups because of the way they prevent privileged groups from accepting the need for change. As Kahan and Braman's (2008) research on cultural cognition demonstrates, people's evaluation of evidence—in particular, their evaluation of the reliability and trustworthy of epistemic authorities—is heavily bound up with their social identities. These results are further supported by Nyhan and Reifler's (2010) research on political misperceptions and the "backfire effect," in which people who hold false political beliefs actually hold their beliefs more strongly after receiving the correct factual information. More information, in the presence of social identity threat, merely serves to exacerbate motivated reasoning and willful ignorance. The danger of the backfire effect is likely to be particularly acute in the case of incidents involving implicit bias, both because individuals are highly motivated not to perceive themselves as biased, and because such incidents are often ambiguous enough to admit of multiple interpretations. Previous research has demonstrated that high-status groups consistently view the prospect of social equality as threatening in-group losses; this holds even when the status is acquired in so-called "minimal group paradigms," where the

higher status is arbitrary, temporary, and experimentally-induced (e.g. by assigning half the group to wear red shirts and the other half blue shirts) (Does, Derks, & Ellemers 2011, p. 562-3). Moreover, Does & colleagues (2011) argue that even though inducing collective guilt among privileged groups has been shown to increase support for compensatory affirmative action (as in the Powell, Branscombe, & Schmitt (2008) study I mentioned earlier), such methods are suboptimal because people are motivated to avoid feelings of guilt and thus are likely to engage in defensive reactions such as victim-blaming and denying past injustices or their relevance to contemporary inequality (p. 563). Additionally, members of privileged groups who are narrowly motivated to avoid feelings of guilt do not support more broad-based initiatives to achieve social equality (Powell, Branscrombe, & Schmitt 2008). Recall also Norton et al.'s (2010) research on "racial paralysis" found that Whites' concerns about appearing racially biased led them to make significant efforts to avoid making cross-race judgments, while Plant and Devine's (2003) and Plant and Butz's (2006) studies on interracial anxiety showed that the fear of appearing biased led White people to avoid interracial interactions.

An appeal to moral ideals, by contrast, might actually encourage people to seek out more diverse environments that allow for greater opportunities to work toward their ideals of diversity and equality. This is supported by an experiment by Trawalter and Richeson (2006), who found that priming participants with a promotion focus ("try to have an enjoyable intercultural dialogue") going into an interracial reaction was less cognitively draining than a prevention focus ("try to avoid appearing prejudiced"), as measured by a follow-up Stroop task. Moreover, Nyhan and Reifler's (2015) backfire effect was reduced when subjects first engaged in "self-affirmation" exercises where they recalled times in which they had successfully lived up to their values. This again suggests the motivational efficacy of moral framing in terms of striving toward ideals, rather than not violating obligations. Effective moral admonition, then, will often open with this kind of self-affirmation, or with an affirmation of the bonds of membership in a common moral community. Indeed, the invocation of a shared moral ideal can perform just this function of priming such an affirmation. Moral criticism aimed at illuminating and correcting implicit biases, then, should for pragmatic reasons be performed in ways that minimize social identity and self-esteem threat that produces hostile and defensive reactions. As I have been arguing, this is much more likely to be accomplished by deploying non-appraising rather than appraisal-based responses. If non-appraising responses can engage the promotion-focused, prescriptive moral

regulation system that encourages an eager rather than vigilant strategy, they may thus help to overcome the risks of interracial anxiety and white fragility.

In addition to moral admonition, let me give another example of a potential non-appraising response: demand for apology. Margaret Urban Walker (2006) argues apologies function to *repair moral relations*, which entails repairing relationships of "default trust" among members of a society that they adhere to common moral norms—in other words, the relationships of trust that form the basis of moral accountability. Empirical studies of the victims of crime have demonstrated that victims of wrongdoing demonstrate a need not only for material compensation and rectification of harm done, but also for such forms of "emotional restoration"[73]. Apologies thus represent a particular "burden" of restoration that cannot simply be transferred to an insurance scheme, and that still serve to *bind an agent to her action* without thereby rendering it attributable to her as a basis for evaluation. Consider in this connection Bernard Williams' (1982) description of agent-regret, which also binds an agent to her action in the absence of any responsibility as attributability. Williams points out that the desire to make amends after causing harm may go *beyond* any compensation that could equally be paid by an insurance company. I believe that this additional "reparative significance" also attaches to actions for which we are accountable but which are not attributable to us (Williams 1982, p. 29). Indeed, part of the importance of material compensation, I believe, is due precisely to the reparative significance—to the recognition of victims who have suffered undue harm—that it normally entails.[74] Jeffrey Helmreich (2012) has also argued that even self-critical apologies need not involve a negative appraisal of oneself. He writes:

> The core of my argument against [the courts'] construal of self-critical apologies – which reads them as legally incriminating – lies in showing that all injurers, even those with no legal or moral culpability, have reason to be self-critical about the harms they caused. Even purely non-culpable injurers have reason to take the stance expressed by remarks such as "I really messed up," or "I did something horrible." That is because harming others, even blamelessly, constitutes a misuse or misfire of one's considerable efforts to avoid doing so, efforts in whose success moral agents are deeply invested. When they harm someone, then, they appropriately regard their injurious conduct with criticism, if not guilt or blame. In fact, it will be compatible with what follows that morally blameless

[73] Cited in Walker (2006, p. 91).

[74] This suggests that certain kinds of apologies from collective agents and public figures, e.g., official apologies for state atrocities, may be particularly important even in the absence of material compensation. While this kind of recognition will not usually constitute sufficient reparation for victims, such recognition matters, and its absence may be conspicuous.

> injurers have no reason to feel responsible for the harms they caused, nor to feel "guilty," if that implies self-blame. (19)

While this may not be true of all apologies, Helmreich is certainly right about some class of these *non-appraising* apologies. Self-criticism here may simply be entailed by the recognition that one has failed to live up an ideal, or even that one has (unintentionally) violated an obligation. This is evident from the appropriateness of saying "I'm sorry!" even for completely no-fault incidents, say, when one accidentally steps on another's foot.[75] By making an apology, one takes on the "burden" of publicly recognizing and disavowing the wrong that occurred (of something that "should not have happened") thereby making an effort to re-affirm shared moral norms and to re-equilibrate relationships of default trust. Thus, I claim, part of being accountable for having (and manifesting) implicit bias is being subject to the expectation or demand for apology and recognition of victims in the event of such "misfires".

There is one last important issue that I have not yet touched on. I want to make clear that my discussion of imperfect and perfect duties, grounded in the "Ideal" realm and the "Ought" realm, is not intended to suggest that harmful actions caused by implicit bias are violations of imperfect rather than perfect duties. Duties of non-maleficence and justice, to avoid harming others and to treat them equitably, are perfect duties. Their violation requires redress and restitution for the victims, and blame, resentment, and perhaps punishment for the perpetrators who are responsible. In the case of implicit bias, however, as I argued in the last chapter, condemnation may not be appropriate because we face epistemic and moral uncertainty in determining whether the violation is genuinely attributable to the agent—in other words, whether she lacks responsibility as attributability. On this understanding, implicit bias (sometimes) provides an *excuse* for admittedly wrongful behavior that otherwise deserves appraisal-based responses; but since the agent still has accountability for the violation, it is appropriate for her to bear instead the material burdens of redress, as well as to be subject to moral admonishment. An alternative way of drawing the same line, however, would be to say that whether harmful actions caused by implicit bias are violations of perfect or imperfect duties depends on which description we affix to the action. While "unjustly harming" or "discriminating against" are violations of perfect duties, "being influenced by an implicit bias" or "unintentionally discriminating" may not be, insofar as they violate "Ought Implies Can." On

[75] I am indebted to Peter Railton for this example.

this reading, non-appraising responses are appropriate while appraisal-based ones are not, because these are failures of imperfect duty rather than violations of perfect duty. This resembles the difficulties raised by causally complex actions in a highly interconnected world: perhaps buying one brand of cereal rather than another unjustly harms those exploited along the way, but "not harming" and "not exploiting" seem like perfect duties while "being an ethical consumer" seems like an imperfect duty. Both of these formulations seems acceptable to me, so long as we are clear that the victims of implicit bias deserve the same redress as victims of wrongdoing that are more clearly a violations of perfect duty.

7. Conclusion

In this chapter I argued that there is a moral-theoretical as well as a psychologically-grounded distinction between the "Ought" Realm and the "Ideal" Realm. Perfect duties, and their appraisal-based responses like blame, resentment, and shame that attend their violation, belong to the "Ought" Realm. Imperfect duties, by contrast, belong to the "Ideal" Realm and thus give rise to different, non-appraising forms of moral criticism like admonition, remonstration, and reprimand, which serve to remind and direct people to strive toward their moral ideals. I drew a distinction between two different functions that imperfect duties can serve: a demarcation function, and an end- or ideal-setting function. I argued that imperfect duties do not yield a precise cut-off for exactly when an agent has done what she can reasonably be expected to do, but that they are still valuable in virtue of a different kind of function: they furnish specific moral ideals that guide agents' patterns of actions in the context of their particular social roles.

The problem of implicit bias encompasses both perfect and imperfect duties. We are accountable for particular actions at particular times that violate our perfect duties, which mandate us to make amends. But we are also accountable for undertaking patterns and habits of action across time that aim at mitigating implicit bias and changing unjust social structures, and failure to do what we can reasonably be expected to do in this regard can be attributable to us. We violate our imperfect duty to mitigate bias when we do not engage in a sufficiently robust pattern of bias-mitigating behavior, but even when we do, we can still on particular occasions fail to perform actions that would help to mitigate bias. In criticizing agents for these failures, we can avail ourselves of moral criticism that frames them in terms of an agent's falling short of her own

ideals, rather than as violating abstract moral obligations. I would like to suggest now that the criterion of reasonable expectation—insofar as it explicitly depends on the existence of a moral community and the basic moral expectations that we are subject to as members—is actually rooted in responsibility as accountability. Further development and refinement of a theory of moral accountability will be my task in the next chapter.

References

Anderson, Elizabeth. Value in Ethics and Economics. Cambridge, MA: Harvard University Press, 1995.

Baron, Marcia. (1987). "Kantian Ethics and Supererogation." *Journal of Philosophy* 84(5): 237-262.

Bell, Macalester. "The Standing to Blame: A Critique", in J.D. Coates & N. Tognazzini (eds.) Blame: Its Nature and Norms. New York: Oxford University Press, 2013: 263–281.

Bennett, Christopher. (2002). "The Varieties of Retributive Experience." *Philosophical Quarterly* 52(207):145-163

Boldero, J.M., Moretti, M.M., Bell, R.C., & Francis, J.A. (2005). "Self-discrepancies and Negative Affect: A Primer on When to Look for Specificity, and How to Find It." *Australian Journal of Psychology* 57(3): 139 – 147.

Calhoun, Cheshire. (1989). "Responsibility and Reproach." *Ethics* 99(2): 389-406.

Coates, Justin and Neal Tognazzini, eds. Blame: Its Nature and Norms. New York: Oxford University Press, 2012.

Deci, E. L., & Ryan, R. M. (2000). "The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-determination of Behavior." *Psychological Inquiry* 11: 227-268.

DiAngelo, Robin. (2011). "White Fragility." *The International Journal of Critical Pedagogy* 3(3): 54-70.

Does, S., Derks, B., & Ellemers, N. (2011). "Thou Shalt Not Discriminate: How Emphasizing Moral Ideals Rather Than Obligations Increases Whites' Support for Social Equality." *Journal of Experimental Social Psychology* 47: 562-571.

Gert, Bernard. Common Morality: Deciding What to Do. New York: Oxford University Press, 2007.

Goodin, Robert. Utilitarianism as a Public Philosophy. Cambridge: Cambridge University Press, 1995.

Fricker, Miranda. (2014). "What's the Point of Blame? A Paradigm Based Explanation." *Noûs* (forthcoming).

Fricker, Miranda. Epistemic Injustice: Power and the Ethics of Knowing. New York: Oxford University Press, 2009.

Hieronymi, Pamela. (2004). "The Force and Fairness of Blame." *Philosophical Perspectives* 18(1): 115-148.

Hale, Susan. (1991). "Against Supererogation." *American Philosophical Quarterly* 28(4): 273-285.

Hardcastle, Valerie, DesAutels, Peggy, and Carla Fehr. (2013). "Report on Site Visit." American Philosophical Association Committee on the Status of Women.

Hardin, E.E., and J.L. Lakin. (2009). "The Integrated Self-Discrepancy Index: a Reliable and Valid Measure of Self-discrepancies." *Journal of Personality Assessment* 91(3):245-53

Helmreich, Jeffrey S.. (2012). "Does 'Sorry' Incriminate? Evidence, Harm and the Meaning of Apologies." *Cornell Journal of Law and Public Policy* 21(3).

Higgins, E. Tory. (1997) "Beyond Pleasure and Pain." *American Psychologist* 52(13): 1280-1300.

Higgins, E. Tory. (1987). "Self-Discrepancy: A Theory Relating Self and Affect." *Psychological Review* 94(3): 319-340.

Holroyd, Jules. (2007). "A Communicative Conception of Moral Appraisal." *Ethical Theory and Moral Practice* 10: 267–278.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). "Proscriptive versus Prescriptive Morality: Two Faces of Moral Regulation." *Journal of Personality and Social Psychology* 96: 521-537.

Kahan, Dan M. and Donald Braman. (2006). "Cultural Cognition and Public Policy." *Yale Law & Policy Review* 24: 147.

Kamins, Melissa L. and Carol S. Dweck. (1999). "Person versus Process Praise and Criticism: Implications for Contingent Self-Worth and Coping." *Developmental Psychology* 35(3): 835-847.

Kant, Immanuel, & Mary J. Gregor. (1996). *The Metaphysics of Morals*. New York: Cambridge University Press.

Legault, L., Gutsell, J.N., & Inzlicht, M. (2011). "Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (but Also Increase) Prejudice." *Psychological Science* 22(12): 1472–1477.

Malle, B.F., Guglielmo, S. & Monroe, A.E. (2014). "A Theory of Blame." *Psychological Inquiry* 25(2): 147-186.

McGary, Howard. (2008)."Psychological Violence and Institutional Racism: The Moral Responsibility of Bystanders" in Laurence M. Thomas, ed., Contemporary Debates in Social Philosophy (pp. 299-311), Oxford: Blackwell Publishing.

Norton, M.I., Mason, M.F., Vandello, J.A., Biga, A., & Dyer, R. (2013). "An fMRI Investigation of Racial Paralysis." *Social Cognitive and Affective Neuroscience* 8: 387-393.

Nyhan, Brendan, and Jason Reifler. (2015). "The roles of information deficits and identity threat in the prevalence of misperceptions". (Unpublished manuscript).

Nyhan, Brendan, and Jason Reifler. (2010). "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32(2): 303-330.

Ozgul, S., Heubeck, B., Ward, J., & Wilkinson, R. (2003). "Self-discrepancies: Measurement and Relation to Various Affective States." *Australian Journal of Psychology* 55: 56-62.

Plant, E.A. and D.A. Butz. (2006). "The Causes and Consequences of an Avoidance-focus for Interracial Interactions." *Personality and Social Psychology Bulletin* 32: 833-846.

Phillips, A.G., & Silvia, P.J. (2010). "Individual Differences in Self-discrepancies and Emotional Experience: Do Distinct Discrepancies Predict Distinct Emotions?" *Personality and Individual Differences* 49(2): 148–151.

Plant, E.A. & Devine, P.G. (2001). "Responses to Other-Imposed Pro-Black Pressure: Acceptance or Backlash?" *Journal of Experimental Social Psychology* 37: 486-501.

Powell, A.A., Branscombe, N.R., & Schmitt, M.T. (2005). "Inequality as Ingroup Privilege or Outgroup Disadvantage: The Impact of Group Focus on Collective Guilt and Interracial Attitudes." *Personality and Social Psychology Bulletin* 31(4):508-521

Railton, Peter. (2010). "Values and Valuing: The Why and How." Paper presented at APA Eastern Division Meeting, Boston, MA. (Unpublished manuscript).

Radzik, Linda. Making Amends: Atonement in Morality, Law, and Politics. New York: Oxford University Press, 2011.

Scanlon, T.M. Moral Dimensions: Permissibility, Meaning, Blame. Cambridge, MA: Belknap Press, 2010.

Scanlon, T.M. What We Owe Each Other. Cambridge, MA: Belknap Press, 1998.

Schroeder, Andrew. (2009). "Divorcing the Good and the Right." (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Smart, J. J. C. (1961). "Free Will, Praise, and Blame," *Mind* 70: 291-306.

Springer, Elise. Communicating Moral Concern. Cambridge, MA: MIT Press, 2013.

Tangney, J.P., Niedenthal, P.M., Vowell, Covert M. and Hill, Barlow D., (1998). "Are Shame and Guilt Related to Distinct Self-discrepancies? A Test of Higgin's (1987) Hypotheses." *Journal of Personality and Social Psychology* 75: 256–268.

Tangney, J.P., Stuewig, J., & Mashek, D.J. (2007). "Moral Emotions and Moral Behavior." *Annual Review of Psychology* 58: 345-372.

Tognazzini, N. & Coates, D. J., "Blame," The Stanford Encyclopedia of Philosophy (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2014/entries/blame/>.

Trafimow, D., & Ishikawa, Y. (2012). "When Violations of Perfect Duties Do Not Cause Strong Trait Attributions." *American Journal of Psychology* 125: 51-60.

Trafimow, D., & Trafimow, S. (1999). "Mapping Imperfect and Perfect Duties on to Hierarchically and Partially Restrictive Trait Dimensions." *Personality and Social Psychology Bulletin* 25: 686-695.

Trawalter, S., & Richeson, J.A. (2006). "Regulatory Focus and Executive Function After Interracial Interactions." *Journal of Experimental Social Psychology* 42: 406–412.

Walker, Margaret Urban. Moral Repair: Reconstructing Moral Relations After Wrongdoing. New York: Cambridge University Press, 2006.

Wick, Warner. Ethical Philosophy. Indianapolis: Hackett Publishing, 1983.

Williams, Bernard. Moral Luck: Philosophical Papers 1973-1980. Cambridge: Cambridge University Press, 1982.

Voiklis, J., Cusimano, C., & Malle, B.F. (2014). "A Social-Conceptual Map of Moral Criticism." In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.) *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1700-1705). Austin, TX: Cognitive Science Society.

Young, Iris Marion. Responsibility for Justice. New York: Oxford University Press, 2011.

CHAPTER 4

Transformative Accountability for Structural Injustice

1. Introduction

Philosophers and activists engaged in movements for social justice have consistently argued that racism, sexism, and other such forms of injustice are not reducible to problems in individuals or individual psychologies. In other words, we might say, racism ain't in the head. Instead, racial and other oppressions are complex systems of disadvantage that encompass legal, political, economic, and scientific institutions, media and aesthetic production, as well as social and cultural practices, expectations, and values—in short, they are *structural*. Indeed, the specific content of implicit biases reflects existing social patterns, hierarchies and relations, sometimes with startling sophistication and detail. For example, Dasgupta et al (2009) found that inducing the emotion of disgust increased subsequent IAT measures of bias against gays and lesbians, but not Arabs; conversely, inducing anger increased IAT measures of bias against Arabs, but not gays and lesbians. Wittenbrink, Judd, & Park (2001) found higher levels of implicit bias against Black faces when pictured against the background of an alleyway rather than a church. Similarly, Barden et al (2004) found that a pattern of higher implicit bias against Black faces than Asian faces in a classroom setting was reversed on a basketball court, and that Black faces depicted as prisoners generated higher levels of implicit bias than White faces, but that the reverse was true for when they were depicted as lawyers. The context- and group-specificity of the implicit biases manifested in these studies—in other words, their *sensitivity to social reality*—suggests that implicit bias is not only a product but also a perpetrator of existing social structures.

Sally Haslanger (2015), however, has recently raised a challenge concerning the relevance of implicit bias to the problem of structural injustice. She writes: "[I]f the best explanation of social stratification is structural, then implicit bias seems at best tangential to what is needed to achieve justice. Why the recent emphasis on implicit bias as a solution?" (2) She makes two claims: first, that racism, sexism, and other oppressions should be understood primarily "in terms of unjust and interlocking social structure, not in terms of the actions and attitudes of individuals," and second, that focusing on individual actions and attitudes is ineffective for

dismantling these structures in part because "people are resentful when they are blamed for problems much bigger than themselves [and r]esentful people are resistant to change" (2). As should be clear from earlier chapters, I am in complete agreement with both of these claims. However, I believe that theorizing and implementing critical moral responses to implicit bias—at least when responsibility is conceived of as accountability rather than attributability—is crucial for collectively organizing to enact change. In this chapter, then, I explore the ways in which practices of accountability figure into wider efforts against structural injustice. I begin, in Section 2, by presenting a set of challenges raised by Haslanger and Chad Lavin (2011) about the whether responsibility is a useful concept for rectifying structural injustice. In Section 3, I draw on Iris Marion Young's (2011) "social connection model of responsibility" to defend the claim that practices of moral responsibility for implicit bias are indeed relevant, though it is *accountability* rather than *attributability* that is best-suited to this task. In Section 4, I develop my own conception of accountability as grounded in role-based ideals, which contains an inherently transformative dimension. Finally, in Section 5, I offer a number of concrete examples drawn from Patricia Hill Collins' (2000) four domains of oppression—structural, disciplinary, hegemonic, and interpersonal—to give examples of what actual practices of accountability for implicit bias look like in each.

2. Two Challenges to Responsibility

Chad Lavin (2008) identifies two important problems with the concept of responsibility[76] that exactly parallel Haslanger's worries about focusing on implicit bias. For Lavin, this is due to a fundamental flaw with liberalism, which assumes an individualist ontology that "reifies subjective desire, distracting from systemic coercions and disciplinary mechanisms that lead to events that we arbitrarily attribute to individual actors" (11). Responsibility for Lavin is thus a framework for praising and rewarding actors that ultimately depends on a metaphysics[77] of sovereign individuals whose actions and choices are performed according to their own acts of

[76] To be precise, Lavin refers to this concept as "liberal responsibility". To minimize the proliferation of technical terms, however, I simply refer to it as "responsibility," and it will imminently be clear that the sort of responsibility at stake is attributability.

[77] To be more precise, it is a framework of "ontopolitical" metaphysics that reflects the way in which moral and political theorists like Peter Strawson (1962) and Arthur Ripstein (1999) have reversed the arrow of justification. According to Lavin, they make claims to an individualist ontology not on purely metaphysical grounds, but on the grounds that practices of holding responsible—which trade in this liberal conception of individual autonomy—are practically and politically necessary.)

will. While I can only flag this issue without fully addressing it here, I myself am inclined to take a broadly but minimalist Kantian stance that there is *some* form of individual autonomy—call it the "merely phenomenological" sense, if you will—which is simply a part or precondition of our experience of ourselves as agents making choices in the world, no matter how severe the structural constraints we find ourselves confronted with. This minimalist sense of individual autonomy, I believe, avoids the ontopolitical commitments Lavin is concerned with, and would exist even under radically different practices of responsibility.

The first challenge Lavin raises is that "liberal responsibility focuses on disruptions to established procedure and neglects conditions that have no immediate cause" (14). This is Haslanger's worry about focusing on particular actors and events—what Haslanger (2004) calls *misuses* of power—at the expense of enduring background social structures, which she calls *misallocations* of power. Haslanger's distinction highlights the difference between a particular use of power by some individual occupying a role or position, on the one hand, and the social organization of power that defines those very roles and positions, on the other. To illustrate, she contrasts the example of a professor who routinely gives lower grades to women of color with a university system that does not allow women of color to receive educations at all. For Lavin and Haslanger, blame and punishment for misuses of power may be "immediately satisfying," but they obscure and distract from the more fundamental problems of structural misallocations of power (Lavin 2008, p. 14). They thus claim that ordinary and theoretical discourse about moral responsibility has relatively few resources for grappling with deep and ongoing problems such as poverty and homelessness, racism and sexism, environmental degradation—in short, with structural injustice.

Lavin's second challenge is that "because liberal responsibility focuses on providing particular indictments, it simultaneously provides general exonerations" (14). Again, the focus on identifying blameworthy *misuses* of power too easily lets off the hook all other parties who are implicated in background misallocations of power. Critical moral responses of this sort are thwarted when, as Haslanger (2004) points out, there may be cases where there is no individual misuse of power:

> If power resides in the relationships created by practices, and no individual agent is responsible for a particular practice, then there is an important sense in which the distribution of power may be unjust and yet the injustice not be properly explicated in terms of an agent's wrongdoing. (104)

In these cases, what usually results is what Iris Marion Young (2011) calls a "round-robin" game of "blame-switching", in which parties point to others' actions, rather than own, as the cause of some unjust outcome. As Young points out, however, the nature of structural injustice means that "such blame-switching is particularly easy because others in fact do participate by their actions in the processes that produce unjust outcomes [and] it is difficult to make blame "stick" to anyone in particular" (117). The failure to identify particular misuses of power, says Young, only "paralyzes efforts to address the problems in a forward-looking way, because we are waiting to isolate the parties who should pay for a remedy" (117). This again raises Haslanger's worry that focusing on the bad attitudes and actions of individual actors is pragmatically ineffective.

3. The Transformative Dimension of Accountability

Haslanger and Young's pragmatic arguments about the resentment and defensiveness should sound very reminiscent of the pragmatic argument I gave in Chapter 2 against appraisal-based responses to actions caused by implicit bias. There I argued that such responses in the case of implicit bias were likely to produce resistance in people who cannot introspectively experience themselves as biased, much like people who in acting under structural constraints do not experience themselves as causing unjust outcomes. Young's solution[78], like mine, is to propose a new kind of responsibility which, I claim, is a particular *conception* of what I have been calling responsibility as accountability. Young describes one concept of responsibility—the one that she is rejecting as inappropriate for structural injustice—as "a concept of responsibility as guilt, blame, or liability [that] is indispensable for a legal system and for a sense of moral right that respects agents as individuals" (98). Young claims that determining this kind of responsibility requires "clear rules of evidence, not only for demonstrating the causal connection between this agent and a harm, but also for evaluating the intentions, motives, and consequences of the actions" (99). However, the nature of structural injustice means that it is generally "not possible to identify how the actions of one particular individual, or even one particular collective agent, such as a firm, has *directly* produced harm to other *specific* individuals" (emphasis mine), even if it can be demonstrated that the agent's actions causally contributed to the harm (96). Recall from Chapter 2 that it was also precisely on these grounds that I gave my moral-epistemic argument against ascriptions of attributability in cases of implicit bias: such rules of evidence are

[78] Lavin approvingly cites Young's account as an example of "postliberal responsibility".

conspicuously lacking when it comes to implicit phenomena that are inaccessible to introspection and detection in particular cases. Just like in the case of structural injustice, it is not easy to establish the operation of implicit bias in any given *particular* case though it can be determined at the aggregate level. Clearly, the reason that rules of evidence for evaluating intentions and motives are important here is that the kind of responsibility in play here is responsibility as *attributability*.

Indeed, Young also distinguishes between two "models" of responsibility, which she calls "liability model of responsibility"[79] and the "social connection model of responsibility". Young characterizes the liability model using classic Aristotelian conditions: a person is responsible for harm if we can establish that they caused it knowingly and voluntarily. Note that her Aristotelian conditions serve to establish when an action is properly attributable to an agent: if a person performed some action unknowingly or involuntarily, then that action did not flow from agent's practical identity because it does not reflect any of the agent's intentions or ends. Like me, Young makes clear that she is not replacing the liability model with her new social connection model, but that the latter is needed when the former is inadequate. This is because the liability model and social connection model are particular *conceptions* of attributability and accountability, respectively.

Young stresses that her social connection model is essentially forward- rather than backward-looking: the point, she says, is "not to compensate for the past, but for all who contribute to processes producing unjust outcomes to work to transform those processes" (109). It is backward-looking only insofar as causal investigation is necessary for constructing "an account of how [unjust outcomes] have come about and operated in the past coming up to the present" (109). Young thus answers Lavin's second challenge by dropping this "singling out" function from her social connection model.[80] Here, Young appeals to Onora O'Neill's notion of "assuming" others in our actions. We assume others in our actions when it would be impossible to perform some action without the prior actions of others: when I buy clothing from a shop, for instance, that action presupposes the actions of everyone who grew, picked, and processed the

[79] In her (2005) "Responsibility and Global Labor Justice," she calls this the "blame model."
[80] Young describes it thusly: "The point of locating guilt or leveling blame is precisely to single out: to say that this person, or these people, by virtue of what they have done, bear direct moral and often legal responsibility for a wrong or a crime, whereas other do not because their actions have not done the needs. The practice of blaming or finding guilty requires singling out some from others, and applying some sanction against them or requiring compensation from them" (76-77).

material, everyone who sewed, transported, branded, displayed it, and everyone who negotiated and oversaw the preceding. On Young's social connection model, each of us is responsible not only for the immediate outcomes of our actions, but for all the outcomes affecting everyone we assume in our actions. This is *not* to say that each of us is expected to perform the action of rectifying structural injustice; indeed, the nature of structural injustice renders that impossible. Rather, each of us is assigned the task of *engaging with others* in efforts to transform our shared social condition. Young points out that even the victims of injustice have a role to play here. To put it in my terms: while it would be appalling to judge sweatshop laborers in any way blameworthy for their conditions, we can reasonably expect them to (collectively!) take steps to organize in resistance. She thus provides another excellent illustration of the difference between attributability and accountability.

By endorsing Young's social connection model as a part of a theory of accountability, my account can also respond to Lavin's challenges. However, I extend Young's account by prioritizing certain social connections over others[81]: that is, by returning to the notion of role-based ideals. Since accountability depends on a moral community's system of demands and expectations that distribute duties and burdens, and since morality involves both perfect and imperfect duties, part of what is distributed are the various role ideals that give shape and content to our imperfect duties. By their very nature, ideals already embody the transformative dimension that is needed to address misallocations of power. It is simply part of *striving to do one's job well* to reflect on the purposes and aims of that position, how the position might be modified to better achieve them, what other positions should be created or modified to help, and how to connect with others who are working with similar aims. For example, it is *part of being a good teacher* to think: "What course offerings are we missing? Does my course need another section or graduate teaching assistant? What national and international trends are shaping higher education today, and what political and economic conditions are affecting my students' lives and ability to learn? What committees or local organizations should I serve on in order to address the problems I see?" To take another example—which takes place outside of formal institutionally-defined roles—consider what Patricia Hill Collins (2000) calls the "motherwork" performed by Black

[81] Young herself acknowledges the need for such prioritization by offering a number of "parameters for reasoning" intended to guide our collective efforts: power (influence over structural processes), privilege (e.g. time, resources, and protections), interest (personal experience of being structurally disadvantaged), and collective ability (membership in civil society, e.g. churches, unions, neighborhood associations). I consolidate these parameters—and add a transformative dimension—through my notion of role-based ideals.

women. Collins points out that Black women resist oppression—as an alternative to riskier, more visible forms of activism that require greater time and resources—when they "resist passing on to their children externally defined images of Black women as mules, mammies, matriarchs, and jezebels" and choose instead to "use their families as effective Black female spheres of influence to foster their children's self-valuation and self-reliance" (210). This example of "how political consciousness can emerge within everyday lived experience" is a testament to the myriad ways in imperfect duties to transform unjust social structures can be discharged across a wide variety of social roles (209). While it is *not* reasonable to sanction or specify in detail which, where, and how these particular tasks should be taken up—since they are infinite and we are finite creatures—it *is* reasonable to expect each of us to adopt these ideals and exercise our discretion in choosing how to strive toward them.

My account can also respond to Lavin's first challenge. Accountability practices involving role-based ideals are applicable not only to specific violations of perfect duty at particular times, but also ongoing failures to take up transformative imperfect duties. Because accountability does not depend on appraisals of an agent's intention, motives, etc., it is more or less[82] always appropriate to engage in moral admonition that invokes and promotes actions that will further an agent's role-based ideals. Because social structures occur at all levels of life, and in overlapping ways, the various role-based ideals they furnish will—as with the good teacher and the good mother above—almost always implicate the occupants of those roles in deep and ongoing problems of structural injustice. And, given the complexity of structural injustice and the multiplicity of roles that contribute to its perpetuation, finding that one person is accountable for some aspect of a problem from the standpoint of that one role does not at all exonerate everyone else from the accountability they have in virtue from *their* roles.

It is clear from some of Haslanger's remarks that her concerns also stem from thinking about responsibility as attributability rather than accountability. She writes:

> In some cases social institutions have relatively costless exit options. But even what might seem to be the most malleable practices depend on background expectations and communicative cues that are not within the control of a single individual; so it would be wrong to think of them, except in the rare instance, as created or directed by an individual (or collective) agent. (104)

[82] By this I mean that it is always *pro tanto* morally justified, though it may, for instance, be inconsiderate or insensitive to engage in moral admonition at moments of grief, healing, and so on. It is also warranted as a protection against bad faith, which I discuss briefly in the next section.

Haslanger's concerns about the lack of exit options—in other words, of conditions under which individuals are deprived of (or which render it very difficult to exercise) the ability to do otherwise—reveals that the concept of responsibility she has in mind is that of attributability. Lack of control or the ability to do otherwise undermines attributability (in at least some cases) because, if the action would have been performed no matter what, it (typically) no longer reflects on the intentions or character of the agent. (Frankfurt-style counterexamples are exceptions that prove the rule here.) In my terms, however, it is not quite correct to claim that no individual agent is responsible for these harms. Lack of control does not necessarily block ascriptions of accountability, since it is often still appropriate for the perpetrator to pick up the costs of her action, even if she does not merit any further condemnation. Thus we can say that individual agents—many of them—are accountable for the harm even if it is not properly attributable to them in ways that would sustain appraisal-based responses like blame and resentment. This, as I see it, is a preferable way of putting Haslanger's following point:

> [A]lthough sometimes structural oppression is intentionally caused, say, by policy makers, it is possible for a group to be oppressed by a structure *without there being an agent responsible for its existence or the form it takes*. Admittedly, individuals play a role in creating and maintaining the social world, but most of the practices and institutions that structure our lives, although made up of individuals and influenced by individuals, are not designed and controlled by anyone individually. (104) [emphasis mine]

Even if no one is responsible in the attributability sense, there are still individuals—in fact, the vast majority of individuals—who are responsible in the sense of being *accountable*. Lest we lose sight of the fact that structural reform ultimately *does* come down to individual actions—individual actions performed in concert with others and that have the aim of revising or introducing new structures[83]—we should remember that we are licensed to say it *is* actual individuals who are responsible (read: accountable) for bringing about the needed changes.

4. A Role-Ideal Conception of Accountability

Young might have a different objection to my role-ideal account, however, along the lines of a critique she levels at Robert Goodin. Goodin offers another potential conception of accountability, which he calls the model of "task-responsibility":

[83] I am grateful to Elizabeth Anderson and Simeon Newman for pushing me to clarify this point.

> On the model of task-responsibility, assigning responsibility amounts essentially to assigning different duties and jobs. Different people have different responsibilities, ex ante, because they are allocated different duties and tasks. And people bear differential ex post responsibilities for outcomes, on this account, depending on the role they played or should have played, pursuant to those ex ante task-responsibilities, in producing or averting those outcomes. (109)

But as Young points out, some injustices are produced precisely when people are performing the tasks they have been assigned: when the system is structured such that it is someone's job to, say, find the factory with the worst working conditions and hence the cheapest labor, while it is someone else's job to make public statements that the company is not responsible for what goes on in the factories run by its contractors. Moreover, certain other tasks which need to be performed—such as organizing a labor union from scratch—are ones for which no job yet exists. She thus writes: "Political responsibility in respect to structural injustice, in other words, often requires transforming institutions and the tasks they assign. This is everyone's task and no one's in particular." (386)

On my account, however, what is distributed in a system of accountability is not merely *institutional* duties, tasks, or jobs in Goodin's senses of those terms. According to Goodin, a "duty" is a specific act which an agent is assigned to perform, whereas "tasks," jobs," or "responsibilities" are specific outcomes that are to be achieved, in multiply realizable ways. My understanding of the duties involved in systems of accountability, however, is somewhat different in that it builds in an inherently normative, moralized element. As I described in the last chapter, following Gert (2004), perfect duties represent constraints on acceptable actions, while imperfect duties do not prescribe specific actions but rather action-guiding ideals. Perfect duties thus resemble "duties" in Goodin's sense, but *imperfect* duties are closer to Goodin's "responsibilities." Yet an ideal goes beyond even ordinary responsibilities and tasks, especially when it is a moralized ideal. We might say a moral ideal is a never-ending and limitless "task," because it by definition *extends beyond* existing definitions of the particular task or job at hand. While it is not printed in a teacher's paid job description that she must perform the tasks I listed above, it is *reasonable to expect her*–because it is part of striving to reach the moral ideal of being a good teacher—to do some of these things, some of the time, at her own discretion. Since these ideals are what give content to imperfect duties, they are not merely the mandates of some social institution, but of morality itself. I thus endorse the spirit of Young's remarks when she says: "We all share this responsibility not by virtue of our particular capacities, institutional roles, relationships, or commitments, but

because…a general responsibility for justice accompanies all of our particular roles and responsibilities; it is not something over and above them". (166-167) However, I believe, it is only through our specific institutional roles and social relationships that this "general responsibility for justice" can acquire any kind of action-guiding significance as to what is reasonable to expect of us; the general responsibility is simply part of what it is for us to have role-based ideals.

Moreover, my account of moral admonition—which reminds and exhorts a person when she falls short of those ideals without attempting to evaluate her—spells out a way to achieve Young's aims of avoiding blame while still directing forward-looking efforts to address structural injustice:

> We who share responsibility ought to take action, but it is up to us to decide what is reasonable for us to do, given our abilities and our particular circumstances. I have argued above that we should not be blamed or found at fault for what we do to try to rectify injustice, even if we do not succeed. As I will discuss shortly, however, we can and should be *criticized* for not taking action, not taking enough action, taking ineffective action, or taking action that is counterproductive. We also have a right and an obligation to criticize the others with whom we share responsibility. (144)

In allowing agents to exercise their own discretion in determining what are reasonable ways for them to strive for their ideals, we always face the potential of bad faith, in which people's own interpretations of what is reasonable align more toward self-serving than idealistic purposes. This is another reason—in addition to the fact that education and motivation are always in need for the ongoing project of structural transformation—why moral admonition for failures of imperfect duty are, as a rule[84], always justified. We are usually not in a position to know whether someone's failure to perform some transformative imperfect duty on a particular occasion is motivated by bad faith; hence, appraisal-based responses like blame and resentment are generally unwarranted. But neither are we in a position to know that it was *not* a case of bad faith, so evoking the ideals involved in moral admonition serves to enlist the forces of moral motivation as a check on bad faith.

On my role-ideal model of accountability, the precise details of what each agent can reasonably be expected to do will differ according to her particular circumstances: it is likely to begin with individually articulating and researching the problem, to communicating those findings to others, to finding collective courses of action that can be taken to address it, to

[84] See Fn. 83.

showing solidarity with others who are undertaking such actions. Moreover, system transformation is not a one-shot deal. It is difficult to conceive of any time in the near future at which further reform would not be needed; and even if we someday managed to instantiate an ideal society as dictated by the ultimately correct principles of justice, there would always be a need to monitor and maintain the stability of such a society in the face of an unpredictable natural and technological world. So even though there might be "jobs" that could be created to address failings in the current system, and more to address failings in the next improved version, and so on—still, there will always be a need for something further. What this means, I have suggested, is that we should understand transformative accountability as task-bound; that is, as grounded in the *internal* standards specific to each task or role. The reason for this is that taking on any one role is not to cease to be a moral agent within a broader moral community.[85] To be committed to a moral norm or principles of justice just *is* to be committed to be committed to the task of reforming a society until it conforms to those principles. Alternatively, we could also say that morality is the ever-present, all-encompassing "institution" within which each of us is assigned the inescapable task of upholding and achieving it, and which grounds transformative accountability.[86]

Before concluding this section, let me just point out that attributability and accountability are more closely linked on my account than on Young's, in at least two ways. Although we both agree that attributability essentially involves causal relations (and other metaphysics), and for that reason look for ways of moving beyond it, attributability for me can—though it *need* not—still have an important role to play in responding to injustice. For recall that even on what I called the "accountability route"—where our starting point is a political, not metaphysical, problem of distributing burdens and duties—we can still arrive *later* at questions of attributability once we

[85] Cf. also Onora O'Neill (1986), who writes: "Kantians generally would play down any distinction between a person's own responsibilities and his or her role responsibilities. They would not deny that in any capacity one is accountable for certain things for which as a private one is not accountable. For example, the treasurer of an organization is accountable to the board and has to present periodic reports and to keep specified records. But if she fails to do one of these things for which she is held accountable she will be held responsible for that failure—it will be imputable to her as an individual. When we take on positions, we add to our responsibilities those that the job requires; but we do not lose those that are already required of us" (291).

[86] Thinking of morality as itself a social institution is most congruent with what Elizabeth Anderson (2014) has called a "naturalistic" account of morality, on which morality is primarily a system of coordination of demands, expectations, and values. This conception would not require the existence of some "external" moral standard by which we measure our existing society. However, it is consistent with my account here to hold that such an external standard of morality does exist, and consequently that transformative accountability is grounded in discrepancies between reality and the standard. But I do not intend to take up any particular meta-ethical stance here.

have already settled the accountability question of divvying up the burdens of dealing with the harmful consequences of some action. Once the appropriate non-appraising responses have been undertaken, we can ask the *further* question whether the agent should be blamed for her violation or failure of duty, and on my own non-consequentialist view this will force us to revert back to questions of attributability. As I put it in Chapter 2, questions of attributability and accountability represent two different ways of getting to the appraisal-based responses traditionally associated with moral responsibility. I am thus open to the use of appraisal-based responses such as blame and condemnation if they can be shown to be warranted, and if they are not counterproductive. It is possible that these conditions can be contexts of structural injustice, e.g. where paper trails make it possible (and politically effective) to demonstrate the racist attitudes and decision-making policies of collective agents such as police departments and governments. The other way in which attributability and accountability are linked, as I emphasized in Chapter 3, is through the criterion of reasonable expectation: ascriptions of attributability for some action depend on whether an agent has sufficiently lived up to the imperfect duties for which she is accountable. My account thus provides an explanation of how the two concepts of responsibility are related, and why they are so often conflated; on Young's, it is less clear why such wholly different concepts—one backward-looking and one forward-looking, one involving praise and blame and the other involving duties to act—should be more than accidental homonyms.

5. Implicit Bias in the Matrix of Domination

Haslanger claims that "although there is space for attention to implicit bias in social critique, it is only a small space" (11). While she admits that we may be responsible for mitigating bias in ourselves, she argues that our vulnerability to social structural realities makes those efforts unlikely to succeed if they are not matched by structural change. She thus concludes:

> Social change requires contestation, organization, and activism. Working to correct our own biases may be a minimum requirement on us. But we are each complicit in the perpetuation of unjust structures, practices, and institutions. Moral responsibility concerns not only what I can and should do, but also what we can and should do together. (12)

My role-ideal model of accountability—on which we can reasonably be expected to strive toward role-based ideals in order to collectively organize for structural change—is clearly compatible with, and indeed entails, this claim. However, Haslanger's own understanding of social structure permits a greater role for considerations for implicit bias than she seems to realize. Explanations

involving implicit bias do not need to *compete* with structural explanations[87], since implicit bias is not only enabled by but is itself an *enabler* of structure: it is the grease that oils the machine, so to speak.

To see this, consider William Sewell's (1992) theory of social structures, which Haslanger explicitly endorses. One of Sewell's primary desiderata for a theory of structure is that it should build in the possibility for structural change. For Sewell this is made possible by the "duality of structure," in which "human agency and structure, far from being *opposed*, in fact *presuppose* each other" (4). Agency, for Sewell, consists first and foremost in knowledge and control over the "schemas" and "resources" that constitute social structures, where genuine knowledge and control necessarily entails the ability to use those schemas and resources in novel and creative ways (20). While I hesitate to push too far into the weeds here, let me just clarify that schemas are culturally shared rules, norms, assumptions, and metaphors—which, as Sewell emphasizes, are "not always conscious"—that can be transposed onto a variety of situations and hence allow us to make meaning out of these situations (8). Resources are simply the material objects[88] that are required to enact schemas in the actual world. For example, our shared "classroom" schema enables the smooth enactment of situations in which the professor goes to the front of the classroom, talks and writes on the board, while the students write down notes, raise their hands, and ask questions—an enactment made possible by the resources of the classroom, chalk and chalkboard, desks, pens, etc. It is the repeated exercise of agency—the professor who knows to pick up the chalk and writes on the chalkboard and the students who know to write down notes with their pens—that give existence to this particular structure. But the knowledge of schemas that constitutes agency also allows for the possibility of exercising it in new ways that can challenge existing configurations of structures and resources, e.g. when a professor decides to hold her class in a coffee shop in order not to cross the picket line of striking graduate student instructors.

Implicit associations[89], on this picture, can be understood as one of the cognitive *mechanisms* by which agents have and make use of their knowledge of cultural schemas. Given the

---

[87] Some of them do, to be sure, as Haslanger clearly demonstrates in her analysis of implicit bias discourse as a source of "Standard stories".

[88] Sewell also discusses non-material, human resources such as e.g. intelligence, strength, skills, willpower, etc. For simplicity's sake I omit these from my discussion.

[89] My shift in terminology here is meant to preserve my preferred usage of "implicit bias," which restricts it to those implicit associations that involve social categories and negative or stereotypic traits.

overwhelming number of overlapping social structures in which we are constantly enmeshed, implicit social cognition makes it possible to navigate, intuit, and respond to the nuances of the structures at play in any given situation without overwhelming our limited cognitive resources. Implicit associations (biases) are thus part of what make agency possible, and as Sewell repeatedly emphasizes, agency and structure mutually entail one another. A concern to understand and address implicit bias, in other words, *is* a concern to understand and address the workings of social structure.

How exactly does agency lead to structural change? Sewell lists five axioms, two of which I will mention here: the *multiplicity* of structures, which "exist at different levels, operate in different modalities" and across diverse religious, economic, aesthetic, epistemic, and other such domains, and the *intersection* of structures, which allows for contestation between different agents seeking to make different uses of the same schemas and resources. When agents have command of this vast repertoire of structures and are capable of deploying them in creative and contesting ways, this means that "reproduction [of structures] is never automatic," and "structures are at risk, at least to some extent, in all of the social encounters they shape" (19). In just the same way, my role-ideal model of accountability exploits these features of structure by capitalizing on the way in which agents always simultaneously occupy numerous roles—worker, parent, sibling, friend, neighbor, citizen—which are multiply realizable in service of different ideals that guide structural transformation.

With this picture of structural change in hand, let me return now to the problem of moral responsibility for implicit bias. Understanding implicit bias as the product of interlocking social structures requires that we develop practices of moral responsibility for implicit bias that can operate at all of the different levels of structure pervading our social lives. Given the innumerable of overlapping social structures that are always present, however, it will be useful to organize them somehow. My organizing principle is Patricia Hill Collins's (2000) "matrix of domination," an overarching framework for "understanding how power is organized and operates," within which Collins explicitly embraces both the work of individual consciousness-raising and institutional transformation (274). Collins identifies four domains of power, along with the strategies for resistance that are most effective in each domain. She summarizes them as follows: "The structural domain organizes oppression, whereas the disciplinary domain manages it. The hegemonic domain justifies oppression, and the interpersonal domain influences everyday lived

experience and the individual consciousness that ensues" (276). Against the background of Collins's matrix of domination, then, let me present a concrete example from each domain. Each exemplifies the kind of critical moral response—of institutional and interpersonal practices of accountability—that I am advocating in response to implicit bias: response that constitutes or has the potential to spur structural transformation. As I hope these examples will make clear, practices of accountability for implicit bias are in fact crucially necessary for exactly the sort of "contestation, organization, and activism" that Haslanger is calling for.

5.1 Implicit Bias in the Structural Domain

The structural domain, which focuses on "large-scale, interlocking social institutions" including "the U.S. legal system, labor markets, schools, the housing industry, banking, insurance, the news media, and other social institutions as interdependent entities," is probably the domain that features mostly prominently in ordinary understandings of structural injustice (277). Collins explains that this structures in this domain "yield only grudgingly to change…in part when confronted with wide-scale social movements, wars, and revolutions that threaten the social order overall" (277). A prominent contemporary example of this kind of "visible social protest" is the #BlackLivesMatter movement, which arose in response to the police shootings of unarmed Black teenagers Trayvon Martin and Michael Brown, amongst many others. As protesters marched on the streets in Ferguson, MO every night for months after Brown's death, a number of youth-led organizations called for connected their work to larger patterns of not only national, but international patterns of state violence—sharing the stage with labor, faith, community, and Palestine solidarity activists—and called on the weekend's participants to go back to their cities, take to the streets, and organize locally. On the night of the announcement that the officer who shot Michael Brown would not be indicted, protesters participated in direct actions across 37 states.

What is remarkable about this storm of activism is the extent to which the guiding principles of the movement have embodied the logic of accountability I have developed here. In one sense, #BlackLivesMatter can read as a critical moral response to the explicit and implicit bias that enables the enactment of racially discriminatory policing and criminal justice structures; this is evidenced by the substantial media coverage not only of racist emails leaked by the Ferguson police department but also of research on implicit racial bias in police and juries. (A

large empirical literature has demonstrated the role of implicit bias in racial violence and policing.[90]) However, the #BlackLivesMatter movement has responded to implicit racial bias by evoking *structural* explanations and ideals. A number of their actions have had the goal of "disruption," where the symbolic significance of disruption is that it targets a *system* in addition to—and indeed, prior to—any particular wrongdoer. In this sense the actions call for not just the application of existing norms but a *transformation* of them. This is evident in the protesters' chants[91], which have coalesced into something of a standard repertoire used across the country:

> "Hands Up, Don't Shoot"
> "I Am Mike Brown"
> "What Do We Want? Justice/When Do We Want It? Now"
> "No Justice, No Peace (No Racist Police)"
> "Hey Hey, Ho Ho, These Killer Cops Have Got to Go"
> "This Is What Democracy Looks Like"
> "Ain't No Power like the Power of the People 'Cuz the Power of the People Don't Stop"
> "Indict. Convict. Send That Killer Cop to Jail/The Whole Damn System is Guilty as Hell"

While a few of the protesters' chants (overrepresented in a sensationalist media) single out "killer cops" or "racist police," they overwhelmingly appeal to universally shared moral and political ideals of *justice* and *democracy* even as they explicitly and implicitly call out the systematic failure of social institutions to uphold those ideals. "Hands Up, Don't Shoot" is a moral admonishment that invokes the role-based ideal of policing, and whose implicit message is that good police officers do not shoot nonviolent persons. And the near-truism that "Black Lives Matter" would seem to be obvious in a regime promising liberty and justice for "all," but the very fact that it does *not* go without saying—that *something* is wrong with the current state of affairs—is an implicit rebuke that thrusts forward a compelling although not yet fully specified moral concern. The demand to indict police offers involved in deadly shootings is clearly articulated, but the more general concern that the system does not treat Black lives as if they matter is more inchoate. Yet this is still a communication of moral concern in Springer's (2013) sense. She writes:

> At one extreme, a concern may seem well-captured by fully articulate claims; at the other extreme, we may be perplexed even about which open-ended claims might be nailed

[90] For a review, see Richardson and Goff (2014).

[91] These are taken from my observation of marches which took place during the #FergusonOctober weekend of action on October 10-13, 2014, as well as a number of marches organized by #aa2ferguson in the streets of Ann Arbor which took place between December 2014 and May 2015.

> down around its edges. Perhaps we are reduced to vagueness: "Something is wrong," or "Something is odd in her tone of voice." (138)

Part of this vagueness, I claim following Springer, is due to the fact that a single concern can be taken up in different ways by different occupants of different roles. This is certainly true in the case of racial biases—explicit and implicit—in policing, which many commentators have pointed out is the result of a confluence of political, economic, and cultural structures such as the mass incarceration of men and women of color, racially segregated and impoverished neighborhoods, and images depicting Black and other men of color as dangerous "thugs". Across these multiple and intersecting domains, many people in many different roles have opportunities to carry out transformative imperfect duties that would contribute to structural change. By recognizing the "millions [who] have answered that call with simple acts of civil disobedience" and "small actions that when woven together, have tremendous impact,"[92] #BlackLivesMatters organizers who have called for nationwide, local organizing have performed exactly the ideal-centered, role-centered collective actions that I have advocated.

5.2 Implicit Bias in the Disciplinary Domain

The disciplinary domain of power, according to Collins, "manages power relations…not through social policies that are explicitly racist or sexist, but through the ways in which organizations are run" (280). It thus encompasses a variety of formal and informal practices that may serve as barriers to equality even though legal and organizational polities are formally nondiscriminatory: in the workplace, for example, these may include evaluation criteria for hiring and promotion that are linked to specific social stereotypes, social networks and informal socializing that tend to exclude women or minorities, and unspoken culture- or class-based social norms in the workplace. It is overwhelmingly likely that many of these forms of disciplinary processes embody or rely on implicit biases that, in generating institutional inertia, function to impede the restructuring that is demanded on paper by legislative directives from the structural domain. According to Collins, addressing injustice in the disciplinary domain requires "insider resistance" that "tries to capture positions of authority within social institutions in order to ensure that that existing rules will be fairly administered and, if need be, to change existing policies"

[92] I take these words from the "About This Movement" of the Ferguson Action website, accessible at http://web.archive.org/web/20150609063336/http://fergusonaction.com/movement/.

(282). Collins calls this "working the cracks," which she explains as follows: "From a distance, each egg appears to be smooth and seamless, but upon closer inspection, each egg's distinctive pattern of almost invisible cracks become visible…represent[ing] organizational weaknesses" (281-282).

One example of efforts to enact change in disciplinary structures can be found in the faculty committee reports of two Canadian universities that have recently undertaken group salary adjustments in order to redress gender pay inequities. The 2012 report of the University of British Columbia Gender Pay Equity Recommendation Committee states that "the unexplained female pay disadvantage [after factoring in gender differences in rank, distribution across high- and low-paying departments, and years of experience] of about $3000 can be considered discriminatory under the assumptions that male and female faculty members are equally productive" (Boyd et al 2012, p. 2). Moreover, the 2014 summary report of the Joint Committee on Gender Pay Equity at the University of Victoria reads: "Given that academic staff is presumed to be equivalent in all attributes relevant to determining the rate of pay, *any* systemic salary differential between men and women raises concern" (Adjin-Tettey 2014, p. 7). However, the report later makes the disclaimer it does not distinguish between "three competing hypotheses" that either "historical discriminations in starting salaries" or "gendered discrimination in performance," or both, caused gender wage gaps wherein men began with lower salaries than women until they reached 6.5 years of experience (the likely point of achieving tenure), after which women received increasingly lower salaries with increasing years of experience. The University of British Columbia subsequently implemented a 2% across-the-board pay increase for all tenured and tenure-track women faculty, while the University of Victoria implemented a modified group award in which all men faculty with less than 6.5 years' experience and all women faculty with more than 6.5 years' experience received a linearly adjusted pay correction.

Let me make three points about these universities' taking responsibility (as accountability) for their gendered pay inequities. First, as evidenced by explicit commentary in university reports of gender pay differences, establishing the existence of bias is methodologically (and perhaps politically[93]) fraught. Aggregate-level data can reveal the operations of disciplinary structures that

[93] Comparable committee reports at the University of California- San Francisco and the University of Minnesota explicitly refrain from the term "inequity" in favor of the term "imbalance" "until such time as any salary differences between groups could not be explained by non-discriminatory legitimate business practices of the University or

serve to disadvantage women—structures that it would be plausible to believe are at once enablers *of* and enabled *by* implicit gender bias. However, if such implicit gender bias does exist, addressing it will require action not just by individuals but by collective agents such as universities that are better placed to manipulate disciplinary structures. Without moving to the collective level, many particular actions by particular people cannot even be described as wrongful actions in isolation. To make the point more clearly: while any hiring decision will be harmful to someone—namely, all the candidates who are not chosen—only those candidates who were rejected on the basis of irrelevant factors such as race, gender, etc. will have been wrongfully harmed.[94] Thus even the establishment of wrongdoing itself which is a necessary condition for critical moral response, much less whether that wrongdoing was caused by implicit bias, may already be missing at the level of individual cases. What this case demonstrates—an upshot of the duality of structure—is that focusing on implicit bias actually demands structural analysis.

Second, the practical effect of group-level remedies is that it more easily avoids the possibility of needing to engage in appraisal-based responses that require the attribution of blameworthy action. The University of British Columbia report noted that the case review method of remedy (case-by-case decisions) "often puts the onus on the individual," lead to small numbers of salary adjustments and "ineffective or poorly executed" outcomes. Given the identifiability of individual cases that could lead to hostility, defensiveness, or retaliation, in addition to the espistemic difficulties noted earlier, this should come at no surprise. By contrast, aggregate-level analysis exposing biased outcomes can be more easily absorbed by the collective agent.

Finally, the case of gendered salary inequities expose the need for *transformative* in addition to restorative measures of accountability. The University of British Columbia committee report points out that "in all cases, universities have acknowledged that a salary correction mechanism does not prevent the problem from reappearing" (Boyd et al 2012, p. 7). They cite the case of the University of Manitoba, which implemented a group salary award but failed to carry out other policy recommendations intended to prevent future inequities; a follow-up study found that salary inequities "reemerged to nearly the same degree" (Boyd et al 2012, p. 6). The committee

---

campus unit," or insist that the results "cannot be considered to provide a *proof* that there is a gender gap in salary" (Office of Academic Affairs and Faculty Development and Advancement 2015; Clayton 2013). The tone of such reports suggests that various forms of insider resistance may well be needed in order to transform the disciplinary domain.

[94] I am indebted to Sarah Buss and Ishani Maitra for discussion of this point.

thus emphasizes the need to take measures such as performing recurring studies to monitor and report salary inequities, training for appointment, tenure, and promotion committees "in systemic bias…and institutional expectations with respect to employment equity," disclosure of starting salary information to shortlisted job candidates, mentoring, and the creation of a senior advisory position. A line of research by Kalev, Dobbin, & Kelly (2006), Dobbin & Kalev (2007), and Kim, Kalev, & Dobbin (2012) indicates this last recommendation may be particularly effective, since they have consistently found that "structures establishing responsibility" (2006, p. 590) and "programs that establish clear leadership and responsibility change" (2007, p. 280) especially when they "engage managers in finding solutions" rather than "treat managers as the source of the problem" (2012, p. 171) in corporate and university settings lead to greater increases in diversity than training or mentoring programs. On my view, these results are explained by the fact that creating a new institutional position also establishes a new role-based ideal (or else incorporates the goal of promoting diversity into an existing role-based ideal), thereby generating new resources for grounding practices of accountability.

5.3 Implicit Bias in the Hegemonic Domain

The primary function of the hegemonic domain is to *justify* systems of oppression. Collins identifies "school curricula, religious teachings, community cultures, and family histories" along with the "growing sophistication of mass media" as examples of institutions with the "ability to shape consciousness via the manipulation of ideas, images, symbols, and ideologies" (284-285). She also introduces the concept of "controlling images," that is, portrayals of social groups depicting them in ways that serve to justify racist, sexist, and other oppressive ideologies. The hegemonic domain, then, can be read as the "home" domain of implicit biases, whose rich content reflects various controlling images based on race, gender, sexuality, disability, and other social identities, as produced and reproduced by the combined operations of educational, cultural, aesthetic, and epistemic structures. How can such hegemonic structures such as controlling images be transformed? Collins writes:

> Racist and sexist ideologies, if they are disbelieved, lose their impact. Thus, an important feature of the hegemonic domain of power lies in the need to continually refashion images in order to solicit support for the U.S. matrix of domination. Not just elite group support, but the endorsement of subordinated groups is needed for hegemonic ideologies to function smoothly. (284)

Collins thus emphasizes the "long-standing existence of Black women's resistance as expressed through relationships with one another, the Black women's blues tradition, and the voices of contemporary African-American women writers," along with the motherwork I mentioned earlier—all of which are forms of Black feminist *self-definition* which that are "highly empowering because they provide alternatives to the way things are supposed to be" (286). A recent vivid example of this is described by Laverne Cox, an openly transgender Black woman who stars in the television show *Orange Is the New Black*. After a series of campus appearances, Cox recounts, "So many students have said—trans students have said: Now I can have a point of reference when I talk about who I am. My friends are like, 'Oh, like Sophia from 'Orange is the New Black?'' and they're like, 'yeah,' and then they just move on and it's not an issue" (Midgarden 2015). By embodying a visible character and public persona, Cox created a culturally shared *image* of a trans woman of color where none had existed. Similarly, when President Barack Obama—to the surprise of many—spoke in the aftermath of Trayvon Martin's death using his own experiences of being under surveillance as a Black man, adding "You know, if I had a son, he'd look like Trayvon," his personal disclosure reframed and refocused the public image of Martin, implicitly admonishing those who failed to see Martin as a boy and a son. Public figures, and the collective agents such as media networks and production companies that support them, thus have an especially important role in exercising their agency to challenge hegemonic structures and replacing them with new and empowering "possibility models"[95].

However, public figures are not the only ones with such agency, though they may have more resources as their disposal. Consider the scores of people who posted pictures of themselves wearing hoodies and declaring "I Am Trayvon Martin," and the Twitter hashtag #IfTheyGunnedMeDown in which people juxtaposed pictures of themselves with hoodies, alcohol, and cigarettes side-by-side with pictures of themselves as college students and military officers in order to rebuke those in the media and general public who subscribed to controlling images of Trayvon Martin and Michael Brown (NPR 2014). By pooling their individual resources into social media aggregators, these people also exercised their agency—collectively and powerfully—to disrupt the hegemonic structures enacted by media portrayal of Trayvon Martin and Michael Brown as deserving to be killed. The logic of such practices lies not in appraisal-based responses that blame or shame, but in holding up alternative images that prompt

[95] Cox states that chooses to refer to herself as a "possibility model" rather than a "role model" because she does not necessarily want other to follow her life choices, but rather to see the possibilities available to them (Couric 2014).

viewers to detect in themselves the discrepancy between their actual and ideal selves as exposed by their immediate perceptions. This development and proliferation of counterstereotypical images—the most direct method of unlearning implicit biases—serves simultaneously to transform hegemonic structures of image and metaphor.

5.4 Implicit Bias in the Interpersonal Domain

The interpersonal domain "functions through routinized, day-to-day practices of how people treat one another" (288). Collins writes: "Because the interpersonal domain stresses the everyday, resistance strategies within this domain can take as many forms as there are individuals" (288-289). My example here, then, is chosen to reflect the complexities and vagaries involved in the kind of activism and organizing against injustice that Haslanger is interested in: a current debate in (online) anti-oppression circles concerning what has been called "call-out culture". To "call out" a person is to publicly point out how something she has said is racist, sexist, heterosexist, ableist, or otherwise oppressive. Given that the practice evolved in anti-oppression spaces, the oppressive attitudes expressed in such behavior is typically the result of implicit biases on the part of participants who are consciously and sincerely attempting to unlearn them. Calling out is typically a non-appraising form of accountability that performs valuable educational, communicative, and motivational functions for all parties: it educates the speaker about problematic language, demonstrates to others in the space that even implicitly racist, sexist, etc. attitudes are not acceptable, and motivates people to "check their privilege" by being mindful and reflective about the various forms of social advantage and disadvantage they all possess. Calling out is also an important way in which a relatively privileged person can be an "ally" to those with a social identity with respect to which she is privileged, since her taking on the work of calling out others—especially those who share the same privilege—relieves some of the burden that would otherwise fall to those directly affected.

A recurring concern[96], however, is that a "call-out culture" has developed in which calling out has become a form of "liberal bullying" (Stallings 2012). Arguably the most high-profile—and most extreme—manifestation of call-out culture occurred when a senior director of corporate communications tweeted a joke about AIDS before a 11-hour flight to Cape Town, during which time tens of thousands of Twitter users had called her out, ultimately costing her

[96] See "Privilege-checking and Call-out Culture" (2012) for a partial list of blog posts and commentary. Here I focus on a number of prominent recent posts.

that job (Uprichard 2012). In another case, a developer at a conference called out two men she heard making lewd jokes behind her, and was herself fired from her job (Romano 2013). In a recent blog post, Asam Ahmad (2015) writes:

> One action becomes a reason to pass judgment on someone's entire being, as if there is no difference between a community member or friend and a random stranger walking down the street (who is of course also someone's friend). Call-out culture can end up mirroring what the prison industrial complex teaches us about crime and punishment: to banish and dispose of individuals rather than to engage with them as people with complicated stories and histories.

In other words, calling out can devolve into *appraisal-based* responses which are used to "police and define the bounds of who's in and who's out" (Ahmad).

This criticism of call-out culture, however, has met with resistance from other quarters. In particular, such critics of call-out culture have been accused of "tone policing". Tone policing, according to blogger NinjaCate (2013), can be defined as follows:

> It is the act of disregarding the substance of someone's argument by focusing on the way it was conveyed. A tone argument focuses on delivery as a means to sidestep the issue at hand. It is a derailment. The reason tone policing is so problematic is that it implies that emotion and logic are mutually exclusive entities. However, it is absurd to expect that people who are discussing their oppression remain calm in the face of challenges to their humanity. Tone policing privileges the feelings of the (implied) bigot, over the humanity of the minority party. It literally requires that the oppressed minority prioritize the majority member's feelings and comfort while fighting to have their own humanity be recognized.

The objection here is that the pushback to call-out culture amounts to a form of tone policing, that is, of (privileged) people not being able to tolerate being held accountable for their mistakes and unfairly asking (marginalized) people to suppress their legitimate feelings of anger and resentment.

I believe these are important conversations to have, notwithstanding the frustration, talking past one another, and lack of closure that is apt to be felt on all sides. Of course, whether a given instance of calling out is really bullying, or whether some critique of a call-out is really tone policing, is not only a matter of considerable discernment but also subject to multiple interpretations about which reasonable people will disagree. Consider "white fragility," which Robin DiAngelo (2011) defines as "a state in which even a minimum amount of racial stress becomes intolerable, triggering a range of defensive moves," and "racial battle fatigue," which

Smith, Yosso & Solorzano (2006) describe as the "stress of unavoidable front-line racial battles in historically white spaces lead[ing] to people of color feeling mentally, emotionally, and physically drained" (301). White fragility and racial battle fatigue are two sides of the same coin: the inevitable product of longstanding White supremacist structures that permeate even well-intentioned cross-racial interactions. It is thus unsurprising that even anti-oppression spaces—where all parties explicitly affirm their commitments to social justice work—are still fraught with such conflicts and contradictions. On the other hand, both critics and critics of the critics (as it were) demonstrate substantial agreement, especially in practical how-to guides, on a number of points which I will highlight here. First, there exist "compassionate and creative" call outs that condemn the behavior and not the person (Ahmad 2015). Second, call outs ought to be taken as such by the person being called out, i.e. as "not a personal attack" but a "way for people to educate others on how systems of oppression operate on a day to day, individual level" ("How to Deal"). Third, call outs require significant emotional resources and labor on the part of the person performing them, in which people "often have to mentally prepare for serious repercussions" ("How to Deal"). And finally, call outs should be performed with serious attention to "what a call out is meant to accomplish" (Ahmad 2015); those calling out ought to be "more interested in helping them change their oppressive behavior than publicly shaming them for it" (Ferguson 2015). In a critical response to Ahmad, Kitty Stryker (2015) writes:

> Call outs are, in my opinion, fundamentally an example of caring about people, as to call someone out is to trust that they will hear your feedback and want to change. To be called out is, in my mind, indicative of people's belief in you, that you're worth improving. It's the opposite of banishment."

These remarks make it clear that call outs are indeed intended to be *non-appraising* critical moral responses in groups working to enact what Carl Boggs (1977) termed "prefigurative politics," or the "the embodiment, within the ongoing political practice of a movement, of those forms of social relations, decision-making, culture, and human experience that are the ultimate goal" (1977). Indeed, (successful) call outs are clearly examples of moral admonition: moral criticism that focuses on actions and not actors, and that point out when people fall short of moral ideals by communicating and exhorting them to fix their mistakes. This debate around call-out culture demonstrates that addressing implicit bias is *particularly* important in activist efforts to organize for structural change. Moreover, it illustrates how attention to implicit bias in anti-oppressive spaces has resulted in the ongoing and real-time evolution of new critical moral responses that anticipate

the structural changes they seek to achieve. As philosophers theorizing about moral responsibility, but also as agents seeking to change social structures, we would thus do well to pay close attention to these on-the-ground projects of developing practices of accountability for implicit bias.

6. Conclusion

In this chapter, I have argued for a conception of responsibility as accountability grounded in role-based ideals that give content to our transformative imperfect duties to collectively organize for structural change, and I have argued that this conception of transformative accountability can handle a number of challenges arising from the nature of structural injustice that responsibility as attributability cannot. I have also defended the importance of understanding and addressing implicit bias as part of the project of combating structural injustice, on the grounds that such biases do not compete with social structures to explain injustice, but rather serve as one of the cognitive *mechanisms* through which—that make it even possible for agents to—continually enact and re-enact unjust social structures. Finally, I have provided a number of concrete examples of non-appraising critical moral responses to implicit biases that function in this way.

To conclude, let me emphasize how my appeal to role-based ideals amplifies the force of Young's claim that addressing structural injustice—the root cause of pernicious implicit bias—is "everyone's task and no one's in particular." Since these ideals are distributed through a system of demands and expectations across a moral community, responsibility as accountability is essentially *collective*. This collective dimension is crucial for addressing structural injustice because the mutual entanglement of economic, political, cultural, aesthetic, and epistemic institutions cannot be unbound by loosening any one particular knot. Achievements in one domain, like the encoding of equal rights and liberties into law, must be matched by progress in other domains, like the elimination of social stigma and group stereotypes. Changing well-entrenched social structures requires multiple—indeed, innumerable—actions taken from diverse roles and positions across a society to transform its structures. What I want to emphasize here is that just the same can be said of moral norms, which serve as one impetus[97] for the transformation of

[97] Structural change can also be occasioned by changes in environmental conditions, technological advancement, etc.

social structures. As Elizabeth Anderson (2014) writes in her analysis of the social movement as an "engine of moral progress":

> A norm can be sustained only if most people believe that most others accept its practical authority. Once enough people demonstrate their repudiation of its authority, even in the face of official sanctions, others who have acquiesced in the norm only from unreflective habit or the expectations of others may waver in their support...Moral reasoning, to be effective in changing social practices, must be done together. The mobilization of large numbers of people supplies a condition for practical social change that individual conscientious reflection does not. (14)

This reveals a deep fact about morality itself: it too is essentially collective. It has the function, as Anderson puts it, of "govern[ing] our interpersonal claim-making." This aspect of morality is best captured in the concept of moral responsibility as accountability, not attributability. To be sure, questions of attributability are morally relevant because the sorts of traits and intentions we ascribe to agents are morally relevant; but recall that attributability arises primarily from questions of metaphysics and philosophy of action. Accountability, on the other hand, is fundamentally concerned with the moral division of labor, with the collective system of interpersonal claim-making. This explains why accountability is naturally suited to the structural problems that concern Haslanger and Lavin. Role-ideal accountability captures a dimension of moral responsibility that is irreducible to individual reasoning and action: one that requires us to reason and act in concert with others around us to collectively change our shared social conditions.

References

Adjin-Tettey, Elizabeth, Aragon, Janni, Brown, Leslie, Hallgrímsdóttir, Helga Kristín, Lesperance, Mary, and Robert Lipson. (2014). "Summary Report of the Joint Committee on Gender Pay Equity at the University of Victoria." University of Victoria. Retrieved from<http://web.archive.org/web/20150318193814/http://www.uvic.ca/vpacademic/assets/docs/genderpayequity/FINALGET_SummaryJuly14th2014.pdf>

Ahmad, Asam. "A Note on Call-Out Culture." Briarpatch Magazine. 2 March 2013. <http://web.archive.org/web/20150319204800/http://briarpatchmagazine.com/articles/view/a-note-on-call-out-culture>.

Anderson, Elizabeth. "Social Movements, Experiments in Living, and Moral Progress: Case Studies from Britain's Abolition of Slavery." The Lindley Lecture, The University of Kansas, February 11, 2014.

Anderson, Elizabeth and Pildes, Richard. (2000). "Expressive Theories of Law: A General Restatement." *University of Pennsylvania Law Review* 148: 1503-1575.

Boggs, Carl. (1977). "Marxism, Prefigurative Communism, and the Problem of Workers' Control." *Radical America* 11(6)/12(1): 99-122.

Boyd, Lara, Creese, Gillian, Rubuliak, Deena, Trowell, Mark, and Claire Young. (2012). "Report of the Gender Pay Equity Recommendation Committee." University of British Columbia. Retrieved from <http://web.archive.org/web/20150318193529/http://www.facultyassociation.ubc.ca/docs/news/GenderPayEquity_UBCV%20Report%20Recommendation%20Committee2012-07.pdf>.

Clayton, Murray K. (2013). "Updated Analyses of Gender Equity in Salaries of Faculty at the University of Minnesota--Executive Summary." University of Minnesota. Retrieved from: <http://web.archive.org/web/20150318192921/http://www.academic.umn.edu/provost/vpoffices/documents/UpdatedAnalysesofGenderEquityinSalariesofFacultyattheUniversityofMinnesota-ExecutiveSumm_000.pdf>.

Collins, Patricia Hill. Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. New York: Routledge, 1999.

Cox, Laverne. "Laverne Cox Opens Up About 'TIME' Cover and 'Orange Is the New Black'" Interview by Katie Couric. Youtube. 10 June 2014. < http://web.archive.org/web/20150318200216/https://www.youtube.com/watch?v=3mgBwCxTRDY>

Curtis, John W. (2010). "Faculty Salary Equity: Still a Gender Gap?" *On Campus with Women* 39(1).

Dasgupta, N. and Greenwald, A.S. (2001). "On the Malleability of Automatic Biases: Combating Automatic Prejudice with Images of Admired and Disliked Individuals." *Journal of Personality and Social Psychology* 81(5): 800-814.

DiAngelo, Robin. (2011). "White Fragility." *The International Journal of Critical Pedagogy* 3(3): 54-70.

Dobbin, F. & Kalev, A. (2007). "The Architecture of Inclusion: Evidence from Corporate Diversity Programs." *Harvard Journal of Law & Gender* 30:279-301.

Frankfurt, Harry. The Importance of What We Care about: Philosophical Essays. Cambridge, MA: Cambridge University Press, 1988.

Ferguson, Sian. "Calling In: A Quick Guide on When and How." Everyday Feminism. 17 January 2015. <http://web.archive.org/web/20150319211302/http://everydayfeminism.com/2015/01/guide-to-calling-in/>

Goodin, Robert. Utilitarianism as a Public Philosophy. Cambridge: Cambridge University Press, 1995.

Haslanger, Sally. (2015). "Social Structure, Narrative, and Explanation." *Canadian Journal of Philosophy*. DOI: 10.1080/00455091.2015.1019176.

Haslanger, Sally. (2004). "Oppression: Racial and Other." in M.P. Levine & T. Pataki (eds.), Racism in Mind (pp. 97-123), Ithaca and London: Cornell University Press.

Haslanger, Sally. (2002). "But Mom, Crop-tops *Are* Cute!" Social Knowledge, Social Structure, and Ideology Critique." *Philosophical Issues* 17: 70-91.

"How to Deal With Being Called Out." Too Young for the Living Dead. Tumblr. <http://web.archive.org/web/20150319211031/http://tooyoungforthelivingdead.tumblr.com/called-out>

I Am Trayvon Martin. http://iamtrayvonmartin.tumblr.com

If They Gunned Me Down. http://iftheygunnedmedown.tumblr.com

Kalev, A., Dobbin, F., & Kelly, E. (2006). "Best Practices or Best Guesses? Diversity Management and the Remediation of Inequality." *American Sociological Review* 71:589-617.

Kim, S., Kalev, A., & Dobbin, F.. (2012). "Progressive Corporations at Work: The Case of Diversity Programs". *Review of Law and Social Change* 36(2):171-213.

Korsgaard, Christine. Self-Constitution: Agency, Identity, and Integrity. New York: Oxford University Press, 2009.

Midgarden, Cory. 13 May 2014. "'Orange Is The New Black' Season 2." MTV. Retrieved from: <http://web.archive.org/web/20150318195651/http://www.mtv.com/news/1824055/orange-is-the-new-black-season-2-peoples-minds-are-going-to-be-blown/>.

NinjaCate. "On Tone Policing." Groupthink. Jezebel. 15 August 2013. <http://groupthink.jezebel.com/a-refresher-course-on-tone-policing-1562126390/all>.

NPR Staff. 16 August 2014. "Behind A Twitter Campaign, A Multitude of Stories." Code Switch. NPR.org. < http://web.archive.org/web/20150318202728/http://www.npr.org/blogs/codeswitch/2014/08/16/340669034/behind-a-twitter-campaign-a-multitude-of-stories>.

Office of Academic Affairs and Faculty Development and Advancement. (2015). "UC San Francisco 2014 Faculty Salary Equity Review." University of California San Francisco. Retrieved from: <http://web.archive.org/web/20150318192532/http://academicaffairs.ucsf.edu/academic-personnel/other/fser/UCSF_FSER_Report_January_2015.pdf>.

O'Neill, Onora. (1986). "The Moral Perplexities of Famine and World Hunger" in T. Regan (ed.), Matters of Life and Death: New Introductory Essays in Moral Philosophy, ed. New York: McGraw-Hill.

Penny, Laurie. "No, Boris, Being Called Racist is Nothing Like Dealing with Boko Haram." The New Statesman. 13 May 2014. <http://web.archive.org/web/20150319205448/http://www.newstatesman.com/laurie-penny/2014/05/laurie-penny-no-boris-being-called-racist-nothing-dealing-boko-haram>.

"Privilege-checking and Call-out Culture." 2 December 2012. Metafilter. <http://web.archive.org/web/20150318203929/http://www.metafilter.com/122432/privilegechecking-and-callout-culture>.

Richardson, L.S., & Goff, P.A. (2014). "Interrogating Racial Violence." *Ohio State Journal of Law* 12: 115-152.

Ripstein, Arthur. Equality, Responsibility, and the Law. Cambridge: Cambridge University Press, 1999.

Romano, Aja. "In Defense of Adria Richards and Call-Out Culture." The Daily Dot. 22 March 2013. <http://web.archive.org/web/20150319204532/http://www.dailydot.com/opinion/adria-richards-pycon-call-out-culture/>.

Scanlon, T.M. What We Owe Each Other. Cambridge, MA: Belknap Press, 1998.

Sewell, William H. (1992). "A Theory of Structure: Duality, Agency, and Transformation." *American Journal of Sociology* 98(1): 1-29.

Shear, Michael D. "Obama Speaks Out on Trayvon Martin Killing." *The New York Times*. 23 Mar. 2012. Web. Accessed 8 Jun 2015.

Smith, William A., Yosso, Tara J., & Daniel G. Solórzano. (2006). "Challenging Racial Battle Fatigue on Historically White Campuses: A Critical Race Examination of Race-related Stress." In C.A. Stanley (ed.), Faculty of Color Teaching in Predominantly White Colleges and Universities. Bolton, MA: Anker Publishing.

Stallings, Ariel Meadow. "Liberal Bullying: Privilege-Checking and Semantics-Scolding as Internet Sport." Offbeat Empire. 15 October 2012. <http://web.archive.org/web/20150318204307/http://offbeatempire.com/2012/10/liberal-bullying>.

Strawson, P.F. (1962). "Freedom and Resentment." *Proceedings of the British Academy* 48:1-25.

Stryker, Kitty, "Who's Afraid of Call-Out Culture? Jerks, Mostly." PurrVersatibility. 5 March 2015. <http://web.archive.org/web/20150319211922/http://kittystryker.com/2015/03/whos-afraid-of-call-out-culture-jerks-mostly/>.

Uprichard, Lucy. "In Defence of Call-Out Culture." The Blog. Huffington Post UK. 27 December 2012. <http://web.archive.org/web/20150318204734/http://www.huffingtonpost.co.uk/lucy-uprichard/call-out-culture_b_4507889.html>.

Watson, Gary. Agency and Answerability: Selected Essays. New York: Oxford University Press, 2004.

Wallace, R. Jay. Responsibility and the Moral Sentiments. Cambridge, MA: Harvard University Press, 2004.

Wolf, Susan. Freedom Within Reason. Oxford: Oxford University Press, 1990.

Young, Iris Marion. Responsibility for Justice. USA: Oxford University Press, 2011.

Young, Iris Marion. (2004). "Responsibility and Global Labor Justice." *Journal of Political Philosophy* 12(4): 365-388.

CHAPTER 5
Concluding Remarks

1. Introduction

In my introduction, I stated that the primary purpose of this dissertation was to re-examine our concepts of moral responsibility—and to develop new ones—in response to the challenges raised by the phenomenon of implicit bias. I interpreted that challenge as the conjunction of three different facets of our current moment: 1) the triumph of social scientific research establishing the existence of implicit social cognition, 2) the shift from first- to second-generation barriers to social equality, both explicit and implicit, and 3) the highly polarized climate of intergroup disagreement and mistrust. In response, I have distinguished between two different concepts of moral responsibility and identified one of them—responsibility as accountability—as a more theoretically fruitful source of solutions to these problems. I have shown my concept of accountability gives rise to its own set of critical moral responses that are distinct from those usually discussed in the moral responsibility literature, and how it can undergird collective practices aimed at structural change.

The purpose of these concluding remarks is "retell the whole story," as it were, in one place, and in two ways. First, I describe as a whole the double-sided framework for thinking about moral responsibility that I have developed here. Second, I summarize the recommendations about implicit bias that this theoretical apparatus yields. Finally, I indicate new directions for further research.

2. Moral Responsibility, Revisited

I used the phenomenon of implicit bias—and the conflicting intuitions about moral responsibility that it generates—as an occasion to peel apart two different concepts of moral

responsibility: attributability and accountability. I showed that these two concepts arise from two distinct philosophical concerns answering to distinct purposes. *Attributing* actions to agents for the purposes of moral appraisal requires answering questions in metaphysics and philosophy of action: "What makes it possible for me to *act*, and what makes my actions my own?" The primary purpose of a concept of responsibility as attributability is to understand and assess moral agents as good or bad exemplars of their kind (and perhaps to delineate what entities do not count as moral agents). By contrast, the primary purpose of a concept of responsibility as accountability is to activate people's agency—in Sewell's (1992) sense—in their capacity as enactors and reformers of social structure. Of course, the former is a prerequisite for the latter—only moral agents can enact and reform social structure—but the sorts of questions that are relevant for determining attributability and accountability are different. Holding agents *accountable* for moral consequences of their actions requires answering questions of moral and political philosophy: "How should we assign duties and distribute burdens across the moral community in line with principles of justice and fairness?"

To further highlight the difference between these two different kinds of concerns, I distinguished between two different kinds of critical moral response: appraisal-based and non-appraising. Appraisal-based responses to some action are only warranted when that action is properly attributable to a person as a reflection of her moral agency, for otherwise that action could not provide grounds for moral appraisal of the person. They include the traditional responses associated with moral responsibility: praise and blame, resentment and the reactive attitudes, and punishment. While the first two kinds of appraisal-based response are typically carried out at the level of interpersonal interaction, punishment is carried out institutionally.[98] Because their reach runs so deep, appraisal-based critical responses require clear standards of evidence and application; they are warranted when people violate their perfect and imperfect moral duties. This is because people care about whether they are good or bad moral agents, making negative appraisal highly threatening.

[98] Here I am endorsing Feinberg's (1970) claim that punishment is a "conventional device for the expression of attitudes of resentment and indignation, and of judgments of disapproval and reprobation," and that it is this condemnatory character that distinguishes punishment from mere penalty and other forms of sanction (97).

	Appraisal-based Responses	Non-appraising Responses
Interpersonal	Blame Resentment	Moral admonition Demand for apology
Institutional	Punishment	Mandate to change policy Compensation/recognition of victims

Table 2: Critical Moral Responses.

Non-appraising responses, by contrast, are not forms of "deep" appraisal and they are not only applicable to things other than violations of duty. Rather, they are critical responses warranted by all actions (and omissions) that pertain to the roles people occupy in carrying out the moral division of labor within a moral community. While institutional forms of non-appraising responses—mandated changes in policy, penalties (but not punishment), and compensation to victims of harm—are familiar, I sought to develop additional *interpersonal* forms of non-appraising response to wrongdoing. These included demands for apology as well as moral admonition, which I characterized as 1) targeting the action and another actor, 2) evoking role-based ideals, and 3) communicating guidelines for improvement. Their purpose is not to appraise moral agents, but to reaffirm moral ideals, educate, and motivate agents to strive toward those ideals in ways that it would be reasonable to expect of them. Mandated change in policy is the institutional equivalent of moral admonition, while compensation and recognition of victims serve the reparatory functions of apology. Moreover, the flexibility and multiplicity of agents' role-based ideals allow them to serve as the basis for collective structural change. ("Think globally, act locally!" as activists are wont to say.) While particular failures to perform an act that would contribute to discharging an imperfect duty do not warrant appraisal-based responses, they may warrant non-appraising responses in virtue of the fact that there is always more work to be done (and because we should be on guard against bad faith and self-serving interpretations of what is "reasonable").

Finally, I also provided an account of how these two concepts of responsibility are related, and hence a diagnosis of why they have so often been conflated. While people deserve appraisal-based responses like blame, resentment, and punishment *directly* on the basis of attributable actions that manifest their moral agency, an agent *may* be subject to appraisal-based responses for accountability reasons, but only after the social burdens of her failure to carry out her assigned responsibility have been appropriately distributed in accordance with fair and just principles.

Because my own view is non-consequentialist but sensitive to pragmatic considerations, this further question of appraisal on the basis of having violated a duty is only warranted when the violation is genuinely attributable to the agent and when it is not counterproductive.

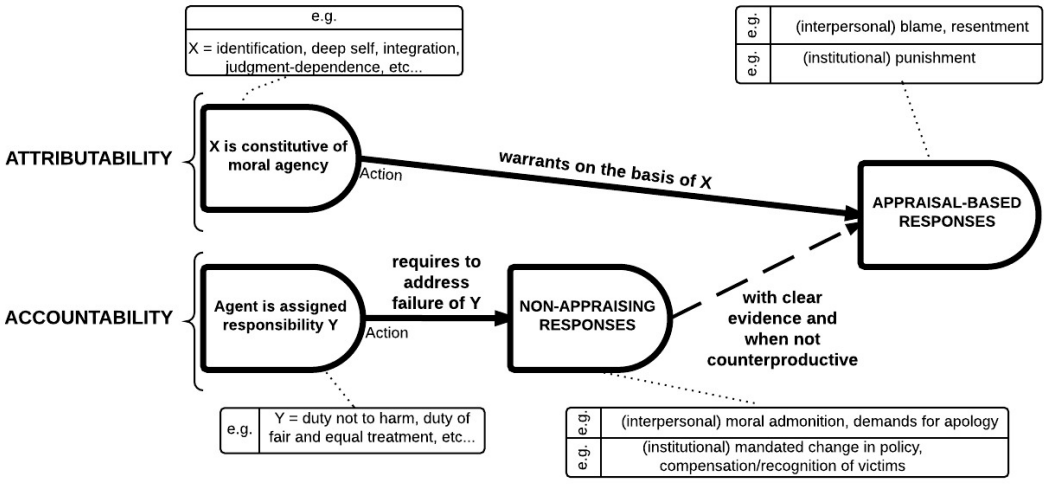

Figure 6: Attributability and Accountability.

A second way in which attributability and accountability are related is that questions of accountability may figure into ascriptions of attributability, i.e. whether some action is genuinely attributable to a person may depend on whether she has done what she can reasonably be expected to do in carrying out her imperfect duties (e.g. of mitigating bias, of working toward structural transformation). In advocating that we shift our focus from attributability to accountability, from metaphysical to social agency, I am in effect advocating for a shift from *ethics* to *politics*, away from questions of demarcating exactly what we are required to do (Oughts) and toward questions of what more we can strive to do (Ideals). At the same time, I have advanced an account of rectifying structural injustice on which the *political* is *ethical*—dependent on everyone's efforts to achieve their own personal role-based ideals—but where the ethical is now infused with the social. This relationship between the ethical and the politics thus mirrors the relationship between agency and structure. Moral responsibility, in its double-sidedness, is the glue that binds the two together.

3. Implicit Bias, Revisited

Another major aim of my dissertation has been addressing the question of what sorts of critical moral response are appropriate to implicit bias. In particular, I have considered critical moral responses to [1] actions caused by implicit bias, [2] actions aimed at mitigating implicit

bias, and [3] actions aimed at transforming the social structures that produce implicit bias. In each case, I have argued that practices of accountability are more appropriate—normatively and practically—than practices of attributability.

I argued that there exist cases in which [1] actions caused by implicit bias are not attributable to us, but that even in these cases we are still accountable for them. Actions caused by implicit bias are not attributable to us if we would not endorse their influence *and* we have done what we can reasonably be expected to do to avoid and respond to bias; when both of these conditions hold, we have an *excuse* for our wrongdoing and are not blameworthy. However, since others were wrongfully treated, and since the burdens must be picked up somehow, it will often still be appropriate for us to be the ones to bear those burdens. (Some burdens in particular, e.g. the burden of apologizing and reaffirming egalitarian norms, fall to us because they cannot be carried out by third-party bystanders.) But even though there are such cases in which we lack attributability, I argued, we should in general refrain from appraisal-based responses, for two reasons. First, it is usually very epistemically difficult to obtain the evidence we need to know in a particular case that some action was caused by an implicit bias in a blameworthy way, and we should err on the side of caution before engaging in morally serious appraisal-based responses. Second, even when we do have grounds for an appraisal-based response, blame and resentment are likely to provoke defensiveness and resistance, and hence to be counterproductive.

I have argued that we should think of ourselves as having an imperfect duty to [2] mitigate bias: that is, to take steps to unlearn and intervene on our implicit biases. These include increasing one's exposure to counterstereotypical exemplars (e.g. by seeking out alternative media and stories), increasing personal contact with outgroup members, and keeping egalitarian goals chronically activated (e.g. by staying involved, aware, and educated about these issues in ways that facilitate the detection of self-discrepancies). I have also argued that part of being responsible (accountable) for implicit bias involves an imperfect duty to [3] work toward structural transformation. What this entails will depend on each person's particular role-based ideals, but it will likely involve such steps as becoming educated and aware of unjust structural processes, further educating and collaborating with others in one's role-based capacities, and collectively organizing to exercise those capacities in order to change existing structures. As with all imperfect duties, because we are creatures of limited time and resources, we have considerable discretion in deciding when and how to discharge this duty. If an agent *clearly* fails to perform a

sufficient number of actions to mitigate bias, then she deserves negative appraisal-based responses like blame and resentment. However, in many cases concerning a particular action or omission that represents a missed opportunity to work toward mitigating bias, it will not be readily apparent whether an agent has otherwise reached this threshold—not only because it is epistemically opaque but also morally indeterminate. This means that we should again, except in clear cases, stick to non-appraising responses such as moral admonition. Moral admonition is generally warranted because unlearning implicit bias is a long and ongoing project that may never be fully achieved, but serves as a regulative ideal. Pragmatically speaking, moral admonition is also preferable to blame and resentment because it engages people's internal, prescriptive moral motivations—encouraging them to approach their own (role-based) Ideal selves—rather than imposing the external threat of sanctions to proscriptively motivate hewing to their Ought selves.

4. Directions for Future Research

In this section I will raise a number of questions and issues that spring from the project I have developed here, which I divided into three kinds: further problems arising from empirical findings, further moral analyses and concepts to pursue, and meta-issues concerning the relationship between empirical and normative inquiry.

4.1 Agency in the Age of Dual-Processing

In this dissertation I proposed a novel approach to understanding moral responsibility for implicit bias. However, implicit bias is but one process countenanced by so-called "dual process" theories of reasoning, which posit the existence of two different types of reasoning: one fast, automatic, and inaccessible to introspection; the other slow, effortful, controlled, and self-reflective. As described in my introduction, implicit bias is a species of *implicit social cognition*, which I defined as information processing about social objects—oneself, others, and social groups—that can occur in some sense without awareness or control, in ways that one might not reflectively endorse. In the introduction[99] I mentioned another example of such a process: situational priming, as in the Kay et al. (2004) study in which participants exposed to a briefcase subsequently judged an ambiguous interaction as being more competitive than participants

[99] See Fn. 2.

exposed to a backpack. But there are other processes of implicit social cognition that are likely to be involved in the persistence of unjust social structures, such as "embodied cognition" and the "fundamental attribution error". For example, Carney, Cuddy, & Yap's (2010) study on embodied cognition—the idea that that cognitive reasoning is influenced by (often metaphorical understandings of) bodily sensations and experiences, showed that taking up wider and more expansive body positions—those traditionally gendered as male—lead to increased confidence, risk taking, and higher levels of testosterone. The fundamental attribution error, moreover, is the psychological tendency to overestimate how much of a person's behavior is caused by stable dispositional traits of the person herself, rather than by contingent features of the situation in which that person is acting. Because the fundamental attribution error directs our attention away from background social institutions and toward the purported internal dispositions of people located in those institutions, it very likely plays an important role in maintaining unjust social inequalities and ideologies that support such hierarchies.[100] Each of these psychological processes, insofar as they contribute to morally significant outcomes, deserves some attention and may call out for different moral prescriptions to act.

In addition to such cognitive processes, however, which might be assimilated to the "cold" heuristics and biases identified by behavioral economists like Tversky and Kahneman (1974), the social domain also introduces a range of human motivations that can similarly distort reasoning and decision-making. Kahan (2011) defines *motivated reasoning* as the "unconscious tendency of individuals to fit their processing of information to conclusions that suit some end or goal". Motivated reasoning was illustrated in Chapter 3 during my discussion of Nyhan and Reifler's (2010) study on the backfire effect, in which people believed false claims even more strongly after they had been given empirical evidence to the contrary—an effect that disappeared when they first performed self-affirmation exercises that reduced their feelings of threat. Another example is Kahan's work on "cultural cognition," whereby people's social group identities shape their judgments of reliable testimony and epistemic authority, and hence their views about the empirical bases of politically controversial debates, e.g. whether climate change is manmade,

[100] Ross (1977), who coined the term, warns that this error represents "a particularly insidious brake upon social mobility whereby the disadvantaged and powerless overestimate the capabilities of the powerful, who in turn inappropriately deem their own caste well suited to the task of leadership." See also Benforado and Hanson (2011), who argue that the ideological chasm between liberals and conservatives in the U.S. is primarily a difference in susceptibility to the fundamental attribution error.

whether the death penalty deters crime, and so on.[101] Another potential form of motivated reasoning relevant for issues of social equality is proposed by "system justification theory". According to Jost, Banaji, & Nosek's (2004) system justification theory, there are three basic motives: to protect the self, the group, and the status quo; while the first two have long been accepted, the existence of this third motive to protect the status quo can explains why disadvantaged groups sometimes support policies that are actually harmful to themselves and their own groups. From a moral perspective, unconscious *motivation* is likely to be relevantly different from unconscious cognitive processing. An empirically-sensitive account of moral responsibility, then, should be able to accommodate not only various forms of implicit social cognition but also motivated reasoning.

More generally, dual process models call for a replacement of the picture of human psychology on which agents reflectively weigh reasons for and against believing and acting in certain ways with one on which most decision-making is performed (or influenced) by automatic processes of which we remain largely unaware. But this old picture—a mainstay of the Western philosophical tradition—formed the basis for our understanding of ourselves as *rational* and *autonomous* agents. If our beliefs and actions are not caused by the reasons we took there to be, then we are not believing and acting upon reasons that we have ourselves endorsed, then we may not be fully autonomous. Moreover, if what we are actually responding to—the real causes of our beliefs and actions—does not correspond to genuine reasons, then we may fail to be rational.[102] I suggest, then, that in addition to reconceptualizing the way we morally respond to actions and judgments, we must also reconceptualize what it means to deem them—and agents—autonomous or rational. A much bigger project, then, would be to develop new conceptions of autonomy and rationality[103] that are sensitive to contemporary empirical findings that cognition is *situated* (dependent on cultural and environmental context), *embodied* (dependent on sensory and physical states of the body), and *social* (dependent on identities and relationships with others).

---

[101] Kahan (2013) defends a "a theory that sees ideologically motivated cognition not as a reasoning deficiency but as a reasoning adaptation suited to promoting the interest that individuals have in conveying their membership in and loyalty to affinity groups central to their personal wellbeing". See also Kahan and Braman (2006) for a representative piece on cultural cognition.

[102] To put this in the helpful terminology of Scanlon (1998) and others: We may be mistaken in identifying the explanatory reasons for our actions: while we may think that the explanatory reasons are the same as our justificatory reasons, it may be that the real explanatory "reasons" are mere causes that are not genuine reasons at all.

[103] For recent important work in this area with regard to automaticity, see for instance Railton (2009), Brownstein and Madva (2012), and Brownstein (2013).

4.2 Morality in an Age of Injustice

There are further complications on the moral-theoretical end of things. Phenomena like implicit bias force us to adopt a more complicated picture of the objects of moral assessment: beyond just particular actions at a particular times, there are also the various kinds of conscious and non-conscious influences that go into any particular action, and there is the particular significance or meaning of that action in the context of larger social structures. Furthermore, I claim, the simple dichotomy between "morally right" and "morally wrong" actions is not adequate for the full range of our moral experience. This is particularly so if "morally wrong" is understood (as is common among philosophers) to entail or to be defined in terms of moral blameworthiness, since—as I argued in this dissertation—some actions, like those caused by implicit bias, may not be morally blameworthy but are still morally *problematic*. Relatedly, feminist and antiracist philosophers have frequently stumbled upon the question of whether various sorts of actions—pornography and prostitution, adherence to oppressive beauty norms, racial endogamy, and so on—are properly condemned as morally wrong or blameworthy. And similar issues crop up in the evaluation of historically and culturally distant others, when we hesitate to condemn or blame them for actions that we would normally deem to be morally wrong. Thus, I believe, it is incumbent on moral philosophers to continue developing more textured understandings of what it means for an action, attitude, preference, practice, institution, or product to be "morally problematic," where this is not synonymous with being morally wrong or morally blameworthy, but instead grows out of real-time, on-the-ground critical practices of identifying and problematizing actions that contribute to injustice and oppression.[104]

As I have already demonstrated in this dissertation, part of this project involves analyzing and conceptualizing a larger repertoire of critical moral responses, especially *non-appraising* responses that are appropriate for actions that are morally problematic but not blameworthy. I have described one example in detail, moral admonition as the non-appraising counterpart to blame, but there are no doubt others. In particular, are there non-appraising counterparts to resentment and the other reactive attitudes, and for first-person analogues[105] such as guilt and

[104] To take just one example already mentioned, Young's theory of liability versus social connection models of responsibility grew out of her (2003) and (2005) work on students' anti-sweatshop campaigns. This bottom-up methodology is also common throughout much feminist and other critical approaches.

[105] I am grateful to Christopher Bennett for suggesting that I consider first-person analogues of blame and resentment. However, since blame is a paradigm example of an appraisal-based response, and since my social

shame? I believe that there are. In particular, as I mentioned in Chapter 3, contemporary social psychological work demonstrates that guilt is a non-appraising moral emotion, while shame is appraisal-based. Tangney, Stuewig, & Mashek's (2007) account, derived from a range of experimental and qualitative methods including case studies, self-reported narratives, tests of attribution and counterfactual thinking, show that guilt invites thoughts of the form "If Only I Hadn't" while shame, "If Only I Weren't". Moreover, Bernard Williams (1982) has also developed the first-personal concept of "agent-regret," a moral emotion that attends the bad consequences of a person's action even when she was in no way at fault: in other words, when she was causally but not *morally* responsible. Agent-regret is thus a non-appraising emotion, and may be warranted in cases when a person learns that an implicitly biased action of hers has caused harm to others; it is simply entailed by being committed to egalitarian norms that one should feel pain at having treated someone in an unjust way even when it does not reflect on one's agency. I suggest that the non-appraising counterpart of resentment that is warranted by, e.g. actions caused by implicit bias, may be something like the third-party analogue of agent-regret, which may be some species of disappointment or dissatisfaction. My hunch is that we will be able to gain insight into these subtle varieties of critical moral response by investigating and theorizing our ordinary lived experiences of noticing, struggling with, and addressing oppression.

4.3 Meta-Issues

In the introduction, I described a discussion by Gawronski, Peters & Lebel (2008) and Gawrsonki (2009) in the social psychological literature on the question of whether implicit attitudes are personal or extra-personal, in which the authors conclude that this is ultimately a non-empirical and *normative* question. Indeed, this is precisely the question addressed by some of my arguments in Chapter 2, along with much current literature of moral responsibility for

---

psychological use of "guilt" is *not*, guilt in my sense is not equivalent to "self-blame," which is a separate phenomenon.

On my view, self-blame may still be an appropriate response to one's own implicit bias even if blame from others is not. An important disanalogy is the shift in epistemic perspective. While it will often be difficult for a third-party to know the patterns of action (and omission) which a given individual may have performed (or failed to) with respect to mitigating bias, the individual *herself* may know full well whether she has sufficiently strived toward the fulfillment of these imperfect duties. Even if she does not know about implicit biases per se, she may know that she has not done what could reasonably be expected of her toward efforts at collective reform of our bias-producing social structures. This much moral knowledge is by no means too advanced for the ordinary person, hence her knowledge that she does not satisfy this second condition for excuse be enough to warrant self-blame.

implicit bias[106]. As should be unsurprising in the case of research with such obvious normative implications, this is not the only such debate. A related debate concerns the question of whether implicit bias is a form of prejudice. While implicit bias research is certainly located within a tradition of research on stereotyping and prejudice, Tetlock and Arkes (2004) argue that the question of what constitutes prejudice is "at root, not just a psychological one" but rather "a deeply political one that requires us to make moral judgments of which aspects of public opinion (implicit or explicit) deserve our censure—and make no mistake here, when you call people prejudiced, they feel censured" (316). Arkes and Tetlock (2004) make this point in the context of criticizing implicit bias research for (among other things[107]) being ideologically biased by politically liberal researchers who assume without justification—or are illegitimately motivated to draw conclusions in favor of—a particular causal explanation of persistent racial inequalities. Their critique is a telling illustration of how, as they put it, "observers with different causal assumptions and value priorities—unless anchored down by ex ante reputational bets—will always be able to put their preferred theoretical spins on the same facts" (320). They thus recommend "an old-fashioned philosophy of science that distinguishes between factual and value judgments, between *is* statements, such as 'reaction time fell from 3,290 to 3,260 millisec,' which can be true or false 'nonperspectivally,' and *ought* statements, such as 'someone with an associative network with these properties should be censured for racism,'[108] which can only be true or false 'perspectivally'" (313). I am skeptical, however, that it is possible to maintain the strict fact-value dichotomy that Arkes and Tetlock prescribe, in the face of decades of work by feminist, social, and naturalized epistemologists and philosophers of science who have argued precisely the opposite. Causal explanations in particular have been shown to be tightly bound up

[106] Ch. 1, Section 3.3.

[107] Here I bracket their discussion of methodological flaws and alternative theoretical interpretations of particular measures and results.

[108] It is worth nothing, however, that leading implicit bias researchers have not made such claims. Banaji and Greenwald (2013) write: "[W]e answer no to the question 'Does automatic White preference mean 'prejudice'?' The Race IAT has little in common with measures of race prejudice that involve open expressions of hostility, dislike and disrespect" (52). Similarly, a Frequently Asked Question on the Project Implicit website reads: "Social psychologists use the word prejudice to describe people who report and approve negative attitudes toward outgroups. Most people who show an implicit preference for one group (e.g., White people) over another (e.g., Black people) are not prejudiced by this definition. The IAT shows biases that are not endorsed and that may even be contradictory to what one consciously believes. So, no, we would not say that such people are prejudiced. It is important to know, however, that implicit biases can predict behavior. When we relax our active efforts to be egalitarian, our implicit biases can lead to discriminatory behavior, so it is critical to be mindful of this possibility if we want to avoid prejudice and discrimination."

with normative judgments of moral responsibility.[109] It would be fruitful, then, to apply these approaches to diagnosing the self-described "part psychological, part philosophical, and certainly part political" disagreement described by Arkes and Tetlock, as well as to larger questions of the role of empirical and philosophical inquiry in determining causal explanations of morally relevant outcomes like large-scale social inequality.

There are other questions in this vicinity, going the other direction. I have explicitly adopted what I called a "pragmatic constraint" on any acceptable theory of moral responsibility: that it must prescribe recommendations that will be practically effective in bringing about desirable social change. This constraint is justified by another assumption I have made only implicitly: that problems of psychology can illuminate issues of morality. For example, I have interpreted psychological recalcitrance—as in studies on the backfire effect and the ineffectiveness of external sanctions—as flagging cases in which something is *morally* amiss: the lack of underlying relationships of intergroup trust, for example, or failure to respect people's experiences of themselves as autonomous and efficacious agents. Such claims are unsurprising given a naturalistic picture on which moral norms and complex human psychologies evolved in tandem in order to overcome problems of social coordination, or given a view of moral theory as existing primarily in order to guide action. But these are only working assumptions which have guided me throughout this project, and require further defense.

5. Conclusion

As I hope to have demonstrated throughout this dissertation, moral philosophers with commitments to social justice have much to be gained by being empirically-informed. Indeed, working out of a genuine concern with real-world injustice—in other words, with my pragmatic constraint—requires a sensitivity to empirical evidence about how different actions are likely to fare in the real world. This brings me to a second point, which is that philosophers would do well (as feminist philosophers have long argued) to pay much more attention to the important of actual relationships, which require time, physical spaces, and embodied, face-to-face contact. Social psychological study has shown that pure argumentation and facts do not change minds.

[109] See for instance Hitchcock and Knobe (2009), Feinberg (1970), and Smiley (1992). These authors provide both philosophical and empirical evidence for the view that people's causal attributions depend on normative judgments of who or what should have—and hence who should have had the *power* to—prevent some outcome. These are irreducibly moral and political questions. See also Björnsson and Persson (2012) for the view that judgments of moral responsibility are actually a type of (causal) explanatory judgment.

This is not to say that we should abandon reasoned arguments, but that we need to consider the background social and psychological conditions that act as barriers or facilitators to the uptake of arguments; and here, as any organizer of a political, community, or social movement would attest, relationships are the beginning and end of all conversations. This means that philosophers should not think their contribution lies only in producing a body of texts, but also in the ways they participate in a living community of scholars in dialogue with others. While there is certainly work to be done from the armchair, this should not preclude us from conversing through the window with those who work on the ground, or from visiting with the neighbors. To be sure, philosophical work that relies on current empirical work is more vulnerable to refutation as new evidence accumulates and shifts. But this, I believe, simply means that we should view ourselves in a new light, more akin to those in the empirical sciences: as offering up tentative and timely investigations and hypotheses, open to further testing, not necessarily permanent moral truths. This way of conceiving of ourselves and our role is not intended to supplant more traditional forms of philosophical work, but to enhance and empower it. In doing so—by examining the normative concepts and assumptions that structure our social world while also paying attention to our best empirical understandings of that world—we can take up a unique and enduring role in changing it. Philosophy can be part and parcel of the struggle for social justice.

References

Arkes, H.R. & Tetlock, P.E. (2004). "Attributions of Implicit Prejudice, or 'Would Jesse Jackson 'Fail' the Implicit Association Test?'" *Psychological Inquiry* 15(4): 257-278.

Banaji, Mazahrin R., and Anthony Greenwald. Blindspot: Hidden Biases of Good People. New York: Delacorte Press, 2013.

Benforado, A. & Hanson, J.D. (2012). "Attributions and Ideologies: Two Divergent Visions of Human Behavior Behind Our Laws, Policies, and Theories" in J.D. Hanson (ed.) *Ideology, Psychology, and Law* (pp. 298-338), New York: Oxford University Press.

Björnsson, G. & Persson, K. (2012). "A Unified Empirical Account of Responsibility Judgments." *Philosophy and Phenomenological Research* 87(3): 611-639.

Brownstein, Michael. (2014), "Rationalizing Flow: Agency in Skilled Unreflective Action." *Philosophical Studies* 168(2): 545-568.

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

Carney, D.R., Cuddy, A.J.C., & Yap, A.J. (2010). "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance." *Psychological Science* 21(10):1363-1368.

Feinberg, Joel. Doing and Deserving: Essays in the Theory of Responsibility. Princeton, NJ: Princeton University Press, 1974.

Gawronski, Bertram. (2009). "Ten Frequently Asked Questions About Implicit Measures and Their Frequently Supposed, But Not Entirely Correct Answers." *Canadian Psychology* 50(3):141-150.

Gawronski, B., Peters, K. R., & LeBel, E. P. (2008). What Makes Mental Associations Personal or Extra-personal? Conceptual Issues in the Methodological Debate About Implicit Attitude Measures. *Social and Personality Psychology Compass* 2: 1002-1023.

Hitchcock, C. & Knobe, J. (2009). "Cause and Norm." *The Journal of Philosophy* 106(11): 587-612.

Jost, J.T., Banaji, M.R., & Nosek, B.A. (2004). "A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo." *Political Psychology* 25(6): 881-919.

Kahan, Dan. "Ideology, Motivated Reasoning, and Cognitive Reflection." *Judgment and Decision Making* 8(4): 407–424.

Kahan, Dan. "What Is Motivated Reasoning and How Does It Work?" *Science and Religion Today*. 4 May 2011. Web. <http://www.scienceandreligiontoday.com/2011/05/04/what-is-motivated-reasoning-and-how-does-it-work/>.

Kahan, Dan M. and Braman, Donald, Cultural Cognition and Public Policy. Yale Law & Policy Review, Vol. 24, p. 147, 2006; Yale Law School, Public Law Working Paper No. 87. Available at SSRN: http://ssrn.com/abstract=746508

Kay, A.C., Wheeler, C., Bargh, J.A., & Ross, L. (2004). "Material Priming: The Influence of Mundane Physical Objects on Situational Construal and Competitive Behavioral Choice." *Organizational Behavior and Human Decision Processes* 95: 83–96.

Nyhan, Brendan, and Jason Reifler. (2010). "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32(2): 303-330.

"Project Implicit." (2011). Retrieved from: <http://www.projectimplicit.net/>.
Radzik, Linda. Making Amends: Atonement in Morality, Law, and Politics. USA: Oxford University Press, 2011.

Railton, Peter. (2009). "Practical Competence and Fluent Agency." In D. SObel & S. Wall (eds.), Reasons for Action (pp. 81-115). Cambridge: Cambridge University Press.

Ross, L. (1977). "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process." In L. Berkowitz (ed.), Advances in Experimental Social Psychology (pp. 174-214). New York: Academic Press.

Scanlon, T.M. What We Owe Each Other. Cambridge, MA: Belknap Press, 1998.

Tangney, J.P., Stuewig, J., & Mashek, D.J. (2007). "Moral Emotions and Moral Behavior." *Annual Review of Psychology* 58: 345-372.

Tetlock, P.E., & Arkes, H.R. (2004). "The Implicit Prejudice: Islands of Consensus in a Sea of Controversy." *Psychological Inquiry* 15(4): 311-321.

Tversky, A. & Kahneman, D. (1974). "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185(4157): 1124-1131.

Williams, Bernard. Moral Luck: Philosophical Papers 1973-1980. Cambridge: Cambridge University Press, 1982.

Young, Iris Marion. (2004). "Responsibility and Global Labor Justice." *Journal of Political Philosophy* 12(4): 365-388.

Young, Iris. (2003). "From Guilt to Solidarity: Sweatshops and Political Responsibility." *Dissent* 50(2): 39-45.