8-2016

The dark side of reactive attitudes: From persons to compatibilism

Mallory A. Parker
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Mallory Parker

Entitled
THE DARK SIDE OF REACTIVE ATTITUDES: FROM PERSONS TO COMPATIBILISM

For the degree of Master of Arts

Is approved by the final examining committee:

Daniel Kelly
Chair

Daniel Smith

Taylor Davis

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Daniel Kelly

Approved by: Rod Bertolet 6/29/2016
Head of the Departmental Graduate Program Date

THE DARK SIDE OF REACTIVE ATTITUDES: FROM PERSONS TO COMPATIBILISM

A Thesis

Submitted to the Faculty

of

Purdue University

by

Mallory A. Parker

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Arts

August 2016

Purdue University

West Lafayette, Indiana

TABLE OF CONTENTS

ABSTRACT

Parker, Mallory A. M.A., Purdue University, August 2016. The Dark Side of Reactive Attitudes: From Persons to Compatibilism. Major Professor: Daniel Kelly.

This thesis contains two independent papers that both address the problems associated with the reactive attitudes. The first paper, presented in Chapter 2, discusses the negativities of the reactive attitudes in debates regarding skepticism about the moral notion of persons. The second paper, presented in Chapter 3, presents the negativities associated with the reactive attitudes in debates concerning compatibilism about moral responsibility. Neither chapter deals solely with presenting the negativities associated with the reactive attitudes. More than present these, both chapters undermine the non-realist or compatibilist philosophical arguments that attempt to save either the moral notion of persons (Chapter 2) or moral responsibility (Chapter 3) from skeptical and incompatibilist arguments, respectively, by appealing to the benefits of the reactive attitudes. Each chapter undermines these arguments by reminding readers that we are using the benefits of the reactive attitudes and ignoring their detriments in order to cling to tightly held philosophical notions.

CHAPTER 1. INTRODUCTION

The unifying theme of chapters 2 and 3 of this thesis is that they both argue against the practices associated with moral responsibility by demonstrating the detrimental effects of the reactive attitudes. The first chapter argues in favor of skepticism about the abstract notion of persons employed in attributions of moral responsibility. The second chapter argues against compatibilist arguments that appeal to the reactive attitudes to justify the practices associated with moral responsibility. In both chapters, the detrimental effects of the reactive attitudes are associated with biases and stereotypes.

The first chapter involves a stronger focus on agency as well as realism and non-realism about moral responsibility. The authors with which the chapter engages are dealing more directly with philosophical concepts. In this chapter, there is appeal to empirical literature, but with an eye toward informing our understanding of philosophical concepts employed in debates concerning moral responsibility skepticism.

The second chapter takes a more pragmatic approach, focusing less than the first chapter on determining the truth of extant philosophical concepts. I abandon discussion of agency and realism or non-realism about moral responsibility, and instead focus on debunking compatibilist arguments that justify our practices associated with moral responsibility by appeal to the reactive attitudes. While the first chapter distances itself from the philosophical concepts to a small degree, the second chapter distances itself considerably.

However, both chapters have much to bring to bear on the philosophical discussions with which they engage. By moving away from the philosophical concepts, the chapters do not abandon the philosophical topics at hand; rather, they avoid being too restricted by the manner in which the topics have historically been construed.

All in all, the two chapters present the development of my views concerning skepticism about moral responsibility in increasing sophistication. While the first chapter remains tied to defending philosophical extremes, the second chapter, in virtue of moving further from the philosophical concepts that have historically been involved in debates about moral responsibility, achieves greater flexibility in parsing out the debate. Each chapter is an instantiation of a skeptical intuition concerning moral responsibility, an attempt to perform in words an argument that is, first and foremost, a disposition. However, they are not mere sophistry. I believe that each chapter comes closer to revealing the causes of this disposition. At the very least, they reveal what this disposition draws attention to.

CHAPTER 2. A REPLY TO THE NON-REALIST DEFENSE OF PERSONS

> "The surest way of ruining a youth is to teach him to respect those who think as he does more highly than those who think differently from him" (Daybreak, Sec. 297).

2.1 Introduction

Mounting evidence in the literature on automaticity suggests that we exhibit much less control over our thoughts and behavior than we attribute to ourselves. In Skepticism About Persons, John Doris argues that this evidence provides a robust challenge to the moral notion of persons. In response, C.D. Meyers has argued that if we accept a non-realist conception of personhood, we need not abandon the moral concept of persons, and, moreover, "there are good reasons why we should retain the concept of personhood in our moralizing, independently of whether agency really exists" (Meyers 194). In what follows, I argue in response to Meyers that if we accept a non-realist conception of personhood, there are good reasons to significantly reduce the use of the concept of persons in our moralizing, my main concern being its use in attributions of moral responsibility. I then conclude by suggesting that a weaker form of skepticism may prove to have an important role to play in our moral psychology.

I first summarize Doris' skeptical challenge to the traditional moral notion of persons. I then explain C. D. Meyers' argument that Doris assumes a kind of realism about persons and moral responsibility and his proposal that adopting a non-realist metaethics allows us to continue using a moral notion of persons. Following this, I explain Kyle Stanford's evolutionary account of moral externalization. I then draw a connection between the function of reactive attitudes on his account and the negative effects of this function, using these as evidence that there are good reasons to reduce our use of the concept of persons in our moralizing, insofar as it grounds attributions of moral responsibility. I conclude by noting that skepticism puts the focus on environmental contributions to misbehavior, forcing us to focus on how to manipulate environments so they give rise to morally permissible behavior rather than on the punishment of individuals. With this in mind, I suggest that a weaker form of skepticism about persons may serve as a means of combating the negative effects that arise from reactive attitudes and the externalization of moral responsibility.

2.2 Summary of Doris

In Skepticism About Persons, John Doris argues that the skeptical challenge raised by the empirical literature on automaticity presents a substantial threat to traditional notions of persons and agency. While he does not think the challenge of skepticism is insurmountable, he aims to demonstrate that it should be taken seriously. Skepticism about persons cannot be easily dismissed, and, if it can be razed, it will be dragging some of our previous philosophical commitments with it.

It is important to note the notion of persons being discussed. Distinguishing the skepticism he is elucidating from skepticism about personal identity, Doris explains that the skepticism he is concerned with suggests "there's reason to doubt… that human beings function as agents in the sense supposed to ground attributions of moral responsibility" (Doris 58). The notion of person is what he calls a "moral honorific;" persons are entitled to a certain manner of treatment and they are obligated to comport themselves in a particular manner. Persons are imputed with these entitlements and expectations at least in part because they are agents. Functioning as a person entails acting as an agent. When someone acts as an agent, it is appropriate to ascribe moral responsibility to at least some of their behavior, and, as a result, it is also appropriate to subject them to reactive attitudes.

Another important component of Doris's argument is his characterization of reflectivism, which he takes to capture the traditional philosophical view of persons threatened by the empirical data. On the reflectivist account, personhood requires accurate self-reflection. Doris explains that "characteristically personal functioning requires that one correctly detect salient facts about oneself, such as what mental states are implicated in one's behavior" (Doris 60-1). To illustrate the point that self-reflection is one necessary requirement for agency under reflectivism, he provides Wanton, an example of an individual who everywhere fails to reflect. This thought experiment is designed to elicit intuitions that Wanton is more akin to an animal than a person or rational agent. To illustrate that not only is self-reflection necessary, but that it must be accurate, Doris provides Clueless, an example of an individual who reflects incessantly, but always does so inaccurately. Clueless is designed to elicit the intuition that, even if we might still count him as a person, we would be in good company if we had doubts about his ability to exercise agency. Doris summarizes reflectivism by explaining that it "maintains that the paradigmatic form of personal functioning is reflective self-direction, which requires that the actor accurately attend to her psychology (introspection) and circumstances (extraspection) (Doris 61). In brief, in order to exhibit agency, we need to do a little bit of thinking before we act, taking into consideration our own abilities and limitations as well as the constraints and opportunities afforded by our circumstances.

With reflectivism on the table, Doris can then clearly explain the skeptical argument against personal self-direction. The skeptical argument relies on the fact that empirical observations appear to have demonstrated that "human beings suffer impoverished self-awareness" (Doris 62). Under reflectivism, accurate self-reflection is required for agency. Yet the empirical data indicate that our capacity for self-reflection is strikingly deficient, suggesting that we "routinely fail the requirements of agency" (Doris 62). If we routinely fail these requirements where we previously thought we demonstrated accurate self-reflection, then we are left uncertain as to how deep the rabbit hole of impoverished self-awareness really is. This uncertainty gives rise to the skeptical argument, which Doris relates in modus ponens form as follows:

> If reflectivism is true (and the empirical observation is defensible), it is not warranted to attribute human beings personal self-direction.
>
> Reflectivism is true (and the empirical observation is defensible).
>
> It is not warranted to attribute human beings personal self-direction.

What he endeavors to establish in his 2009 paper is the parenthetical premise that the empirical observations are defensible, thereby demonstrating that the skeptical challenge ought to be taken as a serious challenge to the ability of human beings to function as persons.

Doris reports empirical observations from literature on the so-called 'automaticity' of cognition. He notes that while disagreement about the interpretations of the standard literature exists in psychology no less than in philosophy, this is not problematic for him. A successful skeptical argument does not require a definitive interpretation of the literature. Rather, it requires only that the interpretation of the empirical observations it employs be defensible, meaning that they be plausible. And so he proceeds to provide the defensible interpretations of several empirical observations. He claims that all of these interpretations invite the same conclusion: “most of a person’s everyday life is determined not by their conscious intentions and deliberate choices but by mental processes that are put into motion by features of the environment and that operate outside of conscious awareness and guidance” (Doris 63). I will assume familiarity with the empirical literature, but I will briefly note that Doris’ tour touches upon the “fragility” of evaluation, studies on implicit prejudice, and implicit egotism.

The brief tour helps to clarify that the skeptical worry can be made more precise. “Cases where actors would reject the causes of their behavior as reasons for their behavior” (Doris 66) serve as defeaters for reflective self-direction. For any proposed instance of responsible agency, such a defeater explanation can be offered in opposition to an explanation that invokes reflective self-direction. Doris indicates that while skeptical arguments in general may be difficult to take too seriously, the skeptical worry about reflective self-direction is a “live” hypothesis judged by experts to be at least as likely as alternative hypotheses. At this point, Doris seems to have established that the skeptical worry is something to be concerned about because, if reflectivism about persons is true, the evidence seems to suggest that personhood is much more scarce than we typically think. In the remaining sections, he proceeds through responses to the worry in order to demonstrate that it will not be easily dismissed. Having explained and argued for the skeptical challenge, as well as having defended it from four potential responses, Doris concludes by suggesting what we might do to address the challenge. Rather than maintaining skepticism about persons, we can adjust the skeptical argument by transforming it into a modus tollens argument against reflectivism as follows:

> If reflectivism is true (and the empirical observation is defensible), it is not warranted to attribute human beings personal self-direction.
> It is warranted to attribute human beings personal self-direction.
>
> ---
>
> Reflectivism is not true.

In summary, while he has argued for the skeptical challenge in order to demonstrate the importance of addressing it, Doris does not align himself with skepticism about persons. Instead, he proposes that the empirical literature suggests "increased skeptical scrutiny of reflectivism" (Doris 79). In the next section, I will introduce Meyers' response to Doris' skeptical argument. It is the response that I will ultimately reject.

2.3 Summary of Meyers

In Automatic Behavior and Moral Agency, C.D. Meyers argues that the skeptical worry introduced by automaticity is not as great a threat to the traditional notion of persons and their agency as Doris' argument would make it seem. Meyers identifies two assumptions in Doris' skeptical argument and attempts to address both. In what follows, I will focus on Meyers' response to the second assumption.

Meyers charges that Doris' account assumes a realist conception of persons. He explains that this follows from the assumption that if human beings lack personal functioning, "then we must abandon our moral concept of persons and reject person-based ethics" (Meyers 204). In response to this, he argues that, if we accept a non-realist conception of persons, we need not abandon the moral notion of persons, and, moreover, "there are good reasons why we should retain the concept of personhood in our moralizing, independently of whether agency really exists" (Meyers 194).

Again, Meyers explains that to accede to the assumption that failing to exhibit personal functioning entails abandoning person-based ethics is to be realist with respect to personhood. By pointing this out, Meyers forces us to consider our metaethical commitments concerning personhood. He claims that doing so shifts the question from whether we exhibit reflective self-direction to whether its absence entails that a moral concept of persons must be abandoned. If we maintain realism with regards to our notion of persons, then the concept of "person" must pick out some mind-independent object or set of properties whose existence is not dependent on our moral attitudes or practices. He explains that while moral non-realism maintains, as does realism, that moral properties supervene on natural properties, it does not hold this fact to be a mind-independent truth. He then contends that if we accept a non-realist metaethics, then "even if human beings do not consciously control their actions or choose on the basis of deliberations, we might still have reason to preserve a moral concept of personhood that supervenes on something other than conscious control or deliberate choice" (Meyers 204). In answer to this, Meyers endorses Strawson's approach of conferring personhood in roughly those cases where an individual may be appropriately subject to reactive attitudes.

Meyers notes that one weakness in this approach is that Strawson gives no account of why we should continue to adopt different attitudes toward children and the insane than we do toward normal adults. Strawson appeals to our practices, but he does not attempt to justify them. So we are left to ask ourselves why we should exclude the incapacitated or children from full personhood or moral responsibility. Meyers resolves this by following Tim Scanlon in proposing that we determine the appropriateness of reactive attitudes by taking their appropriateness to be the product of constructing social norms. Reactive attitudes are appropriate where they are products of contracting between, not ideally rational and fully informed individuals since these do not exist, as is evidenced by the literature on automaticity, but semi-rational and semi-informed individuals. Having addressed this problem with Strawson's account, Meyers takes himself to have demonstrated that a non-realist understanding of persons would allow us to retain the notion of persons in our moralizing.

Before moving forward, I will briefly summarize four conclusions I have drawn from the debate. First, I am convinced by Doris' argument that automaticity reveals the incredible influence of the environment on our behavior. Clearly, our behavior is not only influenced by the environment, but it is influenced by the environment outside of our awareness and to a greater extent than previously thought. Secondly, the skeptical argument rightly takes into account the fact that automaticity does threaten or significantly affect the attribution of moral responsibility. So long as we think of personhood and agency in terms of reflective self-direction, the facts concerning automaticity of behavior should affect how we attribute moral responsibility, whether we are realist or not. Thirdly, Meyers is correct that a non-realist metaethics would allow us to retain a moral notion of persons, but his response does not adequately account for how automaticity should affect the way we attribute moral responsibility. Attributions of moral properties ought to be in some way responsive to empirical evidence about the way in which we actually function. Finally, Meyers assumes there are good reasons to continue employing a moral notion of persons, but he assumes these reasons exist without providing or defending any. It is with this final point in mind that I will present my objection to Meyers' non-realist response.

2.4 Signpost

Recall that Doris presented an argument for a thoroughly skeptical view of persons in order to demonstrate the strength of the skeptical challenge, then suggested we reject reflectivism. Then Meyers broke Doris' argument into two assumptions, and I explained Meyers' response to the second, the assumption of a realist conception of personhood and agency. What I aim to do in the remainder of this paper is to demonstrate that Meyers' non-realist response cannot be assumed to salvage a moral notion of persons. I will argue in opposition to Meyers that there are good reasons for us to significantly reduce our use of a moral notion of persons in attributing moral responsibility, quite independent of its actual existence.

Both Doris and Meyers proceed by using the appropriateness of reactive attitudes as an indication of the appropriate attribution of moral responsibility. Where it appears appropriate to attribute moral responsibility, agency and personhood are also appropriately attributed. We might easily concede that people can be mistaken in the attribution of moral responsibility, but this does not imply that we ought to reduce our use of it, especially where it does appear appropriate. I will argue, however, that the reactive attitudes relied upon for the attribution of moral responsibility on Meyers' non-realist account are not trustworthy because their evolutionary function leaves them highly suspect, and the repercussions are manifest. I draw support for the claim that reactive attitudes are suspect from an unpublished manuscript by Kyle Stanford entitled “The Difference Between Ice Cream and Nazis: Evolution, Cooperation, and the Function of Moral Externalization.” I will first give an overview of Stanford’s account, then I will explain the function of reactive attitudes on his account. Following this, I will explain what I take to be the negative implications of the process described by Stanford. I take these negative implications to demonstrate that we cannot take it for granted that we have good reasons to continue employing a moral notion of persons independent of its actual existence, again my concern being the grounding of attributions of moral responsibility.

2.5 Stanford’s Account of Moral Externalization

The question that Stanford sets out to answer is why we treat the demands of morality as anything more than a matter of our subjective preferences for certain behaviors. According to Stanford, while other animals exhibit prosocial behavior, they do not appear to externalize the demands of morality, which would include judging reactive attitudes as deserved, certain acts as moral transgressions, or certain behaviors to be appropriate. Subjective preferences such as pain can be strongly motivating, so, according to Stanford, the need for effective motivation does not explain why we externalize moral demands. Were some motivation to gravitate toward or away from certain types of behavior all that were necessary, we could treat all such preferences as merely subjective.

To explain why we externalize moral demands, Stanford introduces the idea of biological altruism, exploitable cooperative behavior. He explains that humans, unlike other animals that exhibit prosocial behaviors, appear to be default, domain-general cooperators. He cites multiple studies in which children exhibit helping behavior towards those perceived as members of a well-defined in group. In one such study, Warneken and Tomasello 2006, the researchers found that children as young as 18 months would reliably assist an adult stranger in achieving a goal, such as opening a cabinet if the adult's hands were full, in the absence of any encouragement or reward and without the adult stranger asking for help or so much as looking at them (Stanford). What these studies demonstrate, according to Stanford, is that while we can learn when it is imprudent to cooperate with another individual, our default is cooperation.

As a result of being default, domain-general cooperators, humans have an enormously elevated risk of exploitation and cannot afford to learn with whom to interact by simple trial and error. Humans need to be able to determine in advance who can be trusted to cooperate. Thus, Stanford proposes we demand that others share our views in order to count as desirable cooperative partners. We must demonstrate that we share these views not only in our actions but also as third party observers or in discussing hypothetical cases. In order to count as a cooperative partner, an individual demands of potential partners that they not only share the same attitudes toward certain behaviors, but that they also share attitudes towards those who support or tolerate such behaviors. This serves as an indication of whether or not an individual is likely to be an exploitative partner. Thus, the punishment of non-cooperators consists solely in their exclusion from cooperative interaction by an ever wider range of community members, and the only cost, unlike altruistic forms of punishment, consists in each member simply protecting themselves from exploitation (Stanford). As a result, externalization of moral demands is a very efficient way for default cooperators to preempt exploitation.

2.6 Objection to the Non-Realist Reply

Before explaining the negative implications of the process described by Stanford, I will briefly explain where reactive attitudes fall into his account. What is tested in order to be a cooperative partner is not only one’s behavior, but also the perceived appropriateness of one’s reactive attitudes. Stanford says that we demand others share our views in order to count as desirable cooperative partners, and it is likely that the term “views” here encompasses more than reactive attitudes, but reactive attitudes appear to fall suitably under the term. I suggest that on Stanford’s account the reactive attitudes we express come under scrutiny by potential cooperative partners. Potential cooperative partners demand that we share their reactive attitudes in order to count as desirable cooperative partners. Thus, the ability to be accepted by the in-group is dependent upon expressing the accepted reactive attitudes. Additionally, I assert that being associated with individuals who are subject to reactive attitudes will affect whether one counts as a desirable cooperative partner. I will explain this last suggestion shortly.

If our reactive attitudes do come under scrutiny in this way on Stanford’s account, then the reactive attitudes underpinning the appropriate attribution of moral responsibility are subject to the process Stanford describes. Doris and Meyers both follow Strawson in his claim that the appropriateness of reactive attitudes determines the appropriate attribution of moral responsibility[1]. To be held morally responsible is to be appropriately subject to reactive attitudes. What I intend to demonstrate is that the function that reactive attitudes appear to play on Stanford's account is exclusion from the in-group, and that this entails many negative implications. If we are non-realist about the moral notion of persons, and eventual inclusion with the in-group and correction of misbehavior is our goal in the contracting of moral norms, then this gives us reason to err on the side of skepticism with regards to the moral notion of persons, insofar as attributions of moral responsibility are concerned. I contend that Stanford's account should give us cause to worry about employing reactive attitudes, and since these act as the way in for attributing moral responsibility, we should worry about the extent to which we use the moral notion of persons to ground attributions of moral responsibility.

Stanford maintains that moral demands combine elements from each side of a division between merely subjective reactions to states of the world (such as liking the taste of ice cream) and objective reactions to states of the world (such as seeing a predator) that must be correct on pain of the consequences of being incorrect about them. Since moral demands are a hybrid of this division, moral cognition is likely an imperfect adaptation, but it need not be perfect in order to confer a selective advantage. If we are realist with respect to moral responsibility, then the potential for such imperfection might shake our confidence in our ability to appropriately attribute moral responsibility. However, if we are non-realist, it makes little difference because it is presumed that there is no mind-independent fact of the matter. A worry for Meyers' non-realist reply can be

[1] Doris explicitly expresses reservations about Strawson's account, noting that he follows Strawson "loosely."

found, as I said earlier, in whether or not this process is aligning with the contracting of norms between semi-rational persons. If we value eventual inclusion of moral transgressors, then Stanford's account makes it appear that the function of our reactive attitudes and attributions of moral responsibility is in opposition to the norms we may hope to adopt.

My main concern with this process is the scope of individuals it affects. Stanford explains that selective pressure to identify undesirable cooperative partners is self-reinforcing; it becomes stronger as the pool of undesirable partners already rejected by the community increases. This might seem a desirable feature on first gloss. The more exploitative partners there are, the more important it becomes to detect them and separate them from the in-group and the benefits found therein. However, when this process plays out on a large scale, the in-group with the most power and resources may too easily and unjustly exclude all those that do not or cannot participate in the norms it treats objectively. As a result, both norm violators and those associated with them are excluded from the benefits of inclusion and cooperation. For most humans at this point in our evolutionary history, exclusion from the benefits of cooperation with the in-group that has access to the most resources leads to suffering, but not to death. At least not before procreation. As a result, exclusion from the in-group for one generation has consequences for members of the next. What begins as punishment for potentially exploitative individual partners then becomes punishment of entire groups merely by association with the out-group.

The previous point hinges on the association of individuals with those already subject to reactive attitudes, which I mentioned above I would explain here. I have extrapolated this from Stanford's discussion of gossip, which I maintain gives us insight into the evolutionary roots of stereotyping and implicit bias. I suggest that in anticipation of avoiding exploitative partners, we not only absorb gossip about individuals, but we generalize anecdotes or messages delivered by the media about individuals into stereotypes about entire groups. Just as we exclude individuals who would so much as endorse or demonstrate leniency toward behavior in violation of the in-group norms, we exclude individuals that we would associate with norm violating stereotypes.

My claim is that we subject individuals to reactive attitudes and exclude them from the in-group for merely being associated with individuals we have already subjected to reactive attitudes. In other words, reactive attitudes may proliferate in an uncontrolled manner as the result of their evolutionary function, which is to preempt exploitation by excluding potentially exploitative partners. Again, where I extrapolate from Stanford's account is in saying that we exclude potential cooperative partners, not only based on committing moral violations, not only for endorsing or taking lenient attitudes toward such behavior, but, I contend, for being in any way associated with such behavior. When an individual exhibits behaviors or attributes associated with norm violating behaviors, we find reason to exclude them and may be more inclined to attribute moral responsibility to them in virtue of this association. As a result, the in-group with the most power and resources exhibits oppression of those excluded from the group at an institutional level.

On my understanding of Stanford's account, moral externalization preempts exploitation by excluding members based on norm violation and mere association with norm violation, and this is problematic because it excludes on the basis of anticipated exploitation or norm-violating behavior, not its performance. The deployment of reactive attitudes and the attribution of moral responsibility in principle help individuals avoid exploitation or harm, but in practice they fail to acknowledge the systemic problems that give rise to undesirable behavior in the first place. And to fulfill their evolutionary purpose - that is, to prevent exploitation so that humans may be default, domain-general cooperators - there is no need for them to do so. What happens in practice is that those in violation of the norms externalized by the in-group with the most resources are excluded, and then forced to make due in their own group with fewer resources and less power. As more people are excluded from a privileged in-group for mere association with the objectified norm violations, the selective pressure to identify undesirable partners increases. What we end up with is systematic distrust on the basis of mere association with perceived moral violations.

Given that my extrapolation from Stanford's account is viable, we can see that it cannot be taken for granted that non-realists still have good reasons to employ the moral notion of persons. If our reactive attitudes and attributions of moral responsibility can so easily run amok and, as the literature on automaticity suggests, human beings do not actually function in the manner supposed to ground attributions of moral responsibility, then my interpretation of Stanford's account suggests that we have good reasons independent of its actual existence to significantly reduce our use of the moral notion of persons, insofar as it is used to ground attributions of moral responsibility.

2.7 Conclusion

In summary, I began by introducing the skeptical argument about persons presented by Doris. I then presented C.D. Meyers' response that a moral notion of persons is not incompatible with automaticity[2]. I explained that he proposes a non-realist interpretation of moral agency wherein we have good reasons to retain a moral notion of persons independent of its actual existence. I then assessed the debate between Doris and Meyers, arriving at four conclusions.

In response to Meyers, I argued that there are good reasons for us to significantly reduce our use of a moral notion of persons. I presented Stanford's evolutionary account of moral externalization, and, extrapolating from his account, argued that the function of reactive attitudes and attributions of moral responsibility contributes to stereotyping, implicit bias, and exclusion. While Meyers maintains that there are good reasons for us to retain a moral notion of persons if we are non-realist about personhood, the implications of Stanford's account demonstrate that there are also good reasons to reduce our use of the moral notion of persons. As a result, Meyers' non-realist response cannot be assumed without further argument to have salvaged the moral notion of persons.

2.8 Looking Forward

While I have here argued that skepticism cannot be dismissed by adopting a non-realist understanding of persons, I believe there is more to be said in favor of skepticism about moral responsibility. Skepticism based on the evidence for automaticity points our

[2] It should be noted that Doris agrees with this point, as he gives an account of a moral notion of persons that is compatible with automaticity, the dialogic account, in his book Talking to Ourselves.

attention toward the very real environmental influences on our thoughts and behavior. While the evidence can be taken to undermine the entire notion of moral responsibility, I suggest that, rather than extinguishing moral responsibility, skepticism of a kind not quite as thorough as that proposed by Doris may instead push us to redistribute moral responsibility, or at the very least rethink how and when we attribute it. More specifically, the lesson of automaticity is that we should seriously reconsider our natural tendencies to attribute moral responsibility solely on the individual. While I have not argued this claim here, I propose that in helping us approach what Strawson calls the objective attitude, a weaker skepticism motivated by the empirical data stands to help us correct morally impermissible behavior by making changes to the environment, whereas the attribution of moral responsibility leads (both intentionally and unwittingly) to inflicting punishment.

Skepticism about agency and persons loosens our trust in our attributions of moral responsibility by acknowledging the systemic causes of behavior. Shaking our trust in these attributions may allow us to move away from exclusion-oriented behaviors that ultimately result in injustice and inequality, toward addressing the systems that give rise to undesirable behaviors in the first place. Accepting a weaker skepticism could help us focus on developing environments that include members historically excluded institutionally and generationally from the privileged in-group so that they can benefit from cooperation with the privileged in-group.

CHAPTER 3. NICHOLS' CHALLENGE AND THE DARK SIDE OF REACTIVE ATTITUDES

> "He who fights with monsters might take care lest he thereby become a monster. And if you gaze for long into an abyss, the abyss gazes also into you" (Beyond Good and Evil, Aphorism 146).

3.1 Introduction

Since P.F. Strawson's influential essay, "Freedom and Resentment," many compatibilists about moral responsibility have appealed to the reactive attitudes to justify maintaining our current practices surrounding moral responsibility. In his 2007 paper "After Incompatibilism: A Naturalistic Defense of the Reactive Attitudes," Shaun Nichols argues that we have reasons to retain the reactive attitudes despite our incompatibilist intuitions even if we assume that the universe is deterministic. One of the reasons Nichols maintains that we should retain the reactive attitudes is the role they play in facilitating cooperative behavior. He suggests that an argument from fairness might give us reason to favor the incompatibilist intuition over the intuition that people are morally responsible for their actions, but that, as of yet, the argument from fairness remains too underdeveloped to justify a revolution in our practices. By "argument from fairness," Nichols appears to mean arguments which maintain that it is unfair to attribute moral responsibility in a deterministic universe. My aim is to propose, not an argument from fairness, but a debunking argument against the reactive attitudes which requires no appeal to the incompatibilist intuition and would hold regardless of determinism. In this introduction, I will provide a brief outline of my approach.

To begin, I Nichols uses empirical data to defend the Strawsonian appeal to the "gains and losses to human life." What both he and Strawson argue, each in their own way, is that the benefits of the reactive attitudes far outweigh the costs of ignoring our incompatibilist sentiments. Nichols defends this by maintaining that we should not attempt to suppress the reactive attitudes because of the role they play in facilitating cooperative behavior. Even if the reactive attitudes and our practices surrounding moral responsibility could be influenced by the incompatibilist intuition, they *should not* be. The losses to human life would be too great.

Nichols' acknowledges that we appear to have a universal incompatibilist intuition, but he maintains that arguments from fairness motivated by the incompatibilist intuition remain too underdeveloped to justify a revolution in our practices, especially when we consider the benefits of moral anger in facilitating cooperation. He notes, however, that "a more fully developed argument from fairness might provide an overwhelming moral reason to suppress the reactive attitudes" (Nichols 424). This remark is what I have termed *Nichols' Challenge,* and responding to this challenge is the aim of this paper.

In answer to this challenge, I propose a problem for those compatibilists that appeal to the benefits of the reactive attitudes to justify our practices surrounding moral responsibility: relying on the reactive attitudes results in unfair attributions of moral responsibility as a result of intergroup bias. I first provide empirical evidence for this problem, then I argue that this problem does in fact provide overwhelming moral reasons to suppress the reactive attitudes because, not only are the reactive attitudes influenced by intergroup biases, but they reinforce intergroup biases. Ultimately, this should pose a problem for those compatibilists who justify our practices associated with moral responsibility by appealing to the benefits of the reactive attitudes without thoroughly assessing their costs.

Following this criticism, I conclude by offering two positive approaches to the problem I outline. First, I propose that, rather than using the reactive attitudes as a category to justify moral responsibility, we instead treat each reactive attitude in turn, exploring under what conditions each is either beneficial or harmful. Then I propose a notion of “suppress” capable of flexibly incorporating the information we garner concerning harms and benefits. This notion should allow us to determine the best ways to implement pragmatic approaches that emphasize the benefits of each reactive attitude and reduce their harms.

3.2 Lay of the Land

I will begin by broadly contextualizing the debate into which I am entering. Looming in the background of Nichols’ paper is the debate between free will and determinism and their relation to moral responsibility. A review of terminology relevant to this debate will help facilitate the discussion that follows. The first distinction relevant to the discussion is that between compatibilism and incompatibilism. Compatibilism, on the one hand, is the notion that free will is compatible with determinism (McKenna). Incompatibilism, on the other hand, is the notion that if determinism turns out to be true, we would be without free will (Vihvelin).

Free will is held by many to be a requirement for moral responsibility. One may be held morally responsible for one's actions when they are an expression of one's choosing to conduct them, while one may not be held morally responsible for actions they did not choose to perform. If I rob a bank of my own volition, I will be considered morally responsible for that action; if I rob a bank because someone has threatened to harm my family, then I may be considered either less morally responsible or not at all. However, if the universe is deterministic, then it could be argued that there is no difference between the two cases; robbing the bank in the first case is no different than being forced because in neither case did I choose to rob. This is an expression of the incompatibilist intuition. In response to this, a compatibilist will typically argue that moral responsibility is compatible with determinism, and so we can still distinguish between the two cases of robbery.

Our next concern is what it means to be held morally responsible. To be morally responsible for an action is "to be worthy of a particular kind of reaction—praise, blame, or something akin to these—for having performed it" (Eshleman). In other words, to be morally responsible for an action is to be the proper subject of reactive attitudes for having performed it. The term "participant reactive attitudes" is derived from P.F. Strawson, and they may be understood as "natural attitudinal reactions to the perception of another's good will, ill will, or indifference" (Eshleman). Examples of such attitudes include *admiration* for acts considered morally praiseworthy and *resentment* or *moral anger* for acts performed with malicious intent. Moreover, reactive attitudes are "expressed from the stance of one who is immersed in interpersonal relationships and who regards the candidate held responsible as a participant in such relationships as well" (Eshleman).

In his essay, "Freedom and Resentment," Strawson puts forth a compatibilist argument involving two considerations. The first consideration is an appeal to insulationism. Strawson claimed that the theory of determinism has no bearing on whether or not we employ reactive attitudes in our interpersonal relationships, and in this way the reactive attitudes are insulated from the theory of determinism. Even a hard determinist, someone who holds that determinism is true and that no persons have free will, will still employ reactive attitudes in their interpersonal relationships, holding others morally responsible for their actions (McKenna). The second consideration is an appeal to the "gains and losses to human life" tied up in the reactive attitudes. With this second consideration, Strawson is in some sense arguing that the benefits of the reactive attitudes far outweigh any incompatibilist sentiments we might have, and, as a result, even if the reactive attitudes and our practices surrounding moral responsibility could be influenced by the theory of determinism (i.e. the appeal to insulationism were false), they *should not* be. The losses to human life would be too great. With the terminological and historical context in place, we can proceed to Nichols' empirical defense of these Strawsonian considerations.

3.3 Nichols' Challenge

In this section, I will explain the details of Nichols' argument and their relation to the debate outlined. In defending the practices surrounding moral responsibility, Nichols examines the two Strawsonian considerations previously mentioned, supporting the claims with empirical data. Nichols explains that we are committed to incompatibilism about moral responsibility, as it appears to be a universal folk intuition; however, he sets out to prove that we may and *ought to* ignore this commitment and continue practices that depend on moral responsibility (Nichols 405). He proceeds by assuming that determinism is true then answering two questions about the practices that surround moral responsibility, one descriptive and the other prescriptive: (1) would belief in determinism alter our practices and (2) *should* the truth of determinism lead to revision of our practices? Nichols' answer to both questions is "no," but I will focus on the second question.

Nichols explains that incompatibilism seems to imply that determinism *should* change people's behavior. However, not all incompatibilists promote such a revolution in our practices, so he coins the term "revolutionaries" to describe those incompatibilists that call to extinguish the practices and attitudes that rely on a belief in moral responsibility. In attempting to combat the argument of the revolutionary, Nichols explores the two Strawsonian considerations against revolution mentioned in the previous section. The first is an appeal to insulationism about reactive attitudes — that is, the reactive attitudes are not subject to revision upon considering theoretical viewpoints such as determinism. The second consideration is an appeal to "gains and losses to human life." Nichols believes the first appeal is instructive although it will not necessarily defeat the revolutionary's argument, but that the second appeal provides powerful reason to oppose a revolution in our practices.

With respect to the second appeal, Nichols endeavors to argue that a revolution in the practices associated with the reactive attitudes would generate significant losses to human life. He explains that “[t]he Strawsonian claim of interest is that our lives would be greatly diminished if we uprooted our reactive attitudes” (Nichols 417). Nichols notes that hard incompatibilists such as Pereboom have recently attempted to relieve these worries by arguing that many of the positive reactive attitudes, such as love, are not affected by hard incompatibilism. Nichols explains that Pereboom’s strategy is to claim that the reactive attitudes targeted by the revolution either are not necessary for good interpersonal lives or there are analogues that can replace them (Nichols 417). Nichols considers both in his refutation, taking moral anger as his paradigm.

Nichols begins his account of the benefits that moral anger adds to human life by admitting that from a rational choice perspective, moral anger appears to be irrational. The behavior it motivates generates no immediate benefits for anyone involved. However, he explains that if we take a longer view, the benefits of moral anger appear more clearly. He states that moral anger benefits individuals in the long term because it “signals intolerance for mistreatment, and this plausibly discourages mistreatment” (Nichols 417). Nichols explains that the benefits of moral anger extend to groups in addition to individuals by turning to research in experimental economics. He focuses on public goods games, which have consistently found that anger-driven punishment significantly affects the behavior of participants. In particular, he states that “Fehr and colleagues have consistently found increased cooperation when punishment is an available option” (Nichols 418).

Based on the results found in this literature, Nichols arrives at three conclusions about the relationship between punishment and cooperation. The first of these conclusions is that cooperation deteriorates without the option of punishment. The second conclusion is that cooperation increases with the mere knowledge that punishment is an available option. The third conclusion is that “punishment pushes cooperation near ceiling” (Nichols 419). Nichols takes these experimental results to suggest that moral anger continues to play an important role in securing cooperative behavior.

Nichols returns to Pereboom in his next section, explaining that although moral anger has benefits for both the individual and for groups, the revolutionary may still respond that moral anger can be replaced by an analogue capable of fulfilling the functional role. He explains that on Pereboom’s view the role of moral anger can be served largely by moral sadness. Nichols’ contention is that moral sadness does not affect our behavior for the most part, and so cannot play the role of moral anger in producing behavior that discourages cheating, defecting, and mistreatment. Nichols also considers regret as a substitution for guilt and resolve based on a commitment to doing what is right and opposing wrongdoing, neither of which he finds satisfactory. Since there is no adequate replacement for moral anger, Nichols concludes that there would be significant losses to human life if it were removed from our lives.

Drawing closer to the aim of this paper, Nichols then responds to the objection that we sometimes accept significant losses based on overwhelming moral concerns, in this case the overwhelming moral concern is fairness. Nichols outlines a basic structure for the argument from fairness as follows:

1. It’s unfair to hold people in a deterministic universe morally responsible for wrongdoing.
2. Our universe is deterministic.

---

3. It’s unfair to hold people morally responsible for wrongdoing (Nichols 422).

Nichols maintains that the first premise is grounded in intuition, and he suggests that the import of this intuition, which arguments from fairness have not been explicit about, could be interpreted as objectivist, folk theoretic, or required for reflective equilibrium. I will focus, as he does, on the folk theoretic interpretation of this intuition, that this "intuition reflects a plank in the folk theory of fairness" (Nichols 423).

Nichols revises the first and third premises based on the folk theoretic interpretation of the intuition as follows:

> 1. On the folk notion of fairness, it's unfair to hold people in a deterministic universe responsible.
> 2. Our universe is deterministic.
>
> ---
>
> 3. The folk view has the consequence that it's unfair to hold people responsible for their actions.

After revising the argument from fairness in terms of a folk intuition, he explains that this gives rise to a dilemma between two of our folk intuitions:

> i. It's often fair to hold people responsible for their actions
>
> ii. It's never fair to hold people in a deterministic universe responsible (i.e. the incompatibilist intuition) (Nichols 423).

As Nichols sees it, "it will require a substantial argument to show that we should give up the first view, [and] we have yet to see such an argument" (Nichols 424).

Nichols concludes by summarizing that his aim was to demonstrate that the fairness argument remains too underdeveloped to justify a revolution in our practices but that "a more fully developed argument from fairness might provide an overwhelming moral reason to suppress the reactive attitudes" (Nichols 424). This concluding remark is what I have termed *Nichols' Challenge,* and responding to this challenge is my aim in the remainder of this paper.

My approach in meeting this challenge will be to establish that we do have overwhelming moral reasons to suppress the reactive attitudes. Keep in mind that I will address what it should mean to "suppress" the reactive attitudes in the final section, but it should suffice at this point to note that my proposal for the definition of "suppress" will not be in the terms Nichols has proposed in (i) and (ii) above because the criticism holds regardless of whether the universe is deterministic. In order to demonstrate the unfairness of the reactive attitudes that arises from the nature of their function, I will turn next to an explanation of intergroup-bias. As I have stated, my argument should pose a problem for those compatibilists who rely on the reactive attitudes for justifying the practice of attributing moral responsibility, and it does so regardless of the truth of determinism.

3.4 The Specter of Intergroup Bias

There is an extensive literature on intergroup bias in social psychology, and it is this phenomenon which presents problems for the reactive attitudes. Hewstone et al. define intergroup bias as the systematic tendency to evaluate one's members of one's in-group more favorably than members of the out-group (Hewstone, Rubin and Willis 2002). It can take the form of favoring the in-group as well as derogating the out-group. The authors also clarify that use of the term "bias" involves "an interpretative judgment that the response is unfair, illegitimate, or unjustifiable, in the sense that it goes beyond the objective requirements or evidence of the situation" (Hewstone, Rubin and Willis 2002).

As I have suggested, intergroup bias influences the deployment of reactive attitudes. In what follows, I will provide concrete examples of how intergroup bias influences the reactive attitudes in a manner that leads to unfair moral judgments. My hope is to demonstrate that the reactive attitudes are unfair in virtue of the influence of intergroup bias, and, moreover, that their deployment reinforces this bias.

3.5 The Dark Side of Reactive Attitudes

In this section, my aim is to demonstrate that there are overwhelming moral reasons to suppress the reactive attitudes, independently of determinism. It should be clear, therefore, that I am not addressing the aspect of Nichols' challenge that is responding to the debate between free will and determinism. My sole focus is to motivate suppression of the reactive attitudes, regardless of philosophical commitments or folk intuitions with respect to the debate. Rather than appeal to the incompatibilist intuition, I hope to establish empirically based moral reasons to suppress the reactive attitudes in virtue of their function. While it may appear that I am abandoning the project of the revolutionary, I maintain that my argument gets to the heart of the sentiments that motivate the revolutionary's call to cease the practices surrounding moral responsibility.

We don't have to search very far to see the negative influence of intergroup bias on the reactive attitudes, but it is worthwhile to examine a few illustrative examples. The important point to keep in mind as we proceed through the following examples is that they demonstrate that the formation and expression of reactive attitudes is dependent upon morally irrelevant characteristics of the target of reactive attitudes, such as race or gender. In what follows, I will examine the influence of intergroup bias on the reactive attitudes in the case of racial bias in retributive justice, Islamaphobia, opposition to affirmative action, and gender bias in attributions of merit.

3.5.1 Racial Bias in Retributive Justice

In this section, my aim is to demonstrate that there are overwhelming moral reasons to suppress the reactive attitudes, independently of determinism. It should be clear, therefore, that I am not addressing the aspect of Nichols' challenge that is responding to the debate between free will and determinism. My sole focus is to motivate suppression of the reactive attitudes, regardless of philosophical commitments or folk intuitions with respect to the debate. Rather than appeal to the incompatibilist intuition, I hope to establish empirically based moral reasons to suppress the reactive attitudes in virtue of their function. While it may appear that I am abandoning the project of the revolutionary, I maintain that my argument gets to the heart of the sentiments that motivate the revolutionary's call to cease the practices surrounding moral responsibility.

We don't have to search very far to see the negative influence of intergroup bias on the reactive attitudes, but it is worthwhile to examine a few illustrative examples. The important point to keep in mind as we proceed through the following examples is that they demonstrate that the formation and expression of reactive attitudes is dependent upon morally irrelevant characteristics of the target of reactive attitudes, such as race or gender. In what follows, I will examine the influence of intergroup bias on the reactive attitudes in the case of racial bias in retributive justice, Islamaphobia, opposition to affirmative action, and gender bias in attributions of merit.

3.5.2 Islamophobia

Islamophobia serves as another example of how intergroup bias influences reactive attitudes. Intergroup bias also influences the form and conditions in which reactive attitudes manifest. According to an article published in the Washington Times on November 23, 2015, Tell MAMA (Measuring Anti-Muslim Attacks), a group that monitors Islamaphobia in the U.K., reported that it had been made aware of 115 assaults on Muslims in the eight days following the November 13th attacks in Paris. This was up from 43 assaults reported the year prior in the same time span, an increase of 267% (Blake 2015). The information reported in this article is representative of how negative reactive attitudes manifest as Islamophobic sentiments in the wake of terrorist events. In a research review, Klas Borell discusses sudden, dramatic events and their relationship to prejudicial attitudes and hate crimes towards Muslims. Borell concludes that Islamophobic attitudes are largely "event-driven" and reactive. From Borell's assessment we can infer that negative reactive attitudes such as moral anger manifest as Islamophobic attitudes in the wake of sudden, dramatic events.

To understand how Islamophobic attitudes are an example of how reactive attitudes can manifest under the influence of intergroup bias, we can note how Borell emphasizes that prejudice, including Islamophobia, is a "stereotyped cognition with a strong emotional element" (Borell 411). Borell maintains that it is important to understand Islamophobia as not only a cognitive phenomenon, but also as an affective phenomenon. Islamophobia is not simply a matter of indiscriminate generalizations, but also a matter of affective responses. He explains that sudden, dramatic events such as 9/11 and the terrorist attacks in London in 2005 cause Islamophobic prejudice to flare up but then subside again in calmer times. So, based on the manner in which Islamophobic attitudes present themselves, we can surmise that Islamophobic attitudes are themselves a manifestation of the reactive attitudes as influenced by intergroup biases.

3.5.3 Endorsement of Meritocracy and Desert

Intergroup bias also influences whether reactive attitudes are perceived as appropriate. Cuddy et al. (2009) demonstrated that across multiple cultures, both individualist and collectivist, successful groups are seen as deserving of their success while unsuccessful groups are seen as deserving of their failure.

Endorsing the Meritocratic worldview leads high-status group members to feel deserving of their preferential treatment because they attribute the preferential treatment to their abilities, and, as a result, are more sensitive to perceiving signs of reverse discrimination. For example, in a study of Australian students' discussion of race, "liberal principles such as individualism, merit, and egalitarianism were recurrently drawn upon to justify, argue and legitimate opposition to affirmative action" (Augoustinos et al. 2005). Such opposition is frequently referred to as "the New Racism."

The opposite is true of low-status group members who endorse a meritocratic worldview. Because they feel deserving of their lack of preferential treatment, they are less sensitive to perceiving discriminatory treatment (Jost and Hunyady 2003; Major, Gramzow et al. 2002). For example, working women who endorse a belief in a Just World are less likely to report discontent with the status of women in the working environment (Hafer and Olson 1993).

3.6 Why It’s Not Simply a Matter of Eliminating Bias

There is one significant objection to my argument thus far. That objection would contend that resolving the issues I have posed does not require suppressing the reactive attitudes so much as eliminating our biases. Indeed, this appears to be the focus of many attempts to eliminate our implicit and explicit biases. We want to know how to reduce our biases because they have a substantial and unfair impact on outcomes for many social groups. My response is that we still must suppress the reactive attitudes, especially as part of our attempt to combat intergroup bias. This is because intergroup bias not only influences our reactive attitudes, but the deployment of reactive attitudes *reinforces* intergroup bias. What I mean by “reinforce” here is that the reactive attitudes strengthen, magnify, and intensify intergroup biases, thereby making them more difficult to eliminate. Nichols qualifies that we sometimes accept significant losses based on overwhelming moral concerns, but that, as of yet, there is no argument from fairness compelling enough to overcome the benefits of reactive attitudes such as moral anger. My response to Nichols’ challenge is to argue that this contribution to reinforcing inherently unfair attitudes (intergroup biases) provides overwhelming moral reasons to suppress the reactive attitudes.

It should be made clear that the reactive attitudes do in fact reinforce our intergroup biases in the manner that I suggest. This can be seen in a couple of ways. One way of viewing it might be to take the perspective of evolutionary psychology. In his account of moral externalization, Kyle Stanford describes how the selective pressure for in-group members to identify and exclude undesirable cooperative partners is self-reinforcing; it becomes stronger as the pool of undesirable partners already rejected by the in-group increases. This might seem a desirable feature on first gloss. The more exploitative partners there are, the more important it becomes to detect them and separate them from the in-group and the benefits found therein. However, this suggests that the more we display reactive attitudes toward undesirable cooperative partners, the more vigilant we will become about detecting them, about distinguishing in-group from out-group member and exhibiting differential treatment on that basis.

Further support for the claim that the reactive attitudes reinforce and strengthen our intergroup biases is revealed when we note the methodological issues that arise when conducting experiments on minimal groups. In order to study minimal groups in the absence of independent factors, the classificatory scheme for intergroup categorization must be value-neutral, and there must not be group competition or unequal status between groups, as it has been shown that these increase the strength of intergroup biases (Mullen et al., 2002). When opportunities for reactive attitudes are introduced, intergroup bias increases. So, not only are we inclined to exhibit intergroup biases on the mere basis of arbitrary group division, but these biases are intensified when reactive attitudes are introduced. This suggests that addressing intergroup biases more generally will require suppressing the reactive attitudes.

A final piece of support comes from literature on the relationship between affect and bias. Much of the literature has focused on the emotions anger, anxiety, happiness, and sadness. What has been found repeatedly is that anger, anxiety, and happiness increase reliance on pre-existing stereotypes and attitudes toward out-group members. In their literature review on affect as a cause of bias, Wilder and Simon describe the Distraction Hypothesis. This hypothesis suggests that because attention is a zero-sum game, in intergroup situations, affect will distract perceivers from the behavior of out-group members (Wilder and Simon 160). In order to compensate, when experiencing affect, an individual will rely more heavily on well-learned habits such as stereotypes rather than carefully attending to the features of their environment (Wilder and Simon 160).

The authors explain that this hypothesis has found support in the work of Bodenhausen and his colleagues (Bodenhausen and Kramer, 1990; Bodenhausen, Sheppard, and Kramer, 1994) among others. What the researchers found was that when participants experienced either happiness or anger, they relied more heavily on stereotypes and less on individuating or particular information, while sadness had no effect (Wilder and Simon 161). The authors note that the neutrality of sadness may be due to the fact that happiness and anger are "hotter" emotions than sadness (Wilder and Simon 161). It is not difficult to imagine that reactive attitudes are themselves "hotter" emotions that would distract individuals from carefully attending to individuating factors, resulting in greater reliance on stereotypes. It is interesting to note that Nichols' rejection of moral sadness, described on page 7 of this document, is that it is less motivating. This may be due to "sadness" being a cooler emotion, but the reduction in reliance on stereotypes may count in favor of this emotion as an alternative to moral anger.

In any case, these findings and the theories attempting to describe them suggest mechanisms by which reactive attitudes such as moral anger may reinforce intergroup biases. Because anger has been shown to increase reliance on pre-existing stereotypes, we have some ideas already about at least one mechanism by which a reactive attitude such as moral anger might reinforce intergroup biases. Having a potential mechanism by which we can explain how reactive attitudes reinforce intergroup biases should serve as further support for the claim that we have morally pressing reasons to suppress our reactive attitudes. If our reactive attitudes can so easily run amok, then my argument suggests that we have moral reasons *independent of the incompatibilist intuition* to suppress the reactive attitudes. What I have argued has demonstrated that this is the case, at the very least, in intergroup contexts. However, it should be noted that if we restrict suppression of the reactive attitudes to intergroup contexts, difficulty in determining where group lines are drawn will follow, and, as minimal groups theory has demonstrated, merely drawing a group distinction introduces intergroup biases.

3.7 Concluding and Forward-Looking Remarks

In summary, recall that Nichols posed what I called "Nichols' Challenge." This challenge was that we sometimes accept great losses, such as the benefits of moral anger, when we have pressing moral reasons to do so, but it would require a substantial argument to demonstrate that we have overwhelming moral reasons to suppress the reactive attitudes, and we have yet to see such an argument among those arguments from fairness that appeal to the incompatibilist intuition. In response to Nichols' Challenge, I avoided an appeal to the incompatibilist intuition and developed an argument that claimed we have overwhelming moral reasons to suppress the reactive attitudes, both because the reactive attitudes are influenced by intergroup biases in ways that lead to unfair moral judgments and behaviors, and because our reactive attitudes reinforce our intergroup biases. This suggests that we have overwhelming moral reasons to suppress the reactive attitudes, at least in intergroup contexts, but, again, determining when we are in an intergroup context and restricting suppression of the reactive attitudes to that context is likely to be problematic because merely drawing group distinctions itself introduces intergroup biases, as demonstrated by the literature on minimal groups thinking

As previously stated, what I have argued does not settle the dilemma between the incompatibilist intuition and the intuition that people are sometimes morally responsible for their actions, as the issues I have described remain regardless of the truth of determinism. What my argument has done is present a problem for those compatibilists who attempt to justify our practices associated with moral responsibility by appealing to the reactive attitudes. In other words, this argument undermines the compatibilist's appeal to the benefits of the reactive attitudes *as a whole* to justify our practices surrounding moral responsibility.

Nevertheless, we still must keep in mind the first Strawsonian consideration, insulationism. Even if we *should* suppress the reactive attitudes, we still have to keep in mind that we cannot possibly eradicate them. With this in mind, my argument requires a notion of "suppress" that does not demand that we eliminate the reactive attitudes entirely. The first thing I would suggest with my argument in mind is that, rather than lumping the reactive attitudes together in support of the practices surrounding moral responsibility, we instead treat each reactive attitude in turn, exploring under what conditions each is either harmful or beneficial, and developing systematic changes that accentuate the benefits while minimizing the harms.

Once we have a better understanding of the costs and benefits of each kind of reactive attitude, as well as the kinds of conditions that produce or diminish them, it is conceivable that we could develop systemic changes that accentuate the benefits while minimizing the harms. Doing so would handle the notion of suppress that I propose at the societal level. At the individual level, we might find the notion of suppress more difficult to instantiate. However, I believe there is a notion that we can borrow which would mirror what I suggest for the societal level in the individual: mindfulness.

A number of scholars have responded to the negative effects of our automatic affective responses by suggesting we reign them in with rationality, a controlled and superior process. For instance, Paul Bloom suggests that empathy is detrimental in the moral domain and should be replaced by rational compassion. Josh Greene, in his book *Moral Tribes*, suggest that when the moralities of differing cultures conflict, we settle moral issues by engaging in a utilitarian cost-benefit analysis that engages parts of the mind which involve more controlled cognitive processing. My own notion of suppress is not too far from these, but I would like it to be explicitly distinguished from any outright appeal to rationality in an effort to avoid describing one capacity, rationality or affect, as superior to the other.

Given the historical context of the debate between reason and the passions, I propose a notion of "suppress" at the individual level which is more akin to "mindfulness" than to "being rational." I have several reasons for this, involving both historical and physiological issues. First, it has been historically problematic to privilege reason above the emotions, which are often described as more base or primitive. Feminist ethics has done much work to illuminate the problems with theorizing about morality in a way that privileges reason above the emotions, and, though I will not go into detail about those developments here, I will note that I take their criticisms seriously in my current suggestion for a notion of “suppress.” The second issue motivating my notion is that it has been repeatedly demonstrated that suppressing our emotions physically and socially has detrimental effects on health outcomes. For instance, in general men have a more difficult time dealing with depression due to the social pressure not to discuss their emotions with others, and forcing workers to wear the “service smile” for hours on end, day after day, contributes to worse health outcomes. For these historical and physiological reasons, I advise against any notion of “suppress” that suggests privileging rationality as a superior capacity to the emotions in any way.

In my notion of “suppress,” which may be construed as akin to “mindfulness,” feeling and expressing an emotion need not be construed negatively in and of itself. The research I mention on page 18 of this document suggests that the problem I have introduced for the reactive attitudes is not a matter of being more rational than emotional, but a matter of directing one’s attention. Participants relied on stereotypes because of the cost to attention involved in strong affective responses. With respect to reactive attitudes, a significant amount of attention may be spent on determining whether the emotions are justified, as might be the case with a chronic cognizer, or attention may be focused on the object of the reactive attitudes as the cause of the current emotional experience. These forms of directing our attention with respect to the reactive attitudes express the rationality vs. the emotions dichotomy that we may find that we want to avoid.

A notion of mindfulness about the reactive attitudes would not require attention to be focused to an extreme on whether reactive attitudes are *justified.* This would avoid privileging rationality over the emotions, and it would hopefully sidestep the historical and physiological issues that such a suggestion would involve. At the same time, mindfulness could help reduce the tendency for reactive attitudes to affect behavior in detrimental ways. This is because mindfulness could be construed as focusing attention onto the emotional experience itself. Attention could be paid to the physiological experience of an emotion or exploring the environmental factors, in addition to the object of reactive attitudes, that might relate to the emotional experience. Attending to features of the environment that relate to the experience of a particular reactive attitude would be to incorporate knowledge we gain about the reactive attitudes in general, as well as their benefits and harms.

Fully developing my suggestion of "mindfulness" as a potential notion of "suppress" undoubtedly requires further examination, but, in any case, I believe it illuminates the importance of attention, not only to the argument I have made concerning intergroup biases, but also to arguments criticizing affect in the moral domain more generally. I have argued that we have overwhelming reasons to suppress the reactive attitudes, but I conclude by suggesting that this should not be done by calling to eliminate them by somehow increasing our capacity for rationality.

In summary, I agree with both Nichols and Strawson that we are not capable of eradicating the reactive attitudes, and I have argued in this section that eradicating the reactive attitudes would be detrimental. The notion of "suppress" we do adopt should be capable of incorporating the benefits and harms we find associated with each reactive attitude when we study them more carefully by looking at each in turn. I have suggested "mindfulness" as one possibility due to the relationship between stereotyping and attention. However, whatever notion of suppress we should employ must, first and foremost, take into consideration the nature of each kind of reactive attitude more deliberately.

CHAPTER 4. CONCLUSION

In chapters 2 and 3 of this thesis, I have championed a disposition which is skeptical of moral responsibility in virtue of the ways in which our judgments can be skewed by stereotypes and biases. The first chapter did so by arguing in favor of skepticism about persons. The second chapter worked in favor of this disposition by arguing against compatibilist arguments that appeal to the reactive attitudes to justify our practices surrounding moral responsibility.

Once more, the second chapter presents a more sophisticated argument capable of greater flexibility in conceptualizing the debate because it is less restricted by arguing for or against any particular philosophical concept. While the first chapter loosens itself from the job of defending philosophical concepts to some extent, the second chapter removes their hold to a greater extent, allowing the development of a more practical philosophy.

While I have stated that I believe these arguments to be motivated by a skeptical disposition, I believe that each chapter reveals what it is that a skeptical disposition is attending to. This skeptical disposition, rather than being an incompatibilist intuition that only holds in a deterministic universe or in cases in which an agent could not have acted otherwise, is attending to something inherently unfair about our reactive attitudes themselves. While there has been much in the way of philosophical concepts developed to try to understand and argue about the truth of skepticism about moral responsibility, I believe that much of it has been distracting from what this skeptical intuition is in fact doing in our psychology. In these two chapters, it has been my hope that addressing the relationship between biases and skepticism might help illuminate the reason in which skeptical dispositions exist. Whereas sticking strictly to the abstract philosophical concepts that have historically surrounded the debate has obscured the usefulness of the skeptical disposition, I hope that less philosophically restricted arguments that favor it may help illuminate what features of the world the disposition is attending to.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abrams, David and Bertrand, Marianne and Mullainathan, Sendhil. Do Judges Vary in Their Treatment of Race? (May 28, 2013). *Journal of Legal Studies*, Vol. 41, No. 2 (June 2012), pp. 347-383; U of Penn, Inst for Law & Econ Research Paper No. 11-07.

Augoustinos, M., Tuffin, K., & Every, D. (2005). New racism, meritocracy and individualism: Constraining affirmative action in education. Discourse & Society, 16(3), 315-340.

Blake, Andrew. (2015, November 23). Hate crimes against Muslims in U.K. nearly triple after Paris attacks: report. Washington Times. Retrieved from http://www.washingtontimes.com/news/2015/nov/23/hate-crimes-against-muslims-uk-nearly-triple-after/ on November 27, 2015.

Bodenhausen, G. V., & Kramer, G. P. (1990). Affective states trigger stereotypic judgments. Paper presented at the annual convention of the American Psychological Society, Dallas, TX.

Bodenhausen, G. V., Kramer, G. P., & Susser, K. (1994). Happiness and stereotypic thinking in social judgment. Journal of Personality and Social Psychology, 66, 621–632.

Borell, Klas. (2015) When Is the Time to Hate? A Research Review on the Impact of Dramatic Events on Islamophobia and Islamophobic Hate Crimes in Europe. *Islam and Christian–Muslim Relations*, 26:4, 409-421.

Carlsmith, Kevin M. and Darley, John M., Psychological Aspects of Retributive Justice. Advances In Experimental Social Psychology, M. P. Zanna, ed., Vol. 40, pp. 193-236, San Diego, CA, Elsevier, 2008.

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., … Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. The British Journal of Social Psychology / the British Psychological Society, 48(0 1), 1–33.

Doris, John. (2009). "Skepticism About Persons." Philosophical Issues 19: Metaethics: 57-91.

Eshleman Andrew. Moral Responsibility. (Mar 26, 2014). Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/entries/moral-responsibility/

Hafer, Carolyn L., and James Olson. 1993. Beliefs in a Just World, Discontent, and Assertive Actions by Working Women. Personality and Social Psychology Bulletin 19:30–38.

Hewstone M, Rubin M, Willis H. Intergroup Bias. Annual Review of Psychology 53:575-604.

Jost, J.T., Blount, S., Pfeffer, J., & Hunyady, Gy. (2003). "Fair market ideology: Its cognitive-motivational underpinnings." Research in Organizational Behavior, 25, 53-91.

Levy, Neil (2015). Less Blame, Less Crime? The Practical Implications of Moral Responsibility Skepticism. Journal of Practical Ethics. 3 (2): 1-17.

Major, Brenda, Richard H. Gramzow, Shannon K. McCoy, Shana Levin, Toni Schmader, and Jim Sidanius. 2002. Perceiving Personal Discrimination: The Role of Group Status and Legitimizing Ideology. Journal of Personality and Social Psychology 82:269–82.

McKenna, Michael, Coates, D. Justin. Compatibilism. (Feb 25, 2015). Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/entries/compatibilism/

Meyers, C.D. (2015). "Automatic Behavior and Moral Agency: Defending the Concept of Personhood from Empirically Based Skepticism." Acta Analytica Volume 30, Issue 2: 193-209.

Nichols, Shaun. (2007). "After Incompatibilism: A Naturalistic Defense of the Reactive Attitudes" Philosophical Perspectives Volume 21, Issue 1: 405-428.

Nietzsche, F. W., Clark, M., Leiter, B., & Hollingdale, R. J. (1997). Daybreak: Thoughts on the prejudices of morality. Cambridge, U.K: Cambridge University Press.

Rachlinski , Jeffrey J. and Johnson, Sheri Lynn and Wistrich, Andrew J. and Guthrie, Chris. Does Unconscious Racial Bias Affect Trial Judges? *Notre Dame Law Review*, Vol. 84, No. 3, 2009.

Stanford, Kyle. "The Difference Between Ice Cream and Nazis." Fifth Annual Scholl Lecture Series. Grissom Hall 103, West Lafayette, IN. October 8, 2015. Lecture.

Vihvelin, Kadri. Arguments for Incompatibilism. (Mar 1, 2011). Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/entries/incompatibilism-arguments/

Walen, Alec. "Retributive Justice", The Stanford Encyclopedia of Philosophy (Summer 2015 Edition), Edward N. Zalta (ed.), Retrieved from http://plato.stanford.edu/archives/sum2015/entries/justice-retributive.

Wilder, D. and Simon, A. F. (2003) Affect as a Cause of Intergroup Bias, in Blackwell Handbook of Social Psychology: Intergroup Processes (eds R. Brown and S. L. Gaertner), Blackwell Publishers Ltd, Oxford, UK.