7-2012

Rehabilitating the Consequentialist View of Moral Responsibility

Adam Jared Lerner
College of William and Mary

Rehabilitating the Consequentialist View of Moral Responsibility

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of the Arts in Philosophy from
The College of William and Mary

by

Adam Jared Lerner

Accepted for _____
(Honors, High Honors, **Highest Honors**)

_____
Matthew Haug, Director

_____
Neal Tognazzini

_____
Lee Kirkpatrick

Williamsburg, VA
May 1, 2012

Table of Contents

Chapter 1: What a Consequentialist View Ought to Do

1.1 Introduction

The consequentialist view of moral responsibility has been called many things: "The Influenceability Account" (Arneson, 2003, p. 234), "The Economy of Threats Approach" (Wallace, 1994, p. 57), "Hard Compatibilism" (Arneson, 2003, p. 234), "Consequentialist Compatibilis[m]" (Darwall, 2006, p. 65), "Effect Compatibilism" (Smilansky, 2000, p. 33), "Social Regulation Theor[y]" (Watson, 1987, p. 117) and "Deterrence Theory" (Pereboom, 2001, p. 168). It has also been called "simple-minded" (Smilansky, 2003, p. 280), "comically external" (Wallace, 1994, p. 57), and "positively wrong" (Bennett, 2008, p. 52). The consequentialist view of moral responsibility—according to which the concept of moral responsibility essentially has something to do with the beneficial consequences of holding people morally responsible—is nearly universally despised. Although the complaints are numerous, perhaps the most serious objection is that consequentialism fails to capture a central part of our ordinary practices, namely that when we take someone to be morally responsible, we intuitively believe the appropriateness of that judgment to be entirely independent of whether it has beneficial consequences. In this paper I argue, first, that this objection owes its resilience in part to an ambiguity about what theories of moral responsibility are supposed to do. Resolving this ambiguity requires distinguishing between what it takes to *be* morally responsible from what it takes to permissibly *be held* morally responsible. Arguing that plausible consequentialist views of moral responsibility should aim only to provide an account of the latter, I proceed to spell out one such view and defend it against initial objections. I go on to argue that if the central objection described above remains despite this

clarification, then the stalemate between consequentialists and non-consequentialists about holding people morally responsible can be broken only if we take a closer look at the conflicting intuitions that perpetuate it.[1] Drawing on a wide variety of psychological research, I spend the majority of the paper developing an error theory for non-consequentialist intuitions. I then go on to use this error theory to bolster arguments from moral luck for consequentialism about moral responsibility. I conclude by considering potential obstacles to implementing consequentialism about moral responsibility in the real world.

1.2 How Not to Defend Consequentialism about Moral Responsibility

The most influential consequentialist view of moral responsibility on offer was first articulated by J.J.C. Smart in just a few pages at the end of his (1961) *Mind* article, "Free Will, Praise, and Blame". The characteristic feature of Smart's theory is that he attempts to derive claims about what it means for someone to *be* morally responsible from claims about when it is permissible to *hold* someone morally responsible, where the latter depends entirely on whether holding that person morally responsible is likely to influence his future behavior in morally desirable ways. To illustrate, Smart asks us to consider our differential treatment of a lazy student and a "stupid" student. According to Smart, the reason we say that the lazy student is responsible for not doing his homework is because *holding* him morally responsible—by blaming and punishing him—is likely to influence him to do his homework in the future. On the other hand, the reason we say that the "stupid" student is not responsible for failing to do his homework is because "if he is

[1] I use 'intuition' throughout this thesis to loosely refer to beliefs or judgments that do not seem to have been deduced through any explicit reasoning process. That being said, nothing changes if we understand intuitions to be "intellectual seemings" rather than full-blown beliefs.

sufficiently stupid, then it does not matter whether he is exposed to temptation or not exposed to temptation, threatened or not threatened, cajoled or not cajoled. When his negligence is found out, he is not made less likely to repeat it by threats, promises, or punishments" (Smart 1961, p. 302).[2]

Richard Arneson points out an unintuitive consequence of this view often thought so obvious as to not even be worth making explicit. He puts the problem this way:

> …imagine a Mafia thug terrorizes a small village. He commits many heinous crimes. But as it happens any attempt to punish or reproach him will be unsuccessful, will only make him irritable, and hence will lead him to act more brutally. Even self-reproach would have no effect other than to make him more prone than he otherwise would be to angry, immoral outbursts. The influenceability theory then must say that he is not morally responsible for his misdeeds, which seems odd, for no standard excuses exempt him from blame. (Arneson, 2003, p. 248)

Not only would this particular consequentialist view of morally responsibility forbid us from attributing moral responsibility to people who are intuitively morally responsible, but it would have us attributing moral responsibility to people who are intuitively *not* morally responsible:

> In the circumstances described, holding the Mafia guy responsible for his crimes would be mistaken, whereas holding responsible and punishing a mentally retarded and mentally ill person who entirely lacks rational agency capacity might yield good consequences and so be justified according to the influenceability account. (Arneson, 2004, p. 2003, p. 249)

These objections might be avoided if (a) we understand the relevant sense of influenceability required for moral responsibility to be influenceability *in normal conditions,* and (b) we limit the relevant forms of influence to those that could influence only rational agents. Once made sufficiently precise, the former revision might allow

[2] For similar consequentialist analyses of moral responsibility, see Schlick (1966) and Nowell-Smith (1948).

consequentialists to continue attributing moral responsibility to agents like the Mafia guy, and the latter revision might allow them to refrain from attributing moral responsibility to both mentally ill people and other intuitively non-responsible beings like young children and animals.

Nevertheless, even if these revisions could render the resultant theory co-extensive with non-consequentialist theories, consequentialism would still seem to misidentify what it is about morally responsible agents that makes them morally responsible. As Gary Watson observes, "Apart from the question of its extensional adequacy, consequentialism seems to many to leave out something vital to our practice. By emphasizing their instrumental efficacy, it distorts the fact that our responses are typically personal reactions to the individuals in question that we sometimes think of as eminently appropriate reactions quite aside from concern for effects" (Watson, 1987, pp. 257-258). After all, one might believe that it is merely a contingent fact that morally responsible agents are constituted in such a way that certain forms of treatment will generally influence their future behavior in desirable ways. A non-consequentialist could believe that, if people continued to relate to their actions in the same way as they do now, they could still be morally responsible even if they suddenly stopped being influenceable by traditional means.

According to this objection, what it takes for someone to *be* morally responsible has essentially nothing to do with whether holding that person morally responsible would produce beneficial consequences. I believe the consequentialist about moral responsibility should concede this to the non-consequentialist: *being* morally responsible is not in fact a function of influenceability or anything else having to do with beneficial

consequences. What the consequentialist *should* reject is the claim that *being* morally responsible directly implies anything about the permissibility of *holding* morally responsible. But what does it mean to be morally responsible? And what does it mean to hold someone morally responsible? What is the relationship between the two?

1.3 What Do You Mean, "Morally Responsible"?

There are two ways we might go about answering this question. The first is to provide a list of necessary and sufficient conditions that a person must meet in order to be morally responsible. The second is to explain whether we have any reasons to treat morally responsible people differently than people who are not morally responsible, once we have identified who is morally responsible and who is not. In other words, the goal of the first approach is to fully specify "what it takes" to be morally responsible, and the goal of the second approach is to spell out what being morally responsible "gets us" (if anything). Generally speaking, philosophers have neglected the second question in favor of the first.

P.F. Strawson first brought notable attention to the second question in his seminal (1962) paper "Freedom and Resentment". Among a number of important features of Strawson's discussion is his attempt to shift our focus from questions about what it takes to be morally responsible in some obscure metaphysical sense to questions about what our everyday practice of holding one another morally responsible commits us to. Strawson's account begins, then, with a description of our ordinary practice of holding each other morally responsible, which he takes to be constituted by our experience and expression of the reactive attitudes—attitudes we take on in response to our perception of others' good or ill will towards ourselves and meaningful others. The paradigmatic

reactive attitudes are resentment, which we experience in regard to someone who has wronged us, and resentment's cousin, indignation, which we experience in regard to someone who has wronged someone we care about. (Guilt is another closely related attitude which we experience in regard to ourselves when we acknowledge that we have done wrong.)

Strawson's main thesis is that all it means to be morally responsible is to be the proper object of these reactive attitudes, and all it takes to be the proper object of these reactive attitudes is implied by the structure of the psychological mechanisms underlying these attitudes:

> Inside the general structure or web of human attitudes and feelings of which I have been speaking, there is endless room for modification, redirection, criticism, and justification. But questions of justification are internal to the structure or relate to modifications internal to it. The existence of the general framework of attitudes itself is something we are given with the fact of human society. As a whole, it neither calls for, nor permits, an external "rational" justification. (Strawson, 1982, p. 78)

Like the consequentialists that came before him, Strawson attempts to derive facts about when people are morally responsible from facts about when it is appropriate to hold them morally responsible. Unlike these consequentialists, however, Strawson takes the conditions under which it is appropriate to hold people morally responsible through the experience of reactive attitudes like resentment just to be those situations in which people typically do hold each other morally responsible: "situations in which one person is offended or injured by the action of another and in which—in the absence of special considerations—the offended person might naturally or normally be expected to feel resentment" (Strawson, 1982, p. 64).

The crucial point here is that, unlike the consequentialists before him, Strawson succeeded in bringing attention to the fact that moral responsibility is a *normative* concept and that if we want to give an account of this normative kind of moral responsibility, we must first give an account of the kind of normative implications for *holding* morally responsible that we want *being* morally responsible to have.[3] According to Strawson, we want morally responsible agents to be the appropriate objects of reactive attitudes like resentment, guilt, and indignation. Consequently, what makes people morally responsible is just whatever characteristics render them appropriate objects of reactive attitudes like resentment, guilt, and indignation.

Needless to say, Strawson's theory has provoked quite a bit of debate. The disagreement relevant here lies in whether the normative upshot of being morally responsible that we are after simply consists in being an appropriate object of certain reactive attitudes, and, if so, what *makes* certain people the appropriate objects of certain reactive attitudes. The main problem is that Strawson's theory goes directly from our everyday practices of holding each other morally responsible to normative conclusions about the shape those practices should take, all the while neglecting another crucial part of our ordinary practices: considerations about desert.

---

[3] The sense of moral responsibility at play here maps roughly onto what has sometimes been called the 'accountability' face of moral responsibility. (Watson, 1996, p. 288; Fischer & Tognazzini, 2011, p. 383) There is a sense in which other understandings of moral responsibility are still normative, but they are not normative in the sense of telling us whether it is permissible to make others suffer. For example, on what has been called the 'attributability' face of moral responsibility—according to which being morally responsible simply requires standing in the kind of relationship to one's actions that allows others to make valid assessments of one's character—the only sense in which attributions of moral responsibility are normative is that they make use of normative concepts like 'good' and 'bad'. On the attributability face of moral responsibility, what attributions of morally responsibility do *not* do is provide any immediate guidance as to how to treat people who are deemed morally responsible.

Desert is a notoriously slippery concept. People use 'desert' in many different contexts to mean many different things. What is the relevant sort of desert at stake in the debate about moral responsibility? Derk Pereboom has suggested that it is what he calls 'basic desert': "The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve the blame or credit just by virtue of having performed the action, and not, for example, by way of consequentialist considerations" (2001, xx). Unfortunately, even Pereboom's analysis of desert doesn't seem to have removed all ambiguities. Consider Michael McKenna's recent reflections on basic desert:

> As it is… I have no clear sense of what anyone in the free will debate means by desert. Maybe it is basic, but it need not be mysterious. Consider desert in the case of blameworthiness for a wrong done. Does desert entail that it would be a *pro tanto* good that the wronging party suffer? Does it entail not the axiological judgment that it would be good, but the distinct deontic judgment that it is *prima facie* right to cause the wronging party to suffer? Or is it rather the even weaker notion that it is permissible that others do some harm to the wronging party? Here is an even weaker thesis: the wronging party would have no basis for complaint if those whom she wronged, or relevant others, were to cause the wronging party harm merely by expressing to the wrong doer their moral anger and their moral demands for proper redress. (2009, p. 12)

So if what we want is a theory of *desert-entailing* moral responsibility, which of these many understandings of desert should we focus on? This is a good question, and I'm going to ignore it. Indeed, I'm not sure there is a fact of the matter about what 'desert' means, and so I'm skeptical of any attempt to elucidate the concept of moral responsibility that requires us to first elucidate the concept of desert. Consequently, I want to bypass questions about desert *per se* and focus directly on what many of McKenna's suggestions about desert revolve around: the appropriateness of making someone suffer.

As far as I'm concerned, this is what the debate about moral responsibility is (or

ought to be) about: the conditions under which we can permissibly make people suffer. Just as punishment remains a gripping topic in the philosophy of law because of the suffering involved, I believe moral responsibility remains gripping for the very same reason (cf. Wood, 2010, p. 455). Where holding people morally responsible causes suffering, it stands in dire need of justification. Where holding people morally responsible does not cause suffering—such as when one targets another with resentment but never outwardly expresses this attitude—questions about moral responsibility lose their urgency.

One might take exception to my claim that merely targeting someone with resentment cannot count as an instance of making that person suffer. I concede that there are some understandings of well-being under which this would be false. Undoubtedly, if some version of preference-satisfaction is the correct theory of wellbeing, and if people have preferences not to be thought of in particular negative ways, then targeting someone with resentment will seem to count as holding that person morally responsible in a way that makes her suffer. If this is right, all I can say is that I do not care about this kind of suffering. In what follows, I understand suffering to include any decrease in someone's well-being, where that reduction in well-being must be reflected in that person's experience. On this understanding, truly unexpressed attitudes—attitudes that do not even manifest themselves in subtle shifts in behavior around the target of the attitude and of which their target is completely unaware—do not count as acts of holding people morally responsible in ways that makes them suffer.[4]

[4] Indeed, it is hard to see how someone could reasonably object to simply being thought of in a particular way. In one way, merely thinking of someone as morally responsible seems to consist in no more than accurately representing the features of that person's

On the other side of things, one might doubt my claim that ordinary ways of holding people morally responsible—by expressing the reactive attitudes, giving people the cold shoulder, and so forth—cause them to *suffer*. I disagree. First, remember that I am working with an admittedly broad understanding of suffering, one that encompasses any detriment to one's overall level of well-being (however that is to be understood). Consequently, even if publicly shaming someone does not cause them anything like physical pain, it surely makes them suffer in an extended sense, leading them to feel badly about themselves and to have a much less fulfilling day (or week, or month, or year, or life). Likewise, ostracizing people may prevent them from obtaining improvements in their well-being that they would have otherwise obtained. On my admittedly nuanced understanding of suffering, even this counts as suffering. Perhaps a better word for what I'm getting at would be 'harm'. In any case, I will continue to use 'suffer'.[5]

1.4 The Normative Triviality of Being Morally Responsible

By shifting the object of the consequentialist's account to questions about the permissibility of suffering, it should become clear why I think the consequentialist loses little of significance by conceding that she cannot give an adequate account of what it

---

moral agency, perhaps going on to assess her character in a manner consistent with the attributability face of moral responsibility (See footnote 3). But it would be odd to think that someone could suffer simply because other people reflected upon the quality of her character and the relationship between her mental states and her actions. Does this person simply not want negative thoughts to be associated with her name? This seems more understandable, but I still cannot see the sense in which this person would suffer if these negative thoughts were never made known to her, again assuming that these negative thoughts did not have any behavioral consequences that led to shifts in the quality of her conscious experience. Nonetheless, I can see room for disagreement.

[5] For a comprehensive account of the various ways in which expressions of the reactive attitudes lead to suffering in their targets, see Bennett (2002).

takes to *be* morally responsible. According to my framework, simply *being* morally responsible tells us little if anything about being permissibly *held* morally responsible.[6] That is, *being* morally responsible has no direct implications for how we should treat people who fall under that category. Instead, it becomes a merely classificatory concept whose primary function is to serve as an abbreviation for the set of necessary and sufficient conditions an object must meet to satisfy that concept.

If I am right, there is no more to understanding the concept of being morally responsible than knowing the conditions a person must meet in order to satisfy the concept (e.g., acting from a moderately reasons-responsive mechanism of her own[7]), just as there is no more to understanding the concept of being green than knowing the conditions an object must meet in order to satisfy the concept (e.g. producing a certain sort of phenomenal experience in the viewer). In other words, just as the fact that some objects are green imposes on us no direct obligation to treat those green objects in certain ways, the fact that some people are morally responsible imposes on us no direct obligation to treat those people in certain ways.

This treatment of *being* morally responsible faces some immediate challenges. One might object that the concept of being morally responsible seems to do more than serve as an abbreviation for the set of properties that an agent must have in order to be morally responsible. With all due respect to Kermit, one might think that being green is easy; that an object is green surely imposes on us no direct obligation to, say, make that object suffer when it helps bring about bad actions. On the other hand, one might think that being morally responsible is full of normative implications; that someone is morally

[6] For an argument that the converse also holds, see Smith (2007).
[7] For such an account, see Fischer and Ravizza (1998).

responsible seems to provide us with a reason to make that person suffer if she is guilty or at least a reason to refrain from making her suffer if she is innocent.

I agree that it is misguided to *completely* sever being morally responsible from holding morally responsible. But I have proposed doing no such thing. All I have claimed is that being morally responsible has no *direct* normative implications. This still allows being morally responsible to have *indirect* normative implications, such as providing *prima facie* reasons to hold people morally responsible. All that my account requires is that being morally responsible cannot provide any *pro tanto* reasons to hold someone morally responsible—it cannot by itself generate an independently weighty reason to make someone suffer. All it can do is suggest to us that there might be reasons for us to hold that person morally responsible, such as that doing so might bring about beneficial consequences (e.g., deterring that person from doing wrong in the future).

The plausibility of this suggestion should become apparent in what follows. Taking for granted the relative unimportance of being morally responsible, I shift in the next chapter to providing a consequentialist account of when it is permissible to hold people morally responsible in a way that makes them suffer (from here on: "holding people morally responsible"). I consider and dismiss objections as I go, but I spend a disproportionate amount of time addressing one objection. Adequately responding to this objection requires developing an error theory for non-consequentialist intuitions, and this error theory takes up the majority of the chapter.

Chapter 2: For and Against Weak Consequentialism

2.1 A Consequentialist View of Holding People Morally Responsible

According to what I will argue is the most plausible theory of when it is appropriate to hold people morally responsible, justifiably (and permissibly) holding people morally responsible in a way that makes them suffer (by outwardly blaming them, punishing them, and so forth) requires—and only requires—that there is no available alternative that would produce better overall consequences than holding them morally responsible, where the suffering of wrongdoers is just as bad a consequence as the suffering of anyone else (all else being equal).

This short statement obviously requires some unpacking. First, an act counts as an instance of holding someone morally responsible only when carrying out that act makes that person suffer in a sense broad enough to include decreases in well-being attributable not only to traditional forms of punishment but various expressions of the reactive attitudes. This understanding of what it means to hold someone morally responsible was spelled out in detail in the previous section.

Secondly, the present account largely leaves open the question of what counts as a beneficial consequence and how beneficial consequences are to be weighed against negative consequences in evaluating an overall set of consequences. Although I have a specific account of what makes holding people morally responsible *bad*, I hope to remain neutral as to what kinds of consequences should be promoted by holding people morally responsible and how those consequences should be weighed against the morally undesirable consequences of holding people morally responsible. That being said, there are some consequences that a properly consequentialist view of moral responsibility

cannot countenance, namely those that make essential reference to the suffering of the person being held morally responsible (such as "upholding victim's rights"), or those that require discounting the suffering of the person being held morally responsible simply because that person performed an immoral action or harbors a "guilty mind". In any case, I will frame most of my discussion in terms of human well-being, which virtually all plausible moral theories take to be valuable (even as they provide different understandings of what it consists in and how it is to be weighed against other values).

The last clarification to make at this stage is between two different versions of the consequentialist view just proposed. On the strong version, which is the version I ultimately seek to defend, justifiably holding people morally responsible requires *and only* requires that there is no available alternative that would produce better overall consequences than holding them morally responsible (from here on: "would maximize beneficial consequences" or "would be optimific"). In somewhat looser terms, the strong consequentialist believes that maximizing beneficial consequences is both necessary and sufficient for justifiably holding people morally responsible. The weak consequentialist, on the other hand, takes maximizing beneficial consequences to be merely necessary.

Those familiar with the philosophical literature on punishment will recognize that both strong and weak consequentialists reject something analogous to what is often called 'positive retributivism'. While positive retributivism is typically understood as the view that there is always some genuine reason to make the guilty suffer, I expropriate the term here to refer to a thinner view, according to which making the guilty suffer is permissible even when doing so fails to maximize beneficial consequences. So weak consequentialists say that justifiably holding people morally responsible requires that

doing so would in fact maximize beneficial consequences. This implies that holding people morally responsible even when doing so would fail to maximize beneficial consequences is not justifiable. Furthermore, weak consequentialists of the sort I am describing would also deny that holding people morally responsible could even be *permissible* if doing so would not maximize beneficial consequences.

So why would anyone want to be a weak consequentialist? The central thought driving weak consequentialism is that suffering is bad, and that it ought to only be inflicted when the benefits of doing so outweigh the costs. Weak consequentialists might believe this for any number of reasons; their belief might derive from a prior commitment to something like the principle of utility, or it might be motivated by the perceived truth of determinism (i.e., the claim that the past and the laws of the universe together make only one future possible) or the recognition of the pervasive influence of luck on our everyday lives. Later in this paper, I will argue that we know the suffering of the guilty to be bad in just the same way as we know that the suffering of the innocent is bad—through empathy—and that to the extent traditional arguments against moral responsibility are successful, their success comes from their ability to help us empathize with wrongdoers.

My strategy in the rest of this thesis is as follows. First, I defend weak consequentialism against its most serious objection. Next, I argue that the best reasons to endorse weak consequentialism are also reasons to endorse strong consequentialism. I conclude by considering potential difficulties with practically implementing consequentialism about moral responsibility.

2.2.1 Why Not Weak Consequentialism?

The most serious objection faced by weak consequentialism echoes the objection initially faced by traditional consequentialist views of moral responsibility, except while the initial problem with traditional views was just that they would deny that the irremediable Mafia guy could *be* morally responsible, the further problem that even weak consequentialism faces is that it would advise against *holding* the irremediable Mafia guy morally responsible. In other words, it tells us not to blame and punish people who we intuitively feel ought to be blamed and punished.

To get a firmer grasp on the problem with weak consequentialism, consider an alternative case:

> One day, Newman decides that he has had enough. He is going to kill Jerry. He has deliberated long and hard about whether to kill Jerry, and he has concluded that killing Jerry is his only option. That night when Jerry enters the elevator to get to his apartment, he is met by Newman, who is holding a crowbar behind his back. As soon as the elevator doors close, Newman proceeds to use the crowbar to bash Jerry's face in. Jerry dies of massive head injuries shortly afterward. On thinking about his actions later, Newman completely stands by what he has done. He has no regret whatsoever for his actions.
>
> Following his arrest, Newman undergoes a routine psychological evaluation. At the end of this evaluation, the psychological experts assigned to Newman's case relay their findings to the court. As it happens, Newman turns out to be a fairly mild-mannered man who just happened to hate Jerry. The psychological experts assigned to the case all agree that if Newman were let free, he would *not be at any risk for future offenses.*
>
> Because Newman has no friends (and those involved in his court case are sworn to secrecy), no one outside of court will ever learn that Newman killed Jerry. Consequently, punishing Newman could not deter anyone else from committing similar crimes. Also, no one ever really thought Jerry was funny after his sitcom ended anyway, so no one wants to take revenge on his murderer.[8]

In a case as unusual as this, the weak consequentialist would recommend letting the criminal go free. Since there are no consequentialist benefits whatsoever to be had from

[8] For the record, I do not want Jerry Seinfeld to die.

punishing Newman—and because punishing Newman would involve making him suffer, which is bad—Newman should be let free to enjoy his life as a contributing member of society.

Non-consequentialists will claim that this is ludicrous; if anyone ought to be blamed and punished, it is Newman. Even if punishing Newman would not deter Newman or anyone else from committing crimes in the future, there would be something intrinsically good about making Newman suffer. At least, there wouldn't be anything intrinsically wrong with it—it would certainly be permissible.

The objection here is an intuitive one, and it grows out of a picture of moral philosophy in which the goal of the moral philosopher is to systematize our moral intuitions into a coherent whole that can be captured by relatively few general principles. Roughly, the assumption lying behind this methodology is that most of our moral intuitions are reliable guides to the moral truth, and that the best way to build a justified moral theory is to identify the general principles that can unify and explain as many of these judgments as possible. Consequently, if a proposed principle produces an unintuitive result in a particular case, either the principle can be rejected in favor of one that can capture our intuition about that case, or else we can revise our judgment about the particular case to coincide with the principle. In the case at hand, the objection consists in the claim that the intuition that Newman ought to suffer is more likely to be correct than the weak consequentialist principle that says he ought to be spared.

Many consequentialists have felt the need to reject this methodology, denying its core assumption that our intuitions about particular cases are reliable guides to the moral truth. Others have argued for a less radical position, calling into question only the

reliability of intuitions about unusual cases, as these are the cases that create the most trouble for consequentialism. In what follows, this is the strategy I take. Rather than dismissing all of our moral intuitions about concrete cases out of hand, I argue that we only have good reason to doubt the reliability of our intuitions about specifically unusual cases like Newman's.

Among philosophers who have argued against using intuitions about unusual cases in order to undermine consequentialism, R.M. Hare is one of the most prominent. According to Hare, "[People's] moral intuitions are the product of their moral upbringing, and, however good these may have been, they were designed to prepare them to deal with moral situations which are likely to be encountered; there is no guarantee at all that they will be appropriate to unusual cases" (1981, p. 131). Non-consequentialists have been generally unimpressed with this argument—and with good reason. In particular, one wants to know: why *shouldn't* we trust these intuitions about unusual cases? Even if the psychological processes underlying our moral intuitions were designed for ordinary circumstances, why not think they will remain reliable in unusual cases? If the processes underlying our moral intuitions produce judgments according to some set of rules, why think of these rules as mere heuristics that are likely to lead us astray in unusual cases rather than fundamental moral principles that apply universally? And by the way, what makes a case "unusual"?

Here is a first stab at an answer to these questions: even if the psychological processes underlying people's moral intuitions employ the same reliable criteria in unusual cases as they do in normal cases, it is difficult if not impossible for people to tell whether these criteria are satisfied in unusual cases. In a case like Newman's, the

consequentialist would claim that, even though the psychological processes underlying our judgments of responsibility continue to operate using reliable criteria, these criteria are in fact consequentialist criteria (i.e., whether holding the person responsible would maximize beneficial consequences), and we only think holding people morally responsible is justified when doing so would fail to maximize beneficial consequences because we do not fully *realize* that doing so would fail to maximize beneficial consequences.

The non-consequentialist has an obvious and powerful reply to this proposal: she will simply claim to have no trouble recognizing in the relevant test case that fewer beneficial consequences would ensue from blaming or punishing than from not blaming or punishing. Indeed, she clearly seems to be accurately representing the relevant features of the case when she says to you, "I recognize that punishing this person will have fewer beneficial consequences than alternative courses of action, but I still think we should do it." Needless to say, it will be difficult to make the case that we should distrust her intuition because she is not accurately representing the content of the test case.

My response to the non-consequentialists reply may come as something of a surprise: I agree with it. The non-consequentialist clearly seems to have an accurate representation of the case in mind. But I also think the non-consequentialist has an inaccurate representation of the case in another part of her mind, and it is this inaccurate representation that is responsible for producing her intuition about the case.

Consider that moral judgments are responses to particular states of affairs; we judge a particular state of affairs to be good, bad, right or wrong in virtue of some set of features that state of affairs has or does not have. Consider further that when we issue a

moral judgment in response to a particular state of affairs, we might get things wrong about the state of affairs to which we are responding. We might believe that we are responding to a state of affairs in which the proposition P is true, when in fact P is false.

When we realize that a moral judgment of ours presupposes some false belief about morally relevant considerations or relies on some inaccurate representation of morally relevant facts, we usually correct our moral judgment, assuming we are rational. But not all beliefs and representations are held consciously. Indeed, the psychological processes underlying our judgments make constant use of representations and principles held at the unconscious level, and these unconscious phenomena may be impervious to conscious adjustment (e.g., Nisbett & Wilson, 1977; Kurzban, 2011). If this is right, when we realize our moral judgments may well presuppose some inaccurate representations that are typically held unconsciously, we cannot always consciously correct these representations, and, therefore, *we can no longer trust the output of the psychological processes that rely on them*.

Consider how this abstract schema might be put to work arbitrating the debate about Newman. Even if I consciously believe punishing Newman will have no beneficial consequences whatsoever, I may unconsciously represent the act of punishing him to have a number of beneficial consequences—perhaps more than any alternative course of action. Unbeknownst to me, this unconscious representation of the proposed punishment as having beneficial consequences may play an essential role in driving my conscious belief that Newman ought to be punished. Indeed, if that unconscious representation did in fact play an essential role in generating my conscious belief that Newman ought to be

punished, then, if I were able to correct that representation to reflect the facts of the case, I might no longer believe that Newman ought to be punished.

These claims about unconscious representations may seem like little more than armchair psychologizing. However, there exist both theoretical and empirical reasons to think humans might have evolved to unconsciously represent the expected consequences of blame and punishment (or at least to represent cues to the expected consequences of blame and punishments), that these unconscious representation might play an essential role in making decisions about blame and punishment, and that, when false, they might be impervious to certain forms of conscious adjustment. In the next section, I draw upon emerging psychological research on people's intuitions about punishment to suggest just these things. I argue that this research casts doubt on the claim that non-consequentialists accurately represent unusual cases like Newman's, and that this gives us strong reason to doubt the reliability of intuitions about these cases.

2.2.2 Unconscious Consequentialism

> "The mistake of traditional Schlickian compatibilists was to suppose that individual human beings (and groups of human beings) take up reactive attitudes and engage in retributive practices in order to achieve exclusively utilitarian goals. But that is manifestly false." – David Zimmerman (2008, p. 272)

Recent work in evolutionary psychology suggests that humans may have evolved the capacity to unconsciously represent at least one psychological variable directly related to the expected consequences of punishment. In spelling out what they call the "Recalibrational Theory of Punishment and Reconciliation", Petersen, Sell, Tooby, and Cosmides (2010) argue that, for a number of reasons that need not concern us here, ancestral humans would have had to regularly decide between punishing wrongdoers in some way or reconciling with them. To help solve this recurring problem, they argue that

humans would have evolved a psychological variable they call the wrongdoer's *Association Value index:*

> To solve the problem of choosing between reconciliation, punishment, execution, or ostracism/confinement, our minds evolved to weigh a number of factors against each other. These factors include the relative formidability of the punishers compared to the punished, the likelihood of recidivism and future harm, and the future benefits of continued interactions with the malefactor—including the malefactor's enmeshment with others in deep and productive social relationships. The key decision element is the malefactor's value as an associate: the estimated net lifetime value of maintaining interactions or a relationship with the malefactor from the point of view of the decision maker. We will refer to this as the malefactor's *Association Value* (Tooby & Cosmides, 1996). Hence, we suggest that the human evolved psychological architecture contains subcomponents that are designed to spontaneously compute this index - an *Association Value index,* together with accompanying implicit representations of the degree of uncertainty about the true magnitude of the Association Value. (Petersen et al., 2010, pp. 106-107)

As Petersen et al. summarize elsewhere, "A person's association value is an estimate of how likely that person is to exploit (or confer benefits on) you and those you care about in the future" (in press, p. 28). When the wrongdoer is perceived to have a high association value, you should be motivated to reconcile with him, since maintaining that relationship is likely to maximize your future well-being and the well-being of those you care about. When the wrongdoer is perceived to have a low association value, you should be motivated to instead punish him, since punishment is likely to increase the importance the wrongdoer places on your well-being and the well-being of those you care about. Finally, when the wrongdoer is perceived to have an exceptionally low or negative association value, you should be motivated to execute or ostracize him, since maintaining a relationship with the wrongdoer is likely to only harm your well-being and the well-being of those you care about.

There is some limited evidence supporting the existence of the association value index and its role in regulating judgments about punishment and reconciliation. Petersen, Sell, Tooby, and Cosmides (in press) presented Danish participants with three vignettes in which three different crimes are committed: robbery, vandalism, and rape. Participants then provided their perception of each criminal's association value (operationalized as the criminal's likelihood of recidivism), how serious they considered the crime to be, recommendations for either punishment or rehabilitation (operationalized as an abstract preference for punishing the criminal instead of helping him realize his mistake), and how intense they believed the punishment or rehabilitation should be. Petersen et al. found that, across all three vignettes, participants' decisions to punish or rehabilitate were regulated by the criminal's perceived association value, while decisions about *how severely* to punish or rehabilitate were independently modulated by participants' perceptions of the crime's seriousness.

Petersen et al. also embedded two experiments within their study in order to test predictions about how the association value is calculated. Within the second vignette, half the participants were told that the criminal had committed acts of vandalism three times in the past, while the other half were told that this was the criminal's first offense. This manipulation did have an effect on people's reactions to the crime—people on average preferred punishment for the experienced vandal and rehabilitation for the first-timer—but this effect was completely mediated by differences in the criminal's perceived association value: "The path through perceived association value is so strong that controlling for that variable is, by itself, sufficient to render the relationship between criminal history and reparative preferences insignificant" (in press, p. 19). In other words,

it seems the only reason that participants preferred to punish the experienced vandal is because they believed he was dangerous.[9]

For the third vignette, half the participants were told the rapist was an outgroup member (an immigrant) and the other half were told the participant was an ingroup member (a member of the military). After statistically taking into account participants' varying attitudes towards immigrants, Petersen et al. found that the rapist's perceived status as an ingroup or outgroup member did have an effect on whether participants preferred punishment or rehabilitation, but that this effect "was fully mediated by the way in which these factors influenced perceptions of association value" (in press, p. 23). The fact that group membership, in addition to criminal history, did not *directly* influence punishment decisions—but influenced them only *indirectly* by influencing the perceived association value of the criminal—suggests that what really matters for deciding whether to punish or reconcile with wrongdoers is how much one unconsciously expects to benefit from either strategy.

In a second study, Petersen et al. presented American participants with a vignette in which a man has been arrested for stabbing another man in a fight outside of a bar. Half of the participants were told that the criminal "manages an area auto parts store, expressed deep remorse and apologized for the pain he has caused the victim and his family," while the other half were not told this (in press, p. 25). Like participants in the first study, participants then provided their perception of the criminal's association value (this time operationalized as how likely participants believed "this criminal can someday

[9] The reason that I say it only *seems* that participants preferred to punish the experienced vandal more because they believed he was dangerous is because mediation analyses are necessarily correlational in nature; they can suggest causation, but they cannot definitively establish it in the way that well-designed experiments can.

become a productive member of society"), how serious they considered the crime to be, recommendations for either punishment or rehabilitation (this time operationalized as a recommendation that the criminal either go to prison or enter a job training/college degree program and/or drug and alcohol treatment program), and how intense they believed the punishment or rehabilitation should be (in press, p. 25).

Just as in the first study, Petersen et al. found that differences in the criminal's perceived association value—but not in the perceived seriousness of the crime—influenced participants' choices to punish or rehabilitate. Furthermore, while Petersen et al. found that participants who were told that the criminal was remorseful, had no criminal record, and was a successful employee of a community business preferred rehabilitation over punishment, they also found that this effect was completely mediated by differences in the criminal's perceived association value.

Further evidence for the existence of something like the association value index comes from Burnette, McCullough, Van Tongeren, and Davis (2012). Burnette et al. found in one correlational study that participants' beliefs about how likely wrongdoers were to wrong them again (what they call participants' estimation of 'exploitation risk') significantly predicted participants' willingness to forgive wrongdoers, even when controlling for other predictors of forgiveness. In a second, experimental study, Burnette et al. found that they could increase or decrease participants' willingness to forgive wrongdoers by priming them with questions designed to make them think the wrongdoer was either likely to recidivate or unlikely to recidivate.[10] Importantly, participants'

[10] Importantly, this effect was strongest in conditions in which the wrongdoer was "of high value" to the participant (e.g., friend, family member, etc.). If punishment can ever be intrinsically valuable (i.e., "deserved"), it is implausible to think that whether it is in

conscious estimates of the criminal's exploitation risk did not vary across priming conditions, and yet conscious estimates of the criminal's exploitation risk continued to generally predict forgiveness across conditions. This supports my claim (to be defended below) that people's unconscious estimates of a criminal's likelihood of recidivism can differ from their conscious estimates of the criminal's likelihood of recidivism, even though conscious estimates of this type usually reflect unconscious estimates of the same variable.[11] That being said, these results are also consistent with an alternative explanation, according to which the priming manipulation had its effect by generally activating constructs that were positively or negatively valenced. While there is evidence that trait-related priming effects do not generally work in this way (Erdley & D'Agostino, 1988), only further research can definitively rule this second hypothesis out. In any case, I provide a number of further theoretical and empirical reasons below to think the first hypothesis is more likely to be vindicated.

These studies together suggest that the association value index is a psychologically real representation that plays a distinct causal role in people's decisions

---

fact intrinsically valuable can depend on whether the person being punished is one's friend or family member. Even if being a friend or family member can make a morally relevant difference in some situations, it does not seem like it could make a difference to whether punishment is deserved. At least, it would be a bizarre kind of desert that depended on who was evaluating whether the punishment was deserved. Assuming that we ought to extend our willingness to forgive friends and family to strangers (rather than extending our unwillingness to forgive strangers to friends and family), Burnette et al.'s results lend support to my claim that, if non-consequentialist philosophers were able to accurately represent wrongdoers as nondangerous, they would be less inclined to punish wrongdoers even when doing so would fail to maximize beneficial consequences.

[11] Indeed, as Burnette et al. note: "…although the self-report data here suggest that the outputs of the systems that compute exploitation risk and relationship value are accessible to conscious reflection, this need not be the case. Research methods that assess computations of exploitation risk and relationship value via the activation of particular neural pathways, or via implicit measures, would be valuable as well" (2012, p. 354).

about punishment and rehabilitation. Crucially for my purposes, the association value index represents the likely consequences of continued association with a particular person, and it seems to do this outside of conscious awareness. This provides an explanation for why people continue to punish Newman despite explicitly believing him to be nondangerous: they implicitly believe that he is in fact quite dangerous. In order for my error theory to go through, however, I must not only provide reason to think that these beliefs are realized unconsciously, but that these unconscious beliefs cannot be directly revised via conscious reflection. This is an admittedly difficult task. While the evidence I can provide is far from conclusive, my goal here is only to render the hypothesis plausible enough for us to discount the reliability of intuitions about unusual cases.

2.2.3 Moral Judgment, Modularity, and Encapsulation

According to an increasingly popular view in moral psychology, moral judgments are often the product of unconscious processes, and what looks like moral reasoning more often amounts to mere rationalization of these unconsciously generated intuitions. Spearheaded by Jonathan Haidt's seminal (2001) paper "The Emotional Dog and its Rational Tail", views of this sort do not deny reasoning (or rationalizing) a causal role in moral thought and behavior, but they hold that this role is more often one of a lawyer attempting to craft the most convincing arguments for a preconceived position than of a scientist seeking truth. The goal of reasoning, then, is not to arrive at the most defensible position possible, but to defend and persuade *others* to adopt particular beliefs or perform particular actions.

In an important elaboration of Haidt's view of moral reasoning, Robert Kurzban and Athena Aktipis (2007) postulate the existence of a set of cognitive mechanisms they

call the 'social cognitive interface' or 'SCI', which they liken to a press secretary whose job is to manipulate others' representations of one's traits, abilities, and prospects. According to Kurzban and Aktipis, the SCI "(a) is designed for strategic—especially persuasive—social functions; (b) contains representations that are encapsulated, isolated from many other cognitive systems; and (c) is not necessarily designed to maximize accuracy." (2007, p. 131) Because the SCI has this persuasive function, it is—like a press secretary—strategically ignorant of the reasons why one might have issued a particular judgment or engaged in some other kind of behavior. This strategic ignorance frees up the SCI to coordinate verbal reports and other behaviors that improve one's social standing—*even when such verbal reports and behaviors involve representations that are inconsistent with representations present elsewhere in the brain*.[12]

If something like the SCI exists, it can help explain how it is possible for the non-consequentialist to consciously deny that punishing a particular person will have beneficial consequences, even while continuing to unconsciously represent the opposite state of affairs. It very well may be that the non-consequentialist does believe what she says she does—perhaps because she has quite a bit invested in defending the non-consequentialist position and cannot afford to give that belief up now—but that this belief is instantiated in the SCI and is therefore inferentially isolated from other cognitive processes, leaving her unable to revise the unconscious association value index she holds with respect to the person she believes ought to be punished.

Now imagine that the unconscious belief that punishment would not result in good consequences (or that the wrongdoer remains dangerous) was not in fact

[12] For theoretical reasons to think the SCI would have evolved, see Kurzban and Aktipis (2007) and Kurzban (2011).

inferentially isolated, but that it could in principle interact with psychological mechanisms unrelated to the SCI. There is nevertheless reason to think that this unconscious belief does not reflect our conscious beliefs about the criminal's likelihood of recidivism in unusual cases like Newman's. One reason for this has already been alluded to: modularity. By this I mean only the weak (as opposed to Fodorean) sort of modularity espoused by evolutionary psychologists, the kind of modularity whose core feature is domain specificity and the limited encapsulation that comes along with it. Kurzban and Aktipis (2007) summarize:

> Because evolved specialized mechanisms are designed to process information in particular ways, they process information relevant to the tasks for which the mechanism is designed but not other kinds of information. (Barrett, 2005) That is, functionally specialized systems are necessarily encapsulated with respect to (i.e., do not process) certain kinds of information…Note that by *encapsulation* we do not mean to import the metaphorical entailment that a given representation exists "inside" a system and is, therefore, necessarily unavailable to other systems (Tooby, Cosmides, & Barrett, 2005). We intend the weaker claim that any given mechanisms processes only those inputs that meet the mechanism's formal input conditions (Barrett, 2005). In a similar manner, modularity construed this way should not be taken to entail that any given mechanism necessarily takes only a very narrow range of inputs. A mechanism might function to integrate information, which will necessarily mean that different types of information are taken as inputs. (Kurzban & Aktipis, 2007, p. 132)

Indeed, the sort of encapsulation I am concerned with merely consists in the fact that the inputs to the mental processes that produce our unconscious beliefs about people's dangerousness are limited to ancestrally valid cues to association value, and that, even if conscious beliefs about traits related to people's association values can play *some* part in adjusting unconscious beliefs about the same traits, these conscious beliefs will never be able to unilaterally reverse these unconscious beliefs so long as ancestrally valid evidence

for the opposite conclusions (i.e., non-consequentialist conditions on moral responsibility) remain present in the environment.[13]

2.2.4 Unconscious Representations or Unconscious Beliefs?

> "…a passion must be accompany'd by some false judgment, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgment" – David Hume (1888, p. 416)

There remains a problem with the claim that our unconscious representations of the criminal's association values cannot be altered consciously: if the association value index cannot be (unilaterally) altered via conscious reflection regarding the very same content that it aims to represent, why should we even think of the association value index as a representation of "how likely that person is to exploit (or confer benefits on) you and those you care about in the future"? (Petersen et al., in press, p. 28) If the process that estimates the association value index is sensitive only to *ancestral* cues to association value, why not think it is merely a representation of how likely that person *would be* to exploit (or confer benefits) on you and those you care about *if you happened to live in the time this cognitive mechanisms evolved?* If the association value index represents only the latter, then why think it is inaccurate? And if it is not inaccurate, what reason do we have to think that our intuitions about punishment are unreliable?

So far I have interchangeably referred to the association value index as either an unconscious representation or an unconscious belief. It seems that if I want to justifiably claim that the association value index represents the criminal's association value in a way

[13] If this is right, it is likely a mismatch between the cues to association value that were regularly present in ancestral environments and the cues to association value that are present in the modern environment that accounts for much of the non-consequentialist blame and punishment that we currently see in and out of the court room. I discuss this possibility further in Chapter 4.

that matters for my purposes, I will need to claim that it represents by containing *propositional content* that can either be true or false. One way to make good on this requirement is to show that the association value index plays all of the roles that we typically attribute to beliefs, and that, in light of these considerations, the association value index should be thought of not just an unconscious representation but as an unconscious *belief*.

2.2.5 Stereotypes and Unusual Cases

An intuitive way of understanding my claim about the association value index is on the model of a stereotype. My claim is this: our belief that bad people are dangerous is something like an evolved stereotype, and this stereotype can persist in its influence over our behavior even as we consciously disavow it. The prevalence of this kind of mismatch between conscious beliefs and unconscious stereotypes is one of the central themes of the last three decades of psychological research on stereotypes and implicit attitudes.

The paradigm case from this literature is one in which a person who takes herself to be committed to egalitarian values nevertheless acts in a way that reveals her to unconsciously harbor some racist belief, association, or attitude that she would not consciously endorse. The most common behavioral measure of this sort is the Implicit Association Test, in which participants are asked to correctly group words or pictures into groups like “Pleasant or Black”, “Pleasant or White”, “Unpleasant or Black,” and “Unpleasant or White” (Greenwald, McGhee, & Schwartz, 1998). Typically, participants are shown a computer screen in which two mutually exclusive categories (e.g., “Pleasant or Black” and “Unpleasant or White”) are placed in the upper corners of the screen. Participants are asked to quickly identify words or pictures that appear in the middle of

the screen as belonging to one of the two categories. The dependent measures in these studies are how long it takes participants to correctly group these words and pictures into the presented categories using keys on a keyboard, and how often they make mistakes. The typical pattern of results shows that people typically take longer to correctly categorize a word or picture when the categories are "Pleasant or Black" and "Unpleasant or White" than they do when the categories are "Unpleasant or Black" and "Pleasant or White", and they make fewer mistakes when categorizing words in the "Unpleasant or Black"/"Pleasant or White" condition than they do in the "Pleasant or Black"/"Unpleasant or White" condition.

These results seem to suggest that people either implicitly believe that black people are unpleasant and white people are pleasant or implicitly associate blackness with badness and whiteness with goodness.[14] The question of whether the best explanation for these behaviorial phenomena should be given in terms of beliefs, associations, or some nontraditional mental states like Tamar Gendler's 'aliefs' remains hotly contested (Gendler, 2008a; 2008b). In his (2011) Stanford Encyclopedia article on "Belief", Eric Schwitzgebel notes that "it remains controversial to what extent tests of this sort reveal subjects' (implicit) *beliefs*, as opposed to merely culturally-given associations or attitudes other than full-blown belief."[15] Nevertheless, I favor those who posit the existence of

[14] Or at least, they seem to suggest that people associate blackness with badness more than they associate whiteness with badness, and that they associate goodness with whiteness more than they associate blackness with whiteness.

[15] As evidence for this claim, Schwitzgebel cites a number of articles on the topic: (Wilson, Lindsey, and Schooler, 2000; Kihlstrom, 2004; Lane, Banaji, Nosek, & Greenwald, 2007; Zimmerman, 2007; Gendler, 2008a; Gendler, 2008b; Schwitzgebel 2010).

unconscious beliefs to account for these phenomena. While I do not have the space to fully defend this view here, I will do what I can here to make the view seem plausible.

The arguments in favor of the unconscious-belief model of implicit racism and similar phenomena come almost exclusively from Eric Mandelbaum (in press). Mandelbaum has two arguments: one from the apparently propositional nature of the mental states underlying these phenomena and one from their apparent inferential promiscuity. I will take each of these in turn.

Mandelbaum's first argument revolves around a study originally conducted by Rozin, Millman, and Nemeroff (1986). In the most pertinent version of the experiment, Rozin et al. had participants watch as the experimenter poured "Domino" sugar into two brown 500ml bottles. The experimenter then gave each participant a piece of paper with two labels: "Sucrose (table sugar)" and "Sodium Cyanide". Participants were asked to place one label on each of the brown bottles in whichever arrangement they liked. The experimenter then used two separate spoons to mix sugar from the two bottles into two separate glasses of water, and participants were asked to rate how much they preferred to drink out of each of the cups. On average, participants preferred to drink the water sweetened with the sugar from the bottle labeled "Sucrose (table sugar)" significantly more than they preferred to drink the water sweetened with the sugar from the bottle labeled "Sodium Cyanide", even though they knew that both bottles contained sugar from the same source and that the bottles were arbitrarily labeled. Mandelbaum's argument is simply that, unless the mental state driving people's preferences was a propositional attitude that functioned exactly like a belief functions, it could play no role in explaining their behavior. Here, Mandelbaum is arguing against the claim that the mental states

leading to the behavior are merely associative, which he believes Tamar Gendler's aliefs must be if they are to be importantly different than beliefs:

> The problem for Gendler is that the putative alief looks to be propositional and we need aliefs to be essentially associative in order to underwrite the robust notion. Let us look a bit closer at the Rozin example Gendler uses. Gendler claims that the content of the alief at work is "CYANIDE, DANGEROUS, AVOID" (Gendler 2008a, p. 648). But what is this alief 'telling us' to avoid? To put the question another way, when I token the alief with content CYANIDE, DANGEROUS, AVOID, what am I thinking? If I am just tokening these concepts in succession (which is what Gendler's 'associative state' talk implies), then why would I show any behavior whatsoever toward the *bottle* and not, say, the window, my left foot, or the experimenter's forehead? Since the behavior is bottle specific, the putative alief must somehow bind to the bottle, or else participants would not show the avoidance behavior toward the bottle. Merely saying that the alief's content is associated with the bottle doesn't explain why the alief binds to the bottle alone. (in press, p. 11-12)

This is what Mandelbaum calls "the binding argument": either the mental state driving participants' avoidance of the bottle labeled "Sodium Cyanide" is essentially associative, which means that it cannot explain why participants engage in the specific behavior of avoiding that particular bottle rather than general avoidance behavior, or it has some propositional content, which makes it a belief.[16]

Mandelbaum's second argument is that these mental states seem to be inferentially promiscuous, and only mental states that represent propositionally—namely, beliefs—can be inferentially promiscuous. Here Mandelbaum asks the reader to imagine

[16] One might take issue with the claim that merely having propositional content makes a mental state a belief. This seems right, but these mental states do not only seem to have propositional content—they also seem to presuppose a particular attitude toward these propositions, namely a disposition to evaluate them as true. Even if one nevertheless wants to reserve the term 'belief' for mental states of this kind that are also conscious, one must admit that these mental states are sufficiently belief-like to render them truth-evaluable, and therefore possibly false, rendering any intuitions that are based off of them (when they are false) to be unreliable.

what participants in Rozin et al.'s study would say if they were asked whether *other* people would want to drink from bottle labeled "Sodium Cyanide":

> To see how the putative aliefs can be inferentially promiscuous, imagine that right after you take part in the Rozin study, you are asked a follow-up question about whether other folks would drink from the bottle with the 'cyanide' label. In this case you would probably infer that others would not want to drink from the bottle. (Perhaps you would go through an unconscious chain of reasoning like THAT BOTTLE CONTAINS POISON, PEOPLE DO NOT LIKE DRINKING POISON, SO PEOPLE WILL NOT LIKE DRINKING FROM THAT BOTTLE.) In short, we should expect people to infer from THAT IS DANGEROUS CYANIDE, SO AVOID IT, to other semantically related (and under the circumstances, reasonable-ish) thoughts, such as that others will want to avoid the bottle labeled 'cyanide,' that the bottle would still be labeled 'cyanide' even if the room were a different color, that the bottle will keep its contents even if it is lifted off the ground, and so forth. There are a seemingly unbounded amount of quotidian inferences we would expect the participants to make, but these inferences can only be made from propositional states. Hence there must be belief-like propositional states in play. (in press, p. 14)[17]

In addition to providing this plausible hypothetical extension of Rozin et al. (1986), Mandelbaum goes on to cite just two of what he takes to be many fruitful psychological explanations of *actual* behavior that seem to rely on the existence of unconscious beliefs: cognitive dissonance theory and a particular pattern of behavior observed in insomniacs by Storms and Nisbett (1970).[18]

In cognitive dissonance theory, it is often claimed that people come to express preferences for activities that they previously did not like if (a) they have engaged in that activity, (b) the activity was effortful or costly in some other way, and (c) they lack sufficient external justification (i.e., monetary reward) for engaging in that activity. Think

---

[17] One might claim that the reason these hypothetical participants are able to make these inferences is that they have a conscious belief that they do not want to consume anything that comes out of that bottle, and, since other people are fairly similar, they will not want to consume anything that comes out of that bottle either. This is a fair criticism, which is why I go on to discuss a number of other studies that suggest the existence of unconscious beliefs.

[18] Mandelbaum only discusses Storms and Nisbett (1970) in the section of his dissertation (Mandelbaum, 2010, p. 120) that eventually became Mandelbaum (in press).

of how the hazing process increases commitment to a fraternity that one is initially indifferent towards joining. One might infer from the beliefs, "Only silly people put a lot of effort into things they don't care about" and "I am not a silly person" to "I must care a lot about this fraternity". Of course, since no one *consciously* reasons in this way, these inferences are thought to take place unconsciously (Thibodeau & Aronson, 1992).

In Storms and Nisbett (1970), it was found that insomniacs who took a placebo pill that they were told would increase arousal before bed were able to fall asleep faster than normal, and that insomniacs who took a placebo pill that they were told would decrease arousal before bed took longer to fall asleep than normal. The explanation Storms and Nisbett offer for this reverse placebo effect is that insomniacs in the first group attributed their normal arousal before sleep to the pill rather than their emotionally charged thoughts, leading to fewer tokens of those thoughts and less anxiety at the prospect of falling asleep, while insomniacs in the second group attributed their normal arousal to their emotionally charged thoughts, reasoning that their emotionally charged thoughts must be especially worrisome, since they remained aroused even though the pill should have decreased their arousal.[19] Importantly, these chains of reasoning seemed to

[19] The original study by Storms and Nisbett (1970) has been subject to a fair amount of criticism. For instance, using what they considered to be improved methodologies, both Kellogg and Baron (1975) and Bootzin, Herman, and Nicassio (1976) failed to replicate the reverse placebo effect found by Storms and Nisbett (1970), finding a regular placebo effect in its place. Using a similar manipulation, Singerman, Borkovec, and Baron (1976) also failed to find a reverse placebo effect in a study of speech anxious individuals, finding instead a normal placebo effect. Nevertheless, Brockner and Swap (1983) found that reverse placebo effects are robust—but only among people who pay closer attention to their proprioceptive states. Among those who do closely monitor their internal states, reverse placebo effects are common, but among those who pay closer attention to external stimuli, placebo effects are less common. Similar interaction effects due to individual differences related to proprioceptive awareness occurred between unrestrained eaters and restrained eaters (Heatherton, Polivy, & Herman, 1989) and sexually

be completely unconscious: “When subjects were asked if they had thought about the pills at all before getting to sleep, they almost uniformly insisted that after taking the pills they had completely forgotten about them” (Nisbett & Wilson, 1977, p. 238).

One might easily see how these kinds of results could be used to argue that implicit racism is best explained by unconscious beliefs. It might be more difficult to see how they could be relevant to judgments about the permissibility of holding someone morally responsible even in the absence of beneficial consequences. First, why should we think that a belief about likelihood of recidivism is a part of the criminal stereotype? Even if there are associations between dangerousness and the categories CRIMINAL or WRONGDOER, why think these associations are best understood on the model of belief? That is, what evidence do we have that they are inferentially promiscuous? Secondly, even if these beliefs are part of the criminal stereotype and regulate our decisions about punishment and rehabilitation, why think they are *unconscious* and *uncontrollable*?

2.2.6 The Criminal Stereotype and the Automaticity of Person Perception

What reason is there to believe that we have a stereotype that wrongdoers are likely to recidivate or are dangerous? If this is not obvious from introspection, consider Reed and Reed’s (1973) “Status, Images, and Consequences: Once a Criminal Always a Criminal”. Among all the personality characteristics participants attributed to a hypothetical criminal, “dangerous” was the fourth most attributed with 89.6% of the sample attributing it, following “frustrated” (93.6%), “insecure” (93.1%), and “unhappy” (90.1%) (Reed & Reed, 1973, p. 464). Among a sample of 140 white, African American,

---

functional men and sexually dysfunctional men (Cranston-Cuebas, Barlow, Mitchell, & Athanasiou, 1993). For an overview of these types of studies and related effects, see Bootzin and Bailey (2005).

and Hispanic women from different socioeconomic backgrounds who participated in 18 focus groups and 30 in-depth interviews in and around New York City, Madriz (1997) found that the stereotypical criminal was, among other things described as "cruel", "irrational", "insane", "violent", "out of control", and as being the kind of person who "attacks randomly in the streets" (p. 353). Among a sample of Australians, O'Connor (1984) found that concepts associated with violent criminals include "dangerous, vicious, unintelligent, commits other crimes, immature, and inconsiderate" (p. 260).[20] In the first of two questionnaire studies, MacLin and Herrera (2006) found the ten most frequent responses to the question "What are the first 10 things you think of when you heard the world criminal?" were (from most frequent to least frequent) "jail, murder, police, gun, crime, bad, prison, male, court and drugs" (p. 200). In their second study, MacLin and Herrera found that the most common initial responses to this question were "jail" (at 16%), "theft" (at 11%), and "bad" at (11%) (2006, p. 202). If not all of these concepts directly implicate dangerousness, many do, and certainly none of them preclude dangerousness.

I don't expect that the claim that we have a stereotype that wrongdoers are dangerous or likely to recidivate to be very controversial. Taking this for granted, we now have to ask: why think these stereotypes are (a) unconscious, (b) beliefs, and (c) uncontrollable? Now, it would certainly be misleading to claim that stereotypes about criminals are held *only* at the unconscious level. After all, the stereotypes described in the

[20] I should note that a different kind of criminal—a swindler—was described among participants as "intelligent, smart, well mannered, mature, and inconsiderate" (O'Connor, 1984, p. 260). While none of these concepts directly implicate dangerousness per se, they do not preclude them. In fact, they may exacerbate the extent to which a liar, cheater, or stealer will be viewed as likely to exploit others—another important kind of danger relevant to association value, albeit not a violent one.

studies above were found by asking people to provide self-reports of the kinds of words and concepts they associated with criminals. So these stereotypes can certainly be made explicit under some circumstances. But why think they are ever unconscious at all? And why think they are beliefs?

Consider again Rozin, Millman, and Nemeroff (1986). The reason why Rozin et al.'s experiment provides evidence for the existence of an unconscious belief is that there is a particular behavior—reporting a preference to avoid water that is sweetened with sugar from a bottle labeled "Sodium Cyanide"—that cannot be easily explained by making reference only to conscious beliefs. What kinds of behaviors do humans perform that could only be explained by making reference not only to conscious beliefs but unconscious beliefs about a criminal's likelihood of recidivism? One bad answer would be that humans continue to recommend blame and punishment even when they explicitly believe that criminals are highly unlikely to recidivate. While this belief certainly seems to be true in many cases (Aharoni, 2009), to take this as evidence for the existence of an unconscious belief about criminals' likelihood of recidivism would be to beg the question against the non-consequentialist, for this is precisely the kind of judgment whose reliability is at issue. A better answer would involve a behavior that both the consequentialist and the non-consequentialist can agree could only be explained by positing the existence of an unconscious belief about a criminal's likelihood of recidivism.

One such behavior could in fact be people's reports of their conscious beliefs about a criminal's likelihood of recidivism. While this may sound counterintuitive, consider that all we need to find is a *change* in people's conscious beliefs or behavior that

cannot be explained by any change in any of their other conscious beliefs (or anything else besides an unconscious belief about the criminal's likelihood of recidivism). For this, there is compelling evidence. Consider an early study by Srull and Wyer (1979), in which participants unscrambled sentences that either included words related to a particular character trait (hostility or kindness) or did not, and then evaluated a man named Donald whose behavior was ambiguous with respect to that character trait. Srull and Wyer found that those who were primed with words related to hostility (or kindness) evaluated Donald to be a more hostile (or kind) person than those who were primed with neutral words. Furthermore, Srull and Wyer found that people evaluated ambiguous actors to have certain qualities that were only related to hostility or kindness via implicit personality theories. This provides initial evidence that people have unconscious mental states that contain propositional content about the character traits (e.g., hostility) of other people, and that these mental states are inferentially promiscuous, leading to other beliefs about people's characters that could only be arrived at by inferences via one's implicit personality theories.

Bargh and Pietromonaco (1982) replicated Srull and Wyer (1979)'s results for the priming effects of words related to hostility by using a subliminal priming manipulation rather than a supraliminal scrambled sentence task, further demonstrating that these effects must implicate unconscious mental states. Subsequent studies have since found similar effects by not only subliminally priming words related to individual character traits, but by priming words related to general stereotypes. For instance, in a pioneering study of this sort, Devine (1989) found that subliminally priming participants with words related to the stereotype BLACK would cause them to interpret Donald's behavior to be

more hostile than participants who were subliminally primed with race-neutral words. This lends yet further credence to the idea that representations of people's character traits must at least sometimes exist below consciousness.

These studies taken alone might be unimpressive. Why think these results necessarily implicate unconscious *beliefs*? An alternative explanation might be that the priming task merely makes the construct HOSTILE more accessible, making it more likely that people will use it when they form their *conscious* beliefs about the actor. If this explanation were true, however, it should not matter whether the priming manipulation occurred before or after reading the vignette about Donald—in both cases, the construct is made more accessible before participants are asked to evaluate Donald. And yet Srull and Wyer (1980) found no priming effects whatsoever when participants performed the scrambled sentence task in between reading about Donald's behavior and then evaluating his character. Furthermore, Srull and Wyer found that, in experiments where participants performed the scrambled sentence task before reading about and evaluating Donald, increasing the length of time between the scrambled sentence task and the presentation of the Donald vignette would *decrease* the priming effect, but increasing the length of time between the presentation of the Donald vignette and subsequent evaluations would *increase* the priming effect. As Srull and Wyer conclude, "Evidence from the three experiments reported indicate that priming does not have a direct effect on judgments, and once information has been encoded in a particular way, it is typically not recoded in terms of categories that are highly accessible at the time of judgment." (1980, p. 852)

One would be hard-pressed to explain these results without implicating an intermediary unconscious judgment that contains the information that has been encoded

in a particular way about a particular person. First, it would be difficult to explain why priming only has an effect on participants' judgments of Donald when conducted before encoding—that is, unless participants automatically form unconscious beliefs about Donald's character when they first read about his behavior. Likewise, it would be hard to see why increasing the length of time between encoding and evaluation but not between priming and encoding increases the priming effect—that is, unless encoding requires forming unconscious beliefs that can become stronger over time.[21]

Again, these arguments might be unimpressive. One might accept that the mental states that help explain the effects of sequence and delay mentioned above are beliefs but deny that they are *unconscious* beliefs. Why not think that participants spontaneously form *conscious* beliefs about Donald's character as soon as they read about his behavior? For this to be true, the belief that Donald is hostile would have to spontaneously occur to people upon reading Donald's story, and the related evaluative traits that people often judge Donald to have would also have to spontaneously occur to them, or else they would have to make these inferences quickly and consciously. But this seems unlikely. Introspection suggests that we do not always make conscious judgments about the people we meet or read about as soon as we meet or read about them, especially if their behavior

[21] Another reason to believe that these mental states are beliefs is that priming reliably influences judgments of traits that are only indirectly related to the primed concept, and there is evidence to think that this is not merely an artifact of "spreading activation" that makes these related constructs more accessible. For instance, Todorov and Uleman (2002) found that spontaneous trait inferences were specifically tied to the faces that were presented at the same time as the behavioral description on which the trait inference was based, and that these inferences had to be unconscious, because participants did not remember the behavioral descriptions on which they were originally based. Similar effects have been found in other studies (e.g., Overwalle, Drenth, and Marsman, 1999; Todorov & Uleman, 2004; Crawford, Skowronski, Stiff, & Scherer, 2008; Ferreira, Garcia-Marques, Hamilton, Ramos, Uleman, & Jerónimo, 2011; although see Carlston & Skowronski, 2005).

is ambiguous. Furthermore, psychologists typically consider these inferences to occur too quickly to be consciously mediated (Uleman, Saribay, Gonzalez, 2008; Todorov & Uleman, 2002).

2.2.7 The Futility of Conscious Reflection

The foregoing is far from a knockdown defense of the claim that beliefs can be unconscious. Nevertheless, the hypothesis is not implausible on its face, and there are a number of psychologists and philosophers who do have independent arguments for the claim that beliefs can be unconscious (e.g., Bem, 1970; Gopnik and Meltzoff, 1994; Lycan, 1986; Lycan, 2008; Dretske 1995; Dretske, 2004; Williamson, 2000; Carruthers 2009, 2010).[22] In any case, my error theory for non-consequentialist intuitions requires more than the mere possibility of unconscious beliefs; it requires that these beliefs be impervious to conscious adjustment. Conveniently, further evidence for the claim that the necessary unconscious beliefs exist is also evidence that these unconscious beliefs are impervious to conscious adjustment.

One particularly relevant experiment comes from Gilbert, Tafarodi, and Malone's (1993) "You Can't Not Believe Everything You Read".[23] Gilbert et al. had participants read text from two crime reports as they crawled across a screen (one at a time). Participants were told that some sentences in these crime reports were true and others false—those that were true were presented in black text and those that were false were presented in red text. In one of the crime reports, participants read false statements that, if true, would render the crime more serious; in the other crime report, participants read

[22] This list was compiled by Mandelbaum (2012).
[23] The existence of this experiment and the fact that it seems to support the existence of unconscious beliefs that can come apart from conscious beliefs was brought to my attention by Mandelbaum (2012).

statements that, if true, would render the crime less serious. As the black and red text from the crime report crawled across the screen, a line of blue digits immediately below the text from the crime report also crawled across the screen. Half of the participants were instructed to press a button on a handheld counter whenever they observed the digit "5" appear in the line of digits. The rest of the participants had practice using this handheld counter but were ultimately told they had been randomly assigned to a control condition. After reading the crime reports out loud (and either pushing a button whenever a 5 appeared below the text or else ignoring the line of digits completely), participants were asked to assign a prison sentence to each of the suspects from zero to 20 years and to rate how much they disliked the suspects, how much they thought the suspects could be helped by counseling, and—especially important for my purposes—how dangerous they believed the suspects were.

When the false statements embedded in the crime report were such that, if they were true, they would exacerbate the severity of the crime, participants assigned to the digit-searching condition recommended significantly longer prison sentences, reported that they disliked the suspect more, and judged the suspect to be more dangerous than participants who did not have to complete the digit-searching task. (An effect was found on perceived potential for improvement with counseling, but the difference between the digit-searchers and the non-searchers was not statistically significant.) Likewise, when the false statements embedded in the crime report were such that, if they were true, they would reduce the severity of the crime, participants assigned to the digit-searching condition recommended significantly shorter prison sentences, reported that they disliked the suspect less, and judged the suspect to be less dangerous than participants who did not

have to complete the digit-searching task. (Again, an effect was found on perceived potential for improvement with counseling, but the difference between the digit-searchers and the non-searchers was non-significant.) These results suggest that, although participants explicitly believed the statements to be false when they read them, they nevertheless encoded the statements as true, and then proceeded to unconsciously infer from these unconscious beliefs that the suspect was more or less likeable, more or less dangerous, and deserved more or less time in prison. This is compelling evidence that, despite their conscious beliefs that the exacerbating or extenuating statements they read were in fact false, participants could not help but encode these statements as true, store them as (false) unconscious beliefs, and then infer from them related unconscious beliefs about the criminal's likeability, dangerousness, and deserved punishment.

I should note that at the very end of the experiment, participants were given a set of 60 sentences, eight of which reflected true statements from the two crime reports, 14 of which reflected false statements embedded in the two crime reports, and 38 of which bore no relation to any of the statements in the two crime reports. Participants who completed the digit-searching task did in fact tend to misremember the false statements as true more often than those who did not complete the digit-searching task. One might be tempted to take this as evidence that participants never formed false unconscious beliefs in the first place and that the effects are due instead to their forming false *conscious* beliefs when reading the crime reports. This would be a mistake. First, one would *expect* these unconscious beliefs to be made conscious in the absence of any alternative sources for that conscious belief. Secondly, participants were explicitly told that red statements were false, and it strains credibility to claim that participants simply ignored this part of

the instructions. Still, one might claim, even if this remains good evidence for the existence of unconscious beliefs, it does not suggest that these unconscious beliefs are uncontrollable in any way that matters. After all, it certainly seems like participants would change their prison sentence recommendations, assessments of the suspect's likeability, and judgments about his dangerousness if they were told that the exacerbating or mitigating beliefs they were remembering to be true were in fact false.

This may be true. However, this does not undermine the claim that beliefs about the suspect's dangerousness may be formed unconsciously and that *these* might come apart from people's conscious beliefs about the suspect's dangerousness. Gilbert et al. do not report the exacerbating and mitigating statements they used in their study, but they were presumably statements about the general nature of the crime itself rather than statements directly concerning the criminal's dangerousness. If I am right, these statements likely reference ancestrally valid cues to association value, and I have never denied that conscious beliefs about these ancestrally valid cues can alter one's unconscious estimate of a person's association value. All I have claimed is that, regardless of one's conscious beliefs about a person's association value index (dangerousness, likelihood of recidivism, etc.), if you are presented with ancestrally valid cues to various character traits associated with a person's association value, then you cannot help but unconsciously believe that that person has those traits.

Or can you? This brings us back to the literature on stereotypes. Over roughly the last two decades, there has been an explosion in research on the extent to which people can and cannot control the stereotypes they hold about other people. This research revolves around two themes: control over the activation of a stereotype and control over

the influence of an already-activated stereotype on subsequent behavior. The more mature of these two lines of research is the latter, which cannot help us here. Because the behavior relevant here is whether it is appropriate to make someone suffer—and because the question just *is* whether this stereotype is influencing our intuitions that certain people ought to suffer—the only way of knowing for sure that the stereotype that wrongdoers are dangerous is not influencing our intuitions is to ensure that the stereotype is never activated in the first place, which is the focus of the first area of research. Unfortunately for non-consequentialists, the research that does exist in this area suggests that the prospects for control over stereotype activation are fairly bleak.

In one of the field's seminal studies, Devine (1989) found that even those who consciously espoused egalitarian values were susceptible to being influenced by a subliminal prime involving words related to the category BLACK. Using the Donald vignette from Srull and Wyer (1979), Devine found that those who were subliminally exposed to words related to the category BLACK rated Donald to be more hostile than those who were subliminally exposed to race-neutral words, regardless of whether they espoused conscious prejudice towards blacks or not. Furthermore, Graham and Lowery (2004) found that the same subliminal priming procedure could be used to influence police officers' and juvenile probation officers' beliefs about an adolescent criminal's character, culpability for a crime, likelihood of recidivism, and deserved punishment—all the while having no effect on participants' conscious beliefs about and attitudes toward African Americans. It seems to follow that simply espousing views that run against a stereotype does not prevent that very stereotype from being unconsciously activated and influencing one's judgments and behavior.

However, this conclusion needs to be qualified. Lepore and Brown (1987) ran an experiment that was nearly identical to Devine (1989), but after correcting for some methodological errors, they found that those who were low in prejudice were in fact less likely to have the negative stereotype activated after being exposed to subliminal racial primes. In fact, exposure to the subliminal primes activated *positive* stereotypes. This suggests that even if people cannot avoid activating *some* stereotype, they *do* have the potential to change these stereotypes. Indeed, this hypothesis has been borne out by subsequent studies (Blair & Banaji, 1996; Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000). For instance, Kawakami et al. had participants participate in *counter-stereotype association training*, in which pictures of skinheads were repeatedly paired with words representing stereotypical negative traits or words representing non-stereotypical positive traits, and where participants were instructed to press a "NO" button when the pictures were paired with the negative trait words and a "YES" button when paired with the positive trait words. Participants who underwent this training revealed on two different measures of implicit stereotyping that the training had inhibited activation of the usual stereotype.

It is an interesting question whether procedures like these could be used to curb the stereotyping of wrongdoers in unusual cases like Newman's. However, regardless of whether participating in these training programs could successfully prevent the relevant stereotype from being activated (and for reasons tied up with claims about modularity, I doubt they can), philosophers who object to consequentialism about moral responsibility certainly have not undergone anything like this before judging unusual cases like

Newman's.[24] Therefore, we have no reason to reinvest trust in their intuitions about cases like Newman's. Furthermore, the methods that these philosophers typically *do* employ in order to properly judge unusual cases like Newman's (i.e., consciously denying that the stereotype applies in the case) have been shown to be reliably *counterproductive* when it comes to reducing stereotype activation.[25] For instance, Macrae, Boedenhausen, Milne, and Jetten (1994) found a "rebound" effect in which those who attempted to consciously suppress stereotypes in one task ended up making those stereotypes even more accessible in a subsequent task. That is, participants who made an active effort *not* to think of stereotypes ended up employing those implicit stereotypes even more extensively in judging and interacting with other people than they would have had they never tried to suppress those stereotypes in the first place.[26] Wyer and Budesheim (1987) found that, even when participants were told that a person did not actually perform certain behaviors

[24] In another set of studies, it has been found that stereotypes can generally be inhibited when incompatible goals (e.g., striving to be egalitarian) are themselves implicitly activated (for a review, see Moskowitz, 2010). In any case, it seems unlikely that non-consequentialist philosophers have an implicit goal that can be tapped in this way to reduce stereotype activation in unusual cases like Newman's. Likewise, Galinsky and Moskowitz (2000) found that asking participants to take the perspective of a stereotyped group significantly reduced the activation and employment of stereotypes on both conscious and unconscious tasks. However, as I argue later, initial angry responses to wrongdoers likely function to *prevent* perspective-taking, thus undermining the viability of perspective-taking as an ad hoc strategy for correcting non-consequentialists' representations of Newman-like wrongdoers.

[25] An exception to this claim might be counterstereotypic mental imagery. Blair, Ma, and Lenton (2001) found that participants who engaged in counterstereotypic mental imagery (i.e., imagining a strong woman) revealed significantly diminished implicit gender stereotypes on a number of different measures. While this kind of strategy *might* succeed in getting non-consequentialists to accurately represent the Newman case, I am not optimistic; Blair et al. admit that it may also depend on participants' motivations to diminish the behavior that was associated with the implicit stereotype. Presumably, non-consequentialists do not have the motivation to blame Newman less. Furthermore, the stereotype that women are weak is presumably less resistant to revision than the stereotype that wrongdoers are dangerous.

[26] For further demonstration of the rebound effect, see Galinsky and Moskowitz (2006).

that they were originally led to believe the person had performed, their having originally believed that the person had performed those behaviors still influenced their subsequent impressions of that person's character traits.[27] Johnson and Seifert (1994) found across several experiments that information influenced participants' judgments even when participants were immediately told to disregard that information. They found this effect was increased when the to-be-disregarded information originally provided a causal explanation for some observed phenomenon—as character trait information does in person perception—but that it was decreased when an alternative causal explanation was given to replace the to-be-disregarded explanation. Ecker, Lewandowsky, and Tang (2010) found that going so far as to explicitly warn participants that they might be susceptible to using the to-be-disregarded information in generating inferences did in fact reduce the influence of the to-be-disregarded information, but it did not eliminate it.[28]

These studies suggest that the conditions under which and the extent to which we can revise certain unconscious representations might be quite limited. This is extremely

---

[27] Wyer and Budesheim (1987) found this effect primarily for behaviors that implied favorable character traits. When the to-be-disregarded behaviors implied unfavorable character traits, participants actually reported believing that the target person had even more favorable character traits than those who read vignettes that simply omitted the information that participants in the other condition were told to disregard. Wyer and Budehseim suggest that this reflects an attempt to consciously adjust for the effect participants perceived having initially read the to-be-disregarded information had on their initial impressions. This line of reasoning suggests that the to-be-disregarded information nevertheless had an influence on people's intuitive judgments, which they subsequently revised only if they believed that the to-be-disregarded information influenced their judgments. Because non-consequentialists would deny that their evaluations of people like Newman were influenced by their perception of the person as dangerous *per se*, they would not be motivated to revise their downstream judgments. Consequently, Wyer and Budesheim's results are consistent with the claim that representations of Newman as dangerous might nevertheless persist unconsciously.

[28] For an early and theoretically minded summary of similar effects—which the authors term effects of 'mental contamination'—see Wilson and Brekke (1994).

problematic for philosophers who believe that armchair reflection on unusual cases like Newman's is sufficient to render their intuitions about these cases reliable. Even if such armchair reflection can alter conscious beliefs about the criminal's dangerousness, it may do so at the cost of making the implicit stereotype even more accessible—and it is this implicit stereotype that I argue plays the crucial role in regulating our moral intuitions about these cases.

This concludes my error theory for non-consequentialist intuitions. While the error theory presented here is only tentative, my only goal in laying it out was to render the claim plausible enough to cast doubt on the reliability of intuitions about unusual cases like Newman's.[29] Having done this, I will go on to argue that this is sufficient to undermine non-consequentialist objections to consequentialism, clearing the way for a positive argument for consequentialism.

2.3.1 Why Care About Wrongdoers' Well-Being?

If the error theory presented thus far is successful, then it has provided an undercutting defeater for the claim that we ought to make wrongdoers suffer even when doing so would not maximize beneficial consequences. That is, it has undermined the claim that it is *right* to make wrongdoers suffer even when doing so fails to maximize beneficial consequences. What it has *not* done—at least not directly—is provide a reason to think that it is *wrong* to make wrongdoers suffer even when doing so fails to maximize

[29] At this point, the retributivist might claim that her position never depended on any intuitions about Newman-type cases in the first place. Rather, she might claim that retributivism follows from some set of abstract principles about fairness or desert. It's certainly open to the retributivist to make this move, but I sincerely doubt that any principles she might employ enjoy much intuitive plausibility over and above the fact that they seem to unify and explain our intuitions about Newman-type cases. In fact, once we ignore Newman-type cases, I doubt these principles enjoy any more direct intuitive plausibility than competing consequentialist principles.

beneficial consequences. Therefore, even if we have no reason to think that we *should* blame or punish wrongdoers regardless of the consequences, we have no reason to think we *should not*. In this section, I argue that the same intuition that is the target of the error theory laid out in the previous section is also the one psychological obstacle to otherwise compelling arguments that the suffering of wrongdoers ought to be minimized. Consequently, once we realize that we cannot trust this intuition, the arguments that it used to block suddenly become available for the consequentialist to use against the non-consequentialist.

Here is the most straightforward argument a consequentialist might use to argue that we should minimize the suffering of wrongdoers when making them suffer has no beneficial consequences:

1) Suffering ought to be prevented, except for when it leads to the prevention of even greater suffering or the promotion of beneficial consequences of equal or greater importance.

2) There is no intuitive and trustworthy counterexample to the principle expressed in premise (1).

3) Therefore, the suffering of wrongdoers ought to be prevented, except for when it leads to the prevention of even greater suffering or the promotion of beneficial consequences of equal or greater importance.

My claims are that, in all cases except for those like Newman's, our intuitions support the principle expressed in premise (1), that our intuitions about cases like Newman's are not reliable, and that coherence demands that we extend the principle expressed in premise (1) to cases like Newman's.

This argument follows a familiar form of moral reasoning. Nevertheless, philosophers of a particularist persuasion will likely be unmoved by it. Particularists hold that the non-moral factors that are morally relevant in one case may not always be

morally relevant in the same way (or at all) in other cases. Consequently, particularists would believe that, unless we come to have a properly informed intuition that their suffering is morally problematic, we have every reason to remain indifferent to the suffering of wrongdoers. Although I take coherence to be an essential end of proper moral reasoning, in what follows I take up the particularists' objection and argue against it on its own terms. My claim will be that an appropriately sensitive and adequately informed moral judge will have the intuition that the wrongdoer ought not to suffer when doing so fails to maximize beneficial consequences.

2.3.2 Stalemate and Arguments for Incompatibilism

Incompatibilists about moral responsibility claim that if certain conditions were met (i.e., all events were a function of past events and the laws of nature), then wrongdoers could not be justifiably held morally responsible for their actions. In terms of the present debate, incompatibilists believe that in a world in which determinism is true—or in which what is sometimes vaguely referred to as a 'libertarian free will' is absent—it is wrong to make even wrongdoers suffer if doing so would fail to maximize beneficial consequences. Therefore, conditional on the truth of determinism (or the right kind of indeterminism), incompatibilists deny the legitimacy of non-consequentialist justifications for blame and punishment. (Some incompatibilists also deny the legitimacy of consequentialist justifications for blame and punishment; I briefly address these arguments in the section "The Fairness Problem".)

Incompatibilists have offered a number of arguments for their thesis. Notable examples include Peter van Inwagen (1986)'s Consequence Argument (and its descendants), arguments based on moral luck, and arguments based on the purported

incoherence of a sufficiently robust notion of free will.[30] Currently, the most popular type of argument for incompatibilism is what has come to be called the 'manipulation argument'. Manipulation arguments work by presenting cases in which an agent's actions are deterministically traceable to the activities of other conscious agents, claiming that these manipulated agents are not morally responsible and concluding that, because these cases are morally indistinguishable from cases of pure determinism, agents in normal deterministic worlds must also not be morally responsible. I focus on this type of argument below, but this is primarily because I think it does most effectively what all other incompatibilist arguments attempt to do: to allow people to empathize with and therefore feel compassion for wrongdoers.[31]

---

[30] Technically, people who endorse this last kind of argument are not incompatibilists *per se*, but 'impossibilists'. For an overview of these arguments, see Vihvelin (2011).

[31] It might seem odd for me to claim that incompatibilist arguments are merely tools to get people to help empathize with determined agents, for it seems like this empathy could be elicited even without supposing the agent to be determined. But if it were really empathy that was doing the work here, and empathy could be elicited even for agents that were not causally determined, we would expect libertarians who espouse this argument not to be libertarians but hard incompatibilists—people who deny that people can be morally responsible either because determinism is true, or because only an unhelpful kind of indeterminism is true. So why aren't libertarians hard incompatibilists? First, I am not claiming that libertarians *intend* for their arguments to function in this way, just that this how they do when they are successful. If I had to guess, I would think that the premises libertarians use in their arguments for incompatibilism are largely motivated by theological considerations tied up with the problem of evil rather than considerations of moral luck or empathy. (Indeed, in the recent *PhilPapers* Survey of professional philosophers, the correlation between philosophers who endorsed libertarianism and philosophers who endorsed theism was among the highest at 0.396. This and other correlations can be found at http://philpapers.org/surveys/linear_most.pl.) Secondly, I doubt that libertarians do find it easy to empathize with agents they believe to be undetermined. Consider that libertarians often claim that a full explanation of any act for which an agent is responsible will necessarily involve a notoriously inscrutable sort of agent-causation or indeterminist choice that even many of its most prominent defenders admit to be "puzzling" (van Inwagen, 1986, p. 150), "strange", or "mysterious" (Taylor, 1992, p. 52; van Inwagen, 2000). To speculate wildly, it might be that endorsing these

Consider the following thought experiment:

> Diana [a goddess with special powers] creates a zygote [Newman] in Mary. She combines [Newman's] atoms as she does because she wants [Newman to kill Jerry] thirty years later. From her knowledge of the state of the universe just prior to her creating [Newman] and the laws of nature of her deterministic universe, she deduces that a zygote with precisely [Newman]'s constitution located in Mary will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to [kill Jerry] and will [kill Jerry] on the basis of that judgment, thereby bringing about [Jerry's death]. (adapted from Mele, 2006, p. 188)

Now imagine the same case, except this time Newman's genetic constitution and environmental situation is not traceable to Diana, but to blind chance. Incompatibilists and compatibilists—those who believe that determinism is compatible with people being morally responsible—have more or less arrived at a consensus that, when these two cases are adequately described, there is no morally relevant difference between them (McKenna, 2008; Pereboom, 2008; Fischer, 2011; although see Kearns, 2011). Therefore, if Newman is morally responsible in the second case, he must be in first as well; if Newman is not morally responsible in first case, he must not be in the second, either.

As might then be predicted, compatibilists and incompatibilists have both ended up taking manipulation arguments and running them in opposite directions, with compatibilists starting with the case in which the wrongdoer is intuitively morally responsible and arguing that he must also be responsible in the other case, and incompatibilists starting with the case in which the wrongdoer is intuitively not morally responsible and arguing he must also not be responsible in the other case (Pereboom, 2008; McKenna, 2008). With this the perpetual stalemate in the debate about free will and moral responsibility returns with a vengeance, and manipulation arguments seem to

---

obscure metaphysical claims helps libertarians prevent themselves from acting on their temptation to empathize with wrongdoers.

be of no help to either the incompatibilist or the consequentialist about moral responsibility.

I believe the error theory for non-consequentialist intuitions laid out earlier in this paper can help settle this debate in favor of incompatibilism (and consequentialism). The relevant claim is just this: those who believe even causally determined agents can be morally responsible in a way that makes it permissible to make them suffer even when doing so would fail to maximize beneficial consequences may unconsciously believe that nondangerous agents are in fact dangerous (among other things) and so misconstrue the case in such a way that their intuitions would only *seem* to speak against consequentialism. In other words, cases like Newman's only work to undermine consequentialism if certain conditions (such as Newman's not being dangerous) are reflected in the processes generating our moral intuitions. If the processes that produce our moral intuitions do not accurately represent these conditions, we have no reason to trust these intuitions as reliable indicators of the moral truth in these cases. These intuitions are simply responses to different kinds of cases—cases in which the consequentialist can agree that blaming and punishing the wrongdoers might be permissible or even required (e.g., because the wrongdoer is dangerous).

If the compatibilist already agrees that there is no morally relevant difference between manipulation cases and cases of pure determinism, then once we provide a reason to doubt the compatibilist-friendly intuition in the pure determinism case, we get the incompatibilist-friendly intuition about the manipulation case to transfer to the pure determinism case for free. So if we believe something like determinism is true—as most opponents of consequentialist theories of moral responsibility do—then we are

committed to only making wrongdoers suffer when doing so would maximize beneficial consequences.

There is a problem here, and that is my argument presupposes the validity of arguing from one case to another, a claim that I have already noted a particularist might deny. Rather than ignoring the particularist, I can easily reformulate my argument in particularist terms. The reformulation is simply that, rather than claiming the incompatibilist intuition strictly *transfers* from the manipulation case to the pure determinism case, I should claim that a moral judge who was capable of accurately representing a case like Newman's would in fact come to believe from direct moral perception that the wrongdoer ought not to be made to suffer when doing so would fail to maximize beneficial consequences. This is because removing the false belief about the criminal's likelihood of recidivism removes an obstacle to empathizing with the wrongdoer, and then when one empathizes with the wrongdoer, one will work to remove their suffering, unless that suffering is itself necessary to relieve even more suffering or to bring about other beneficial consequences in the future.[32]

Although these claims are largely speculative, there is some limited empirical evidence in support of them. For instance, McCullough, Worthington, Jr., and Rachal (1997) found naturalistic and experimental evidence that ability to feel empathy for a wrongdoer strongly predicts whether someone will feel forgiveness for that wrongdoer. McCullough, Rachal, Sandage, Worthington, Jr., Brown, and Hight (1998) found further evidence for empathy's role in forgiveness. Not only did empathy strongly predict

[32] Furthermore, Todd (2012) convincingly argues that the purpose of manipulation arguments is to help people recognize the nature of determinism in a way that they could not before considering deterministic *manipulation*.

forgiveness in McCullough et al.'s study, but it fully mediated the effect of other variables (e.g., closeness to wrongdoer, apologies from the wrongdoer) on forgiveness. (p. 1598)

In a neuroimaging study, Singer, Seymour, O'Doherty, Stephan, Dolan, and Dirth (2006) found activation in empathy-related brain areas of participants who observed painful shocks being delivered to the hands of confederates who had played fairly with them in an economic game beforehand, but they found significantly less activation in these areas in men who observed painful shocks being delivered to confederates who had played *unfairly* with them in the earlier economic game. Furthermore, these men who observed unfair players receive painful shocks also had increased activation in brain areas related to reward processing, and this increased activation was correlated with a verbally reported desire for revenge. This suggests that one of the reasons that men (at least) might be unwilling to refrain from making wrongdoers suffer is that they fail to empathize with wrongdoers, desiring instead that they suffer. If I am right, they have that desire to make wrongdoers suffer primarily because they (unconsciously) believe they are dangerous. Therefore, if we could revise this unconscious belief in the person's dangerousness, we would expect decreased vengefulness and thereby increased empathy.

The challenge in vindicating these claims is that, per the error theory laid out earlier in the paper, revising unconscious beliefs about wrongdoers' dangerousness is very difficult. This follows from the fact that perceiving a person doing wrong activates from the bottom-up a criminal stereotype that includes as one of its core features a belief in the propensity of criminals to do wrong again in the future. Nevertheless, it may be possible to have some direct but necessarily incomplete top-down control over the

activation of this stereotype.[33] Operating on this assumption, several studies suggest that the suppression of this stereotype and the recognition of something like the truth of determinism—which may increase people's ability to empathize by making it easier for them to take another person's perspective—may in fact decrease people's desire to punish wrongdoers.

The first attempt to test the general hypothesis that determinism will reduce non-consequentialist blame and punishment was conducted by Nahmias, Coates, and Kvaran (2007). Nahmias et al. asked participants whether causally determined criminals in general could deserve to suffer solely "because they deserve it for what they have done" and whether the causally determined criminals in that particular study's vignettes deserve to suffer for what they have done (Nahmias et al., 2007, p. 239). With the exception of questions asking whether humans have nonmaterial souls, these questions elicited more "I don't know" responses than any other question in the survey: approximately 20% to 30% of responses (Nahmias et al., 2007, p. 240). This presumably reflects a greatly decreased certainty on the part of participants concerning the permissibility of non-consequentialist punishment in deterministic worlds.

The decreased ability for participants in Nahmias et al. (2007) to decipher their motivations for punishment in deterministic worlds reflects people's general inability to

[33] This may be especially true when the wrongness of the agent's acts is in fact ambiguous, as it was in Bargh, Chen, and Burrows (1996): "In this study, the construct-relevant behaviors were unambiguous and clearly diagnostic of the trait in question, and accessibility of the trait concept did not influence impressions of the target (though it did affect ability to process the information). The top-down effects of accessibility should influence impressions only when the informational input is sufficiently ambiguous (i.e., a relatively weak bottom-up effect; see Higgins, 1989)" (Bargh, Chen, & Burrows, 1996, p. 235).

accurately identify their own motives for their actions, even in universes they take as actual (e.g., Nisbett & Wilson, 1977). Furthermore, Carlsmith (2008) has shown that this lack of self-knowledge extends specifically to people's knowledge of their own motives for punishment. It is for this reason that Lerner (2011) followed previous studies on motives for punishment in taking a "policy capturing" approach, a method in which the experimenter systematically varies features of cases that should or should not have an effect on the dependent variable depending on whether participants have motives of one kind or another (Darley, Carlsmith, & Robinson, 2000, p. 661). In Lerner (2011), this required varying a feature of the case relevant to forward-looking, consequentialist motives for punishment (the criminal's likelihood of recidivism), varying a feature of the case that may be relevant to backward-looking, non-consequentialist motives for punishment depending on whether the folk are natural compatibilists or incompatibilists (i.e., whether causal determinism is true in the universe in which the crime is committed), and then observing how these variations affected recommended punishment severity (measured abstractly and in terms of recommended length of confinement in a prison or a mental institution) as well as abstract attributions of moral responsibility, blameworthiness, and free will.

Lerner (2011) was designed to test the hypothesis that participants would feel more empathy and therefore recommend less severe punishment for wrongdoers when the wrongdoers were (a) not dangerous, and (b) causally determined. Indeed, Lerner found that participants recommended sentences of equal length to criminals who were not causally determined regardless of whether they were highly likely to recidivate or highly unlikely to recidivate, but that they recommended significantly shorter sentences to

causally determined criminals when they were unlikely to recidivate compared to when they were highly likely to recidivate. This suggests that both a disbelief in the wrongdoer's dangerousness *and* a belief in the truth of causal determinism are necessary in order to reduce non-consequentialist punishment.

The results of Lerner (2011) suggest that people might become consequentialists about blame and punishment if they were to come to believe that determinism were true. Nevertheless, the majority of participants in Lerner (2011) who read about a criminal who was both causally determined and nondangerous nevertheless recommended that he receive *some* substantial amount of punishment. Even though participants were told that it was known that the criminal they were judging would never commit another crime again if freed, and that the outcome of the trial would not be made public (and so no deterrent effect would be foregone by assigning a less severe punishment), the average participant still recommended approximately 7 years of detainment, with approximately half of that spent in prison. There are two possible explanations for this result. The first explanation is that even though participants might have believed the causally determined, nondangerous criminal deserved *less* than the nondangerous, undetermined criminal, they may nevertheless have judged him *somewhat* deserving. This would mean that participants are not *complete* consequentialists when it comes to punishing causally determined criminals.[34] The second explanation is that participants did not believe the conditions described in the vignette (low likelihood of recidivism, non-publicity of the trial) could actually be met, or else they did not see how meeting these conditions

[34] For an argument that something like this is in fact the response we should take to determinism, see Capes (in press). For an argument that someone who holds this position cannot ultimately resist the stronger conclusion that determinism should undermine all non-consequentialist punishment, see Todd (2011).

precluded all possible consequentialist justifications for punishment. Indeed, many participants reported in response to manipulation checks that they believed the criminal was still somewhat likely to commit similar crimes in the future and that the results of the trial would eventually be made public. If something like this second explanation is right, then non-consequentialists about blame and punishment have no reliable intuition to offer against consequentialism about blame and punishment.

A second study also suggests that determinism reduces people's desires for non-consequentialist punishment. Shariff, Greene, and Schooler (2011) found that people recommended less retributive punishment for criminals when they were first primed with articles about neuroscientific research into human decisionmaking. Although neuroscientific research does not demonstrate the truth of determinism, it does provide insight into the mechanisms underlying people's decisions and behaviors, which, like determinism, presumably increases people's ability to take the perspective of wrongdoers and thereby empathize with them.[35]

The last study that suggests determinism might undermine people's support for non-consequentialist punishment comes from Petersen (2010). Petersen found that people's tendency to feel compassion rather than anger for criminals and their tendency to agree with the statements (a) "Deep down, many criminals are regular people like you and me" and (b) "Criminals are victims of a hard upbringing" rather than the statements (c) "Most criminals commit crimes because they know they can get away with it" and (d) "Most criminals are psychopaths who do not care about others at all" interacted to predict people's support for harsh punishment over rehabilitation (p. 360). While agreeing with

[35] For evidence that perspective taking increases empathy, see Lamm, Batson, and Decety (2007).

(a) and (b) is not exactly the same as recognizing the influence of moral luck on people's actions and accepting (c) and (d) is not exactly the same as denying the influence of moral luck one people's actions, the fact that agreeing with the former two items strongly predicts preferences for rehabilitation over punishment lends some support to the idea that recognizing moral luck might help people feel compassion for wrongdoers and ultimately reject retributive punishment.[36]

In this section, I have provided reason to think that an ideal moral judge would refrain from making wrongdoers suffer when doing so would fail to maximize beneficial consequences. This constitutes an argument for what I have called weak consequentialism. However, this argument depends crucially on the assumptions that (a) the error theory I laid out earlier for non-consequentialist intuitions is right, and (b) there is no equally good or better error theory for weak consequentialist intuitions. In the next section, I briefly present and dismiss two possible error theories for weak consequentialist intuitions.

2.3.3 Two Error Theories for Weak Consequentialist Intuitions

In the last few years, Eddy Nahmias and Dylan Murray as well as Chandra Sripada have independently offered error theories for the intuition that Newman-type

[36] It might be thought that classic experimental philosophy studies like Nichols and Knobe (2007) are relevant here. While Nichols and Knobe did find that people would deny that causally determined agents could be held morally responsible, this effect was only found when the relevant question was framed in the abstract. (In cases where participants were asked to consider a particular causally determined agent, they reported believing that he *could* be held morally responsible.) Furthermore, the dependent measure used by Nichols and Knobe (and many following them) was simply whether participants thought causally determined agents could be "morally responsible". As recognized by Nahmias et al. (2007), questions of this sort are too ambiguous to provide any evidence about people's motivations for punishment. Lastly, Nahmias and Murray (2010) provide strong evidence that even the effect Nichols and Knobe did find was likely due to participants misunderstanding the nature of determinism.

wrongdoers should not be made to suffer.[37] Both error theories share the claim that the only reason people report incompatibilist intuitions is because they misunderstand critical features of the relevant thought experiments. For Nahmias and Murray (2010), the claim is that people only have incompatibilist intuitions when considering cases of causally determined agents because they misunderstand determinism to imply *bypassing*. According to Nahmias and Murray, bypassing involves either *fatalism*—the claim that a particular event in one's future is inevitable no matter what one does, or that one's future would be the same even if one's past had been different—or *epiphenomenalism*—the claim that one's mental states are causally inert—and the truth of either of these claims undermines moral responsibility because they entail that the types of mental states compatibilists tie to moral responsibility do not play their proper causal role in the production of one's actions. For Sripada (2011), the claim is that people only have incompatibilist intuitions in response to manipulation cases like the one considered earlier because they falsely believe either (a) that the manipulated agent is in possession of false, distorted, incomplete, or otherwise "corrupted" information, or else (b) that the attitudes that the manipulated agent acts on are somehow unreflective of his deeper values, those values that reflect his real self. Both Nahmias and Murray as well as Sripada offer impressive survey results that seem to strongly support their theories.

[37] In point of fact, both Nahmias and Sripada offer their error theories as error theories for incompatibilist intuitions about free will and moral responsibility generally. But insofar as weak consequentialist intuitions just are a kind of incompatibilist intuition, and insofar as the weak consequentialist is only concerned with the kind of moral responsibility at stake in the Newman case, I explore here the potential for Nahmias and Sripada's error theories to undermine the intuition that wrongdoers like Newman should not be made to suffer on account of their immoral behavior.

Although Nahmias, Murray, and Sripada explicitly offer their error theories as debunking explanations for non-philosophers' ostensibly incompatibilist intuitions, there are ways these theories might be extended to account for *philosopher's* incompatibilist intuitions. The strategy is this: 1. Show that non-philosophers only have incompatibilist intuitions because they misinterpret determinism to entail bypassing or manipulation to involve corrupted information or deep self-discordance. 2. Boldly declare that philosophers were once non-philosophers. 3. Argue that, despite their conscious disavowal of the belief that determinism entails bypassing, philosophers nevertheless maintain an unconscious belief that determinism does entail bypassing, and either (a) this unconscious belief continues to sustain their consequentialist intuition, or (b) their consequentialist intuition is nevertheless traceable to this misunderstanding from their pre-theoretic days. 4. Conclude that because incompatibilist philosophers would have these incompatibilist intuitions even if incompatibilism were false, they are not justified in believing that their intuition is justified. (5. Humbly assume that no such story can be told about compatibilist intuitions.)

Assuming Nahmias, Murray, and Sripada's studies show what they say they have, this argument turns on the plausibility of the third step of the argument. Why should we think that the intuitions of incompatibilist philosophers—professionals who make no mistakes about what determinism does and does not involve—are somehow infected by a mistake they have long since corrected? There are at least three ways this might work. First, one might think that budding incompatibilist philosophers are only attracted to incompatibilism because they commit the interpretive errors postulated by Nahmias, Murray, and Sripada, and that mulling over the compatibility question many times early

in their career causes their initial intuition that the causally determined or manipulated agent is not responsible to become in some way psychologically tied (perhaps through some sort of classical conditioning mechanism) to considerations of determinism. Consequently, even when these interpretive errors are eventually corrected, reading descriptions of determinism or manipulated might still provoke the incompatibilist intuition.

Alternatively, one might argue that that incompatibilists' current intuitions are the product of motivated reasoning. Again, incompatibilist philosophers, so the story might go, may have been initially attracted to incompatibilism because they committed one of these interpretive errors. Nevertheless, on the occasion in which they first had an intuition about a deterministic world or a manipulated agent, they may have misdiagnosed the source of their prima facie incompatibilist intuitions, citing the determined agent's lack of alternative possibilities or ultimate sourcehood rather than bypassed mental states, corrupted information, or deep self-discordance. Having to constantly defend their explicit explanations against philosophical opponents, cognitive dissonance mechanisms may have recruited incompatibilists' initial intuitions in order to help sustain their explicit declarations of non-responsibility, even when those explicit declarations of non-responsibility are accompanied by explicit disavowals of the relevant interpretive errors.[38]

Consider one final way in which these interpretive errors might infect professional incompatibilists' intuitions. Perhaps the psychological mechanisms

[38] In fact, humans probably have no general motive to eliminate cognitive dissonance. However, humans likely have a motive to *appear* consistent. For an argument to this effect, see Kurzban and Aktipis (2007, p. 140)

underlying incompatibilists' intuitions simply cannot properly discriminate between determinism and bypassing, or manipulation and corrupted information/deep self-discordance. It seems doubtful that, evolutionarily speaking, there ever would have been selection pressures to be able to tell between two agents who were the same in all relevant respects except that one had libertarian free will and the other did not. (In fact, even if there were such pressures, there would be no phenotypal cues a mechanism could use for purposes of discrimination.) However, there certainly would have been pressures for humans to develop a way to identify and excuse or exempt agents who were not acting normally from their true values with uncorrupted information, either because they were coerced, developmentally immature, or sick; blame and punishment risk retaliation, and blaming and punishing the wrong person carries especially large risks. Consequently, it may be that even though incompatibilists have no trouble consciously understanding determinism and manipulation, the psychological mechanisms underlying their intuitions continue to represent determined and manipulated agents to be coerced, developmentally immature, or sick.

The first two suggestions considered above are not implausible on their face, but I do not believe they will ultimately be vindicated. The first hypothesis requires an egregious and habitual error on the part of professional philosophers, and the second hypothesis could well be adapted to provide an equally plausible explanation of compatibilist's intuitions (if it is plausible at all). The last suggestion, however, closely resembles the kind of error theory I provided above for non-consequentialist intuitions. If it worked above, why wouldn't it work here? Unlike in the error theory for non-consequentialist intuitions laid out earlier, the "unusual" feature of the case that

compatibilists claim incompatibilists are failing to understand—the fact that a causally determined agent has a perfectly normal relationship to his behavior—is not unusual in any significant sense. In order for causally determined but psychologically normal agents to be unusual in the relevant sense, it would have to have been the case that whenever we encountered agents in the past that we believed to be causally determined, we also believed they were psychologically abnormal. But there is good evidence that people generally believe that humans are *not* causally determined (Sarkissian, Chatterjee, De Brigard, Knobe, Nichols, & Sirker, 2007). It follows that people have probably never had regular interactions with agents that they believed were both causally determined *and* psychologically abnormal, the kinds of interactions that would be required to build the stereotype that all causally determined agents are psychologically abnormal. In fact, the first time most people ever consider such an agent is in the context of a philosophical thought experiment. But if the first time people encounter such an agent is in philosophical contexts where the psychological normality of the agent is made explicit (perhaps after a few initial confusions), the stereotype will never have the chance to develop. Unless there is some reason for thinking that humans come "pre-equipped" with the stereotype that causally determined agents are psychologically abnormal, there seems to be no plausible way we could have obtained the kind of stereotype needed for the present error theory to go through. But there seems to be no reason to think humans have this "innate" stereotype. Therefore, it seems completely reasonable to assume that incompatibilists are representing the case correctly when they say they are.

There are further, methodological reasons for doubting that the results Nahmias, Murray and Sripada obtained undermine *anyone's* incompatibilist intuitions. First, the

manipulation cases Sripada used in his study left open the possibility that the agent was not causally determined. Likewise, Nahmias and Murray did not check to make sure that participants who did not mistakenly understand determinism to entail bypassing or fatalism nevertheless understood determinism to imply what determinism *does* imply. These gaps in the designs of the two studies leave open the possibility that participants believed the manipulated agent was morally responsible because he was not in fact causally determined. Secondly, both of the studies simply asked whether participants believed the wrongdoers were "morally responsible". As was argued in the first section of this thesis, the answer to this kind of question is too ambiguous to bear on the kinds of issues that the consequentialist takes to be at stake in the free will debate.

I have argued in this section that the error theories on offer for weak consequentialist (incompatibilist) intuitions are significantly weaker than the error theory laid out above for non-consequentialist intuitions. Consequently, weak consequentialism should be preferred to non-consequentialism. The next task for the consequentialist is to show that strong consequentialism should be preferred to weak consequentialism. In the next chapter of this thesis, I extend my argument for weak consequentialism to strong consequentialism, which encompasses weak consequentialism but also claims that we should be willing to impose costs on the innocent just as frequently as wrongdoers when doing so would maximize beneficial consequences.

Chapter 3: Why Strong Consequentialism?

3.1 Why Weak Consequentialism Collapses into Strong Consequentialism

So far I have argued only that we should be weak consequentialists. Weak consequentialism is weak because it simply imposes a constraint on when we can

permissibly hold people morally responsible. This leaves open the possibility that there are *further* conditions that must be met in order to permissibly hold someone morally responsible. For example, weak consequentialists may believe that, in order to permissibly hold someone morally responsible in a way that causes that person to experience a given amount of suffering, that person must have committed some sufficiently bad act or harbored some sufficiently bad quality of will. This is a demand not to punish the innocent, and not to punish the guilty more than is proportional to their crime. This demand constitutes what is sometimes called 'negative retributivism'. In this brief section, I argue that the combination of weak consequentialism with negative retributivism is unstable; the same reasons that should make us weak consequentialists should also make us strong consequentialists.

In an earlier section, I argued that we should seek to prevent wrongdoers from suffering when their suffering would not maximize beneficial consequences. My argument for this claim was straightforward: suffering is usually bad, we have no good reason to think that the suffering of wrongdoers isn't bad, and so we should try to prevent wrongdoers from suffering except when allowing them to suffer prevents even more suffering in the long run. To make this into an argument for strong consequentialism, we need only make one small change to this argument: the second premise should not just read "we have no good reason to think that the suffering of wrongdoers isn't bad" but "we have no good reason to think the suffering of wrongdoers *is any less bad than the suffering of anyone else*." If this newly revised premise is right, then we have lost all reason for prioritizing the well-being of the innocent over that of the guilty.

Here might be an appropriate point to revisit the reasons offered in an earlier section for thinking that the suffering of wrongdoers is bad and ought to be prevented. The argument offered earlier was that certain psychological mechanisms prevent us from empathizing with wrongdoers in cases in which punishing them would typically maximize beneficial consequences, but that these mechanisms continue to prevent us from empathizing with wrongdoers even in unusual cases in which punishing them would *not* maximize beneficial consequences because they misrepresent crucial parts of the case (e.g., the criminal's unusually low likelihood of recidivism). Because these psychological mechanisms can prevent us from empathizing only by relying on these misrepresentations, we have no reason to think that preventing us from empathizing is appropriate in these unusual cases. Since we have no reason for thinking that empathy is inappropriate in these unusual cases, we should do what we can to make ourselves empathize with the wrongdoers so that we can gain full information about the case. Perhaps by reflecting on how constitutive and circumstantial luck conspired to make the wrongdoers perform their wrongful acts, we should come to recognize the badness of their suffering and consequently become motivated to eliminate it.

The key for transforming this argument for weak consequentialism into an argument for strong consequentialism is recognizing that constitutive and circumstantial luck are ultimately responsible for *all* human behavior. My claim is not that constitutive and circumstantial luck *per se* put the suffering of the guilty on par with that of the innocent, but that constitutive and circumstantial luck should help us empathize with the guilty *just as much as we do with the innocent*, and that once we do this we will see that the suffering of the guilty is just as bad as the suffering of the innocent. But—and here is

the crucial claim—if we think the suffering of the innocent is no worse than the suffering of the guilty, and if we think that we can nevertheless make the guilty suffer when doing so would maximize beneficial consequences, *we should also think that we can make the innocent suffer when doing so would maximize beneficial consequences.*

Compatibilists will likely balk at this conclusion. They will claim that even if they can fully empathize with the guilty, and even if doing so shows that their suffering is bad, doing so does not provide any reason to think that suffering of the guilty is *just as bad* as the suffering of the innocent. They may claim that constitutive and circumstantial luck should only make us recognize that there is *some* disvalue in making the guilty suffer, not that there is *just as much* disvalue in making the guilty suffer as there is in making the innocent suffer.

Properly assessing this argument requires returning to the question of constitutive luck touched on above. The philosophical literature most germane to this question has its origin in Gary Watson's (1987) discussion of Robert Harris. One day in 1982, Robert Harris brutally murdered two teenage boys and stole their car in order to rob a bank. The details of the case reveal just how evil Harris really was. For instance, Harris is reported to have laughed hysterically after shooting one of the boys in the head at point-blank range with a Luger. Soon thereafter, Harris proceeded to finish off the hamburgers that the boys had been eating in their car when they were hijacked.

Harris's case becomes philosophically interesting when we start to look at his history. As might be expected, Harris's early life was marked by brutal physical, emotional, and sexual abuse both at home and in juvenile detention centers. Robert's

sister describes just one of a number of recurring instance of neglect from his early childhood:

> He'd come up to my mother and just try to rub his little hands on her leg or her arm. He just never got touched at all. She'd just push him away or kick him. One time she bloodied his nose when he was trying to get close to her. (Corwin, 1982, as cited in Watson, 1987, p. 273)

The challenge here is to explain why we intuitively feel less inclined to make Harris suffer, if we do, when we learn about his history. Consequentialists and others who wish to challenge the legitimacy of non-consequentialist justifications for holding people morally responsible will claim that Harris ought to be spared because he was merely a victim of unbelievably bad luck. Non-consequentialist compatibilists, however, will want to claim that whatever inclines us to empathize with Harris and exempt him from blame is not something so general as his bad luck *per se*. Rather, they will claim it has something to do with the intelligibility of his response to his horrible upbringing:

> Thus, someone who had a supportive and loving environment as a child, but who was devoted to dominating others, who killed for enjoyment, would not be vicious in just the way Harris is, since he or she could not be seen as striking back at "society"; but such a person could be just *as* vicious. In common parlance, we sometimes call such people "bad apples", a phrase that marks a blank in our understanding. In contrast to Harris, whose malice is motivated, the conduct of "bad apples" seems inexplicable. So far, we cannot see them as victims, and there is no application for thoughts about sympathy and moral luck. (Watson, 1987, p. 277)

But why shouldn't considerations of moral luck mitigate our punitive responses to these more banal types of wrongdoers? If one's genetic makeup and environment are together sufficient to produce all of one's behavior, why shouldn't we think that anyone who commits wrong is merely the victim of bad luck? And if all wrongdoers are merely victims of bad luck in genes and environment, why shouldn't we seek to prevent their suffering as much as we seek to prevent the suffering of people like ourselves who

benefit from good luck in genes and environment? Watson claims that the kind of perspective taking that underlie these empathetic responses verges on incoherence:

> …the counterfactuals that underlie thoughts about moral luck must be constrained by the conditions of personal identity. It may be that no one who had been exposed to just the internal and external conditions of some given individual could have been me. To make sense of a counterfactual of the form, "If *I* had been in *C*, then *I* would have become a person of type *t*," *C* must be supposed to be compatible with *I*'s existence as an individual (*I* must exist in the possible world in which *C* obtains). For example, it is widely held that genetic origin is essential to an individual's identity. In that case, the counterfactual, "If I had had Harris's genetic origin and his upbringing, then I would have been as evil as he," will not make sense. Now it might be that Harris's genetic origins are among the determinants of his moral development. Thus, even if this is a deterministic world, there may be no true counterfactual that would support the thought that the difference between Harris and me is a matter of moral luck. There is room for the thought that there is something "in me" by virtue of which I would not have become a vicious person in Harris's circumstances. And if that factor were among my essential properties, so to speak, then that difference between Harris and me would not be a matter of moral luck on my part, but a matter of who we essentially were. That would not, of course, mean that I was essentially good or Harris essentially evil, but that I would not have been corrupted by the same circumstances as those that defeated Harris. To be sure, to suppose that this difference is in itself to my moral credit would be odd. To congratulate me on these grounds would be to congratulate me on being myself. Nevertheless, this difference still might explain why it is to my credit, such moral virtues as I may possess. This will seem paradoxical only if we suppose that whatever is a ground of my moral credit must itself be to my credit. But I see no compelling reasons to suppose this. (Watson, 1987, p. 134)

Watson's response here is troubling on several counts. First, some people might think it is intuitively plausible to think that whatever is a ground of one's moral credit—or at least whatever is an essential component of the ground of one's moral credit—must itself be to that person's credit. Secondly, if it is odd to congratulate Watson on being himself, and there is something about Watson that makes his having certain moral virtues essential to his personal identity, why isn't crediting him with these moral virtues equally odd? Lastly and most importantly, one might doubt that our ability to empathize with wrongdoers

really presupposes the thought that *we literally could have been them in some metaphysically possible world.* Andrew Latus is excellent on this point:

> Is it really the case that in order for me to be able to say "You're lucky to be clever," it must have been possible for you to exist without being clever? Surely not. There is nothing odd about saying both that we think being clever is an essential element of your character and that you are lucky to be clever. The way we think about chance is not in terms of the chance of you being constituted differently but in terms of the chance of a person being constituted that way (that is so as to be clever). You're lucky because you're constituted in this *unusual* way. (2003, pp. 471-472, emphasis added)

But even if we do think that luck with respect to a particular trait or circumstance is not merely a matter of how common that trait or circumstance is among the human population, we need not agree with Watson that it is metaphysically impossible for us to have been evil like Harris. As Michael McKenna writes:

> …considerations of personal identity might show that it is not metaphysically possible, in many cases, for one to share the determining causes of another's evil behavior. But that does not dispel the worry that for any individual, there might be *some* determining causes *consistent* with her identity which also would have led her to equally evil ways. Hence, Watson's worries about moral luck and determinism remain unanswered. (1998, p. 136)

McKenna himself goes on to dismiss these skeptical worries by claiming it nevertheless remains impossible for us to *now* share *just any* of the possible determining causes of another's evil behavior, even if we *could have* shared those determining causes in some possible world:

> [H]ad many of us had quite different formative years we might be quite different people—possibly morally contemptuous ones. But this thought does not seem to me to have the effect upon our reactive attitudes which Watson suggests it does. The thought that we ought not blame does not seem compelling in these cases since it is hard to take seriously that we might have been like that—for *we are now not anything like that*. Furthermore, for many of us, perhaps most of us, it is *essential* to our present selves that we *could not* be like that. The intuition that we ourselves ought not cast blame seems only to gain a purchase upon our sentiments when the condition of our present selves shares the same kind of moral fault or (minimally) shares the *potential* for the same kind of fault. (1998, p. 140)

We might deny McKenna's intuition here; the very fact that we *could have* had some combination of genes and environment consistent with our identities that would have led us to evil seems sufficient to justify our empathetic responses to wrongdoers.

But even supposing it is *false* that we could have had some combination of genes and environment consistent with our identities that would have led us to evil, this does not seem to affect my argument for consequentialism about moral responsibility. In fact, the thought that I could have been an evildoer plays no crucial *justificatory* role in my argument for consequentialism at all. Rather, the thought that I could have shared the causes of another's bad behavior plays a merely *rhetorical* role in helping me come to *empathize* with that person. And surely the empathy at stake here—understanding another's suffering and the kinds of factors that led to it—does not logically require that I could have actually experienced that person's suffering or been subject to the factors that put that person in the position to suffer. Rather, it simply requires an ability to recognize that *that person there is suffering*. If I am right, it is only because we are less likely to recognize that wrongdoers suffer in being blamed or punished that we are tempted to think that their suffering is less problematic than the suffering of the innocent. While it may be that we do in some sense recognize that wrongdoers suffer when we blame and punish them—indeed, it is often our goal to make them suffer—it is likely that when we represent those people as dangerous, that representation of their suffering gets co-opted by the process that regulates our decisions about how much to blame and punish them. When we accurately represent those people as nondangerous, however, that representation of their suffering may get recruited by the process underlying compassion instead of the process underlying decisions about how much to punish.

I have argued that recognizing that a wrongdoer is nondangerous and that determinism is true should lead us to refrain from making that wrongdoer suffer unless doing so would maximize beneficial consequences. A crucial part of this argument is that constitutive and circumstantial luck together make the suffering of wrongdoers intrinsically bad, or rather prevent it from being morally neutral or good. But if the guilty are subject to constitutive and circumstantial luck, then surely the innocent are as well. However, if this is the case, both parties are ultimately just as "innocent" as the other. Consequently, if reflecting on the consequentialist benefits that blaming and punishing can engender is sufficient to motivate us to punish the guilty despite our tendency to empathize with them, it must be sufficient to motivate us to punish the innocent despite our tendency to empathize with them. I have argued that if we cultivate the sensitivities of the ideal moral judge as described in previous sections, then we will empathize just as much with the misery of the guilty as we will with that of the innocent, and if consequentialist benefits can nevertheless motivate the ideal moral judge to blame and punish the guilty, then there is no reason to think they cannot motivate the ideal moral judge to blame and punish the innocent. Contra weak consequentialism, there is no additional psychological or moral obstacle to surmount before we can blame and punish the innocent in the name of consequentialist benefits.

There is at least one other debunking explanation available for our aversion to punishing the innocent. Just as it is difficult to accurately represent unusual features of cases like Newman's, it may be difficult to accurately represent the unusual features of cases in which punishing the innocent would maximize beneficial consequences. It may be that when we think of hurting innocent people, we automatically activate a

representation of the bad consequences that would ensue—that person's suffering, the suffering that might be brought upon us by their family and friends if we were to punish them, the immediate costs we would endure in carrying out the punishment—in a way that thinking of hurting the guilty does not. Although I do not have the same kind of psychological evidence to offer in support of this error theory for negative retributivist intuitions as I did in support of my error theory for positive retributivist intuitions, I believe the large body of evidence I marshaled in favor of the latter provides us with reason to be optimistic in assessing the prospects of the former.

In this section, I have argued that, because the most plausible reason to reject positive retributivism is also a reason to reject negative retributivism, the most plausible form of weak consequentialism collapses into strong consequentialism. Having now dismissed all theoretical objections to consequentialism about holding people morally responsible, I need now only to address practical objections to the view. This is the goal of the final chapter.

Chapter 4: Practical Objections and Implications

4.1 Why This Chapter is in Some Sense Irrelevant

Before discussing the practical implications of accepting strong consequentialism about moral responsibility, I would like to pause to emphasize that nothing of what I have said so far depends on the success of what's to come. If my arguments thus far have been successful, they have merely established strong consequentialism about moral responsibility at the *foundational* level, and the success of those arguments does not depend in the least bit on the viability of strong consequentialism about moral responsibility at the *normative* level, the level of everyday moral decisionmaking. My

goal thus far has been to identify the kinds of values we ought to ultimately promote in our practices of holding one another morally responsible; it has not been to identify the kinds of decision procedures we ought to employ when deciding whether to hold one another morally responsible. Although these questions are closely related and we can use questions about what decision procedures we should use at the normative level (i.e., in various thought experiments) to help answer questions about what values we ought to promote at the foundational level, an answer to one of these sets of questions does not necessarily imply any answer to the other set of questions.

That being said, I do believe that there is room for reform in our practices of holding one another morally responsible, and there may be a number of situations where we should specifically move in the direction of direct cost-benefit analyses. While it is surely true that consequentialist values are not always best promoted by strictly consequentialist decision procedures, I will explore in this section the extent to which consequentialism at the foundational level might be most effectively promoted by consequentialism at the normative level.

The most natural way to assess consequentialism's viability at the level of practical decisionmaking is to consider the objections that opponents of consequentialism have offered over the years. These objections to consequentialism about moral responsibility all attempt to show that consequentialism is somehow self-defeating. That is, they attempt to show that promoting consequentialism at the normative level would somehow fare worse in promoting consequentialist values than promoting certain non-consequentialist normative theories. These sorts of self-defeat worries come in two general flavors: the psychological and the game-theoretical. Because the psychological

objections are both weightier than and in some sense prior to the game-theoretical worries, I will treat them first.

4.2.1 Psychological Objections

What I am calling 'psychological' objections to consequentialism all claim that it is psychologically impossible, or at least impractical, for people to simultaneously endorse consequentialism at the foundational level and nevertheless maintain the kinds of beliefs and motivations that a practicable consequentialist theory of moral responsibility would need them to. The first objection of this type is what Saul Smilansky calls the *present danger of the future retrospective excuse* (2008, p. 245). The worry here is that the same considerations about moral luck that motivate us to endorse consequentialism in the abstract will nevertheless undermine our personal willingness to accept responsibility for our actions in everyday life. Because we know we will always be able to attribute our bad behavior to moral luck, we will never feel guilt, and because we know we will never feel guilt, it will be difficult for us to anticipate guilt and so act as to avoid it.

The present danger of the future retrospective excuse suffers from an unrealistic picture of moral agency according to which moral behavior is always motivated by a desire to avoid guilt, and that merely reflecting upon moral luck will immediately free us from guilt. While it may be true that guilt avoidance is often an essential component of moral motivation, it is surely not the only one. Presumably, moral behavior is often motivated not (just) by guilt-avoidance, but by a direct desire to help others and oneself (where that self-interest involves more than just avoiding guilt).[39] In any case, it is highly

[39] For an argument that morality does not require the maintenance of guilt—and perhaps requires the rejection of guilt—see Harman (2009).

unlikely that we have the sort of direct control over guilt that Smilansky thinks we do. Even if we *were* to think guilt to be irrational or the suffering that follows from guilt to be automatically unjustifiable, it is unlikely that simply reminding ourselves of this would diminish our guilt. Indeed, most people do not have to look far to find an example from their own lives where they attempted but failed to relieve themselves of guilt by telling themselves the guilt was irrational. For instance, consider the last time you had to make a choice between two terrible options. Did the fact that you chose the lesser of the two evils diminish your guilt? To speculate from the armchair, I would guess not.

Consider now a closely related worry that Smilansky calls *the danger of retrospective dissociation* (2008, p. 246). According to Smilansky, reflections on the pervasiveness of moral luck will make it impossible for us to adequately engage in remorse, to feel *compunction* in response to the evils we have committed. My response to this worry is much like my response to the *present danger of future retrospective excuse*: it seems to presuppose an unrealistic picture of human psychology whereby merely thinking an emotion to be irrational is alone sufficient to dispel that emotion. In any case, it's not even clear that a consequentialist has to believe guilt, shame, remorse, or any other moral emotion *is* irrational. Even if consequentialists believe we should rid ourselves of these emotions when they cause needless suffering, this does not mean they believe these emotions are always *irrational* when they occur. A consequentialist can believe that a particular wave of guilt simply causes needless suffering and should therefore be eliminated while nevertheless considering it, just like any ordinary compatibilist, to be a perfectly comprehensible response to having done wrong. As Shaun Nichols observes, guilt simply does not seem to presuppose much more than the fact that

one has harmed someone she cares about (2007, p. 241).[40] And even if the wide psychological profile of guilt does in some sense presuppose that the experience of guilt maximizes beneficial consequences, it will often be the case that experiencing guilt *does* maximize beneficial consequences, and so the consequentialist will be motivated to see the guilt through its natural course.

Another psychological objection due to Saul Smilansky claims that this motivation—to experience guilt because of its beneficial consequences—is unavailable to the consequentialist:

> …the seriousness of moral appraisal depends on our not viewing judgements merely as manipulative ways of influencing people, which can in principle be applied to the blameless if it is socially useful to do so. People would not be willing to be blamed, would not accept blame as appropriate, were it not assumed that they deserve blame on account of their freely taken actions. (2008, p. 241)

P.F. Strawson voices a similar sort of worry:

> What *is* wrong is to forget that these practices, and their reception, the reactions to them, really *are* expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. Our practices do not merely exploit our natures, they express them. Indeed, the very understanding of the kind of efficacy these expressions of our attitudes have turns on our remembering this. (Strawson, 1982, p. 80)[41]

These objections simply presuppose that people will not find consequentialism compelling. Surely, if people did not find consequentialism compelling, they would be unwilling to follow its recommendations. In this case, consequentialists would have consequentialist reasons to promote an alternative normative theory of moral

---

[40] As Nichols notes, others that share his analysis of guilt include Baumeister, Stillwell, and Heatherton (1994), Haidt (2003), and Prinz and Nichols (2010). Tooby and Cosmides (2008, p. 134) offer a similar analysis grounded in evolutionary considerations.

[41] For a sophisticated elaboration of this claim—and an argument for the claim that the reactive attitudes provide the foundation for a happy, "restorative" middle ground between equally crude consequentialist and retributive approaches to criminal justice—see McGeer (2011).

responsibility. However, assuming people did find consequentialism compelling and were more or less morally motivated, wrongdoers would not object to being blamed any more than they do now. Furthermore, consider that wrongdoers do not typically accept blame and demand their own punishment—which they often do—because they simply want to receive what they deserve. Rather, they often want to signal their recommitment to the people they harmed and to the values of the moral community at large.[42] This quotidian motivation is undeniably consistent with consequentialism.

All of the psychological objections canvassed so far share the basic claim that the theoretical considerations underlying consequentialism cannot motivate the sorts of practices we typically associate with holding one another morally responsible. These objections only have force against the consequentialist so long as these sorts of everyday practices do in fact tend to maximize beneficial consequences compared to available alternatives. A different sort of objection to consequentialism begins with the thought that our everyday practices of holding each other morally responsible are far from optimific and should in fact be discouraged. However, if the responses I've offered to the above objections are sound, they would seem to show that we are "stuck" with our non-optimific practices whether we endorse consequentialism or not. This isn't a worry that consequentialism is self-defeating so much as it's a worry that it can play no role in guiding our actions. In other words, if the worries expressed earlier revolve around the thought that consequentialism motivationally defeats the very types of practices it needs

[42] For an overview of the research concerning the psychological worries that lead Saul Smilansky to reject hard determinism and consequentialism in favor of what he calls 'illusionism', see Nadelhoffer and Matveeva (2007). Nadelhoffer and Matveeva convincingly argue that Smilansky's worries are unsupported by the available evidence.

to engender, a different set of worries revolve around the thought that consequentialism cannot have any influence on our behavior whatsoever.

What I'll call the *normative impotence* objection claims that a theory like consequentialism can have no inhibitory effect on our retributive practices of blame and punishment. When combined with what I'll call the *non-optimific* objection, the normative impotence objection suggests that attempts to move our practices in a more consequentialist direction are hopeless. Fortunately, there is reason to think that neither the normative impotence objection nor the non-optimific objection are generally true. There may be *particular* domains in which endorsing consequentialism cannot lead us to revise our practices in a consequentialist direction, and there may be *particular* domains in which our moral responsibility practices are not optimific, but it does not seem that the normative impotence objection and the non-optimific objection apply in *all* contexts. In fact, I will argue below that, by and large, the domains in which the non-optimific objection does apply are precisely the domains in which the normative impotence objection does not apply, and the domains in which the normative impotence objection *might* apply are precisely the domains in which we would expect the non-optimific objection not to apply. In other words, it seems to be a convenient feature of our existing responsibility practices that we have less control over the practices that are optimific and more control over those that are not.

4.2.2 When and Where Are the Reactive Attitudes Optimific?

Consider two different spheres in which the reactive attitudes and associated practices operate: within everyday interpersonal relationships and outside of everyday interpersonal relationships. In what follows, I argue that we have reason to believe the

reactive attitudes and associated retributive practices will be more or less optimific in the context of everyday interpersonal relationships, but that, once we leave this sphere, these responses to wrongdoing run the risk of failing to maximize beneficial consequences. I also argue that we have more control over our responses to wrongdoing outside the context of everyday interpersonal relationships than we do within them. In other words, my claim is that the modern environment just so happens to be constituted in such a way that the normative impotence objection applies only in the domain in which the non-optimific objection does not, and the non-optimific objection applies only in domains where the normative impotence objection does not.

Why think that the reactive attitudes and associated practices are generally optimific within interpersonal relationships? Motivations for this claim have their origin in Aaron Sell's "Recalibrational Theory of Anger", the emotional analogue and theoretical ancestor of the Recalibrational Theory of Punishment and Reconciliation discussed at length earlier in this paper (Sell, 2011). According the Recalibrational Theory of Anger, anger is an adaptation designed to motivate behaviors that increase the weight others place on your welfare and the welfare of those you care about. (Henceforth, whenever I speak of the individual's welfare, I am speaking both of the individual's welfare and the welfare of those he cares about.) Anger is triggered, then, when people exploit you—that is, when they impose an unjustifiably high cost on you in order to obtain an unjustifiably small benefit for themselves. Importantly, what counts as an "unjustifiably high cost" to you relative to the benefit received by the wrongdoer will depend on your relative bargaining power compared to the wrongdoer. Relative bargaining power depends on several factors, including your relationship to the

wrongdoer and the ability the wrongdoer has to harm or benefit you. Ancestral cues to this second factor include how physically strong the wrongdoer is, how many friends he has (and how strong they are), his social status, any special skills he may have, and his access to resources. Consequently, whether one experiences anger and its concomitant retributive motivations will depend on one's perception of these various cues and one's (unconscious) assessment of one's relative bargaining power. This design feature plays the crucial role of allowing one to not only initiate conflicts when one is likely to benefit from doing so, but to avoid conflicts when one is likely to lose.

Notice the consequentialist aim of anger on the recalibrational account. If the Recalibrational Theory of Anger is true, then anger was designed to be optimific (in the sense of motivating only behaviors likely to solicit the maximum amount of benefits in terms of welfare for oneself and those one cares about) in everyday interpersonal relationships, contexts in which human ancestors spent most if not all of their time (Sell 2011b, p. 382; Sell, 2005, pp. 30-31; Sell, 2011, p. 60; Petersen et al., 2010, pp. 107-109).[43] Furthermore, even though anger is only designed to maximize the welfare of

[43] This, of course, does not mean that anger is the best possible solution to the adaptive problems it evolved to solve; in order to evolve, anger needed only to be "good enough" compared to other strategies that were evolving. Furthermore, even a small advantage in a given trait can confer differential reproductive success sufficient for the genes that code for that trait to proliferate and become dominant in future generations. Nevertheless, given an environment in which other people are already predisposed toward anger and guilt, and where adopting an alternative disposition to employ more or equally effective but less costly strategies (i.e., calmly explaining to the other person why their acts are wrong) may be psychologically impossible or extremely costly itself, anger may very well have been an optimific response to wrongdoing for ancestral humans. It follows from this that anger may also be optimific for modern humans in contexts that are relevantly similar to ancestral environments. Of course, this would be false if there were a psychologically available alternative that could bring the benefits of anger without its costs. Nichols (2007) considers one such alternative strategy—Derk Pereboom's sadness, sorrow, and concern combined with resolve to bring about moral improvement—and

oneself and those one cares about, it was typically the case in ancestral environments that the only interests at stake in these contexts were the interests of oneself and those one cares about. Consequently, insofar as anger would have maximized beneficial consequences for oneself and one's close associates, it would have ipso facto maximized beneficial consequences from an impartial moral perspective as well.

Nevertheless, it would be a mistake to infer from these considerations that anger necessarily remains optimific (in this impartial sense) in the modern environment. All the Recalibrational Theory of Anger predicts is that natural selection would have shaped the anger system to motivate behavior that *would have been optimific in humans' ancestral past*. It follows that we would only expect anger to remain optimific today to the extent that certain features of the modern environment resemble the relevant features of the ancestral environment. If anger and associated practices would have been adaptive in our ancestral environment, and the modern environment differs in one of these relevant ways from our ancestral environment, then we might expect to see the occasional "misfire" in which our anger and associated practices fail to maximize beneficial consequences in the modern environment.

Does the modern environment differ in any way that would prevent that from continuing to be true? There are at least four features of the modern environment that threaten the optimific nature of anger: (1) that one-shot interactions are much more frequent than they were in the past, (2) that we presumably care about (and are capable of

---

dismisses it as inevitably ineffective. Pereboom (2009) argues that Nichols is wrong to think that sadness, sorrow, and concern cannot be just as effective as anger and punishments. Pereboom's most compelling argument is that "[i]t is very difficult to calibrate angry responses, such as anger-motivated punishment, optimally" (2009, p. 176). In what follows, I argue that there are some cases in which Pereboom is right, but these are primarily cases outside the context of normal interpersonal relationships.

affecting) the wellbeing of more people than our ancestors did (i.e., all members of the moral community), (3) that cues to blame and punishment's likely consequences (e.g., relative bargaining power, strength of deterrence, and likelihood of recidivism) have in some cases changed dramatically from ancestral environments to the present, and (4) evolutionarily novel alternatives to punishment now exist that did not in ancestral environments. Of course, these four "mismatches" between the ancestral environment and the modern environment do not occur in all modern contexts. The question is: are there any modern contexts that sufficiently resemble ancestral contexts such that we would expect anger to remain optimific? I will argue that we can expect anger and associated practices to remain more or less optimific in ordinary interpersonal relationships, but that we should be much more skeptical about the optimific nature of blame and punishment in other contexts.

To see how various contexts resemble or do not resemble the ancestral environment in ways that make anger and associated practices either optimific or non-optimific, let us take the four mismatches listed above in reverse order. Consider first that anger and associated practices like punishment (which may involve more complex algorithms, as described earlier) remain optimific only insofar as there are no alternatives available in the modern environment that can deliver the same (or more) benefits at lower costs. Insofar as anger does seem to make our disappointment in others' actions immediately clear in a way that gently telling them how they have wronged you does not, there are probably few realistic alternatives to anger in everyday interpersonal relationships in which the mistreatment in question is local and relatively minor. In cases of chronic mistreatment, however, that may no longer be true; the Hatfields and the

McCoys would have presumably done well to renounce their anger. Likewise, in legal contexts in which effective rehabilitation programs are available instead of prison or execution, it will no longer be the case that anger will always motivate optimific responses to wrongdoing. Furthermore, even when rehabilitative programs are not available, it may be that simply shortening sentences and improving conditions in prisons could reduce crime more than either maintaining or increasing the severity of the current system (Levy, 2012; Pizarro, Stenius, Vanja, & Pratt, 2006; Kennedy, 2008; Kleiman, 2009).

Another potential mismatch between the modern environment and ancestral environments is that many ancestral cues to punishment's likely consequences no longer apply in many modern contexts. For instance, ancestral cues to bargaining power included, among other factors, physical strength. Except in certain interpersonal relationships and cultures in which physical aggression remain a live option, physical strength is no longer relevant to considering whether punishment will maximize beneficial consequences. Nevertheless, those of us who are on the small side may continue to find ourselves less prone to anger in response to mistreatment by others, and strong people may find themselves *more* prone to anger (Sell, Tooby, & Cosmides, 2009). Likewise, the likelihood that men will get angry and wish to punish wrongdoers may continue to depend on how strong they perceive the wrongdoer to be (Jensen & Petersen, 2011). To the extent that these differences in size and strength continue to influence people's tendency to get angry in contexts where physical aggression truly is not an option, anger and its associated practices may fall short of optimality.

Besides physical strength, ancestral cues to future wrongdoing—such as past

wrongdoing—are increasingly imperfect predictors in the modern environment. While the fact that a friend has slighted you in the past may be a highly reliable indicator that she will slight you again in the future, the fact that a criminal has committed some crime is not always a reliable indicator that she will commit similar crimes in the future. With the advent of sophisticated rehabilitative programs and novel institutional incentives to stay out of trouble, the fact that someone has done wrong in the past no longer remains as reliable an indicator that this person will do wrong again in the future as it did in the ancestral past. Indeed, my entire error theory for non-consequentialist intuitions about moral responsibility is predicated upon the claim that ancestral cues like previous wrongdoing are imperfect predictors of future wrongdoing in the modern environment but nevertheless continue to strongly influence our decisions about blame and punishment as if they were reliable predictors of future wrongdoing. That being said, previous wrongdoing that meets certain conditions (e.g., being intentional) does remain a highly reliable indicator of future wrongdoing when sophisticated rehabilitation programs or novel institutional incentives are unavailable to shape behavior—that is, in ordinary interpersonal relationships.

Another set of cues that may not always be relevant to the modern consequentialist are cues related to whether anger-driven punishment would be publicized. Publicity would have been important in the ancestral environment, for punishment cannot deter future exploitation of the punisher if potential exploiters do not connect that punishment with the punisher, nor can it deter future exploitation of third parties if potential exploiters (other than the wrongdoer who is being punished) do not learn of the punishment at all. Fortunately, publicity remains a relevant consideration for

the modern consequentialist, and recent studies have shown that people's decisions about punishment are highly sensitive to whether other people are present to observe the punishment. For instance, Kurzban, DeScioli, and O'Brien (2007) found that participants would become angrier and pay approximately three times more to punish third parties who defected in a sequential prisoner's dilemma when they were told their punishment would be publicized compared to participants who were told their punishment would remain anonymous. Piazza and Bering (2008) found that participants in a similar economic game who were told that their punishment would remain anonymous paid less to punish third parties than those who were told their punishment would be publicized, regardless of whether they would be personally identified as the source of the punishment. Even more strikingly, Bourrat, Baumard, and McKay (2011) found that they could increase the severity of people's condemnation of an immoral action simply by inserting a picture of human eyes into their survey questionnaire.

Even though the presence of witnesses in fact remains a reliable indicator of whether punishment will have a deterrent effect, the modern environment provides additional cues to deterrence that would not have been present in ancestral environments and so do not always have the effect on anger that the consequentialist would hope them to have. For instance, Carlsmith, Darley, and Robinson (2002) found that explicitly telling participants that a particular act of punishment would either be highly publicized or not publicized at all had no effect on their decisions about punishment. Similarly, both Baron and Ritov (1993) and Baron, Gowda, and Kunreuther (1993) found that explicitly telling participants about the likely deterrent effects of an act of punishment in the context of tort law had no influence whatsoever on their decisions about whether to carry

out that punishment.

What should we make of these considerations? As before, I believe they give us reason for optimism about the consequences of anger in ordinary interpersonal relationships but pessimism about the consequences of anger outside of ordinary interpersonal relationships. While cues to publicity and the possibility of deterrence in interpersonal relationships remain more or less limited to ancestrally valid cues like the presence of witnesses, cues to publicity and the possibility of deterrence in more general contexts like tort law are no longer limited to these ancestrally valid cues. Therefore, we should worry about the reliability of anger in motivating optimific behaviors outside of interpersonal relationships, but we can remain cautiously optimistic about the benefits of anger in small-scale contexts with repeated interactions with the same people.

Consider another potential mismatch between ancestral and modern environments that might decrease the chances of anger remaining optimific today. While the anger system was designed to maximize benefits to oneself and those one cares about, the category of people one cared about in the ancestral environment was usually limited to friends, family members, and allies. In the modern environment, however, morally motivated people typically take the interests of all members of the moral community into account. Furthermore, if my arguments from moral luck are sound, morally motivated people should take into account the interests of *wrongdoers*.

This may seem to present a problem for consequentialist justifications for anger across all modern contexts. Aren't we necessarily discounting the interests of wrongdoers when we punish them? Yes, but only temporarily. Recall that, in interpersonal relationships, anger is typically a response to friends or family *undervaluing* your

interests and the interests of others you care about in favor of their own. The effect of anger in these contexts is typically to motivate the least costly behaviors required to restore your relationship with wrongdoers to a point of maximal mutual benefit, assuming that you and the wrongdoer should place equal weight on each other's interests. Tooby, Cosmides, Sell, Lieberman, and Sznycer (2008) explain how anger generally works (whatever the weights people might place on each other's interests), by using the concept of a welfare-tradeoff ratio (the amount of benefits one person is willing to forego to bring about a given amount of benefits for another person):

> There is an equilibrium WTR value, at which the marginal increase in price P would pay, in the form of a higher WTRyou, is exactly offset by the marginal increase in benefits P would gain by doing so, through increased cooperation from you. If P's WTR toward you is below this equilibrium value, the marginal decrease in your cooperation that this elicits will make P worse off than he could be. When this is true, there is the possibility of raising P's WTR toward you. By threatening to lower your level of cooperation with P—or even withdraw it by switching to a partner who values your welfare more highly (i.e., whose WTR toward you is higher)—it should be possible to raise P's WTRyou to a value closer to P's equilibrium point. (p. 264)

If we take seriously the idea that friends and family should weigh each other's interests equally, then anger should only be triggered when one person imposes costs on another greater in size than the benefits she received.[44] Furthermore, when anger is elicited, it is

[44] It may be objected that we do not in fact place equal weight on the interests of friends and family. For instance, I do not as a matter of fact share all of my money with all of my friends all of the time (of course remembering to give more to those who are poorer and less to those who are richer). Nevertheless, I suspect that those of us who are morally motivated think that the reason we do not constantly share our resources in this way is not because we do not care equally about our friends and family, but because doing this would remove the incentives our friends and family need to motivate them to undertake the sorts of tasks they must in order to maximize their well-being in the long run. For that reason, we are likely justified in placing less weight on the interests of our friends and family than we do on our own, at least in everyday circumstances. In any case, whatever weight we do place on their interests in everyday circumstances, we do expect them to place the same weight on our interests. It follows that if we are all on the same page

designed to motivate the least costly behaviors possible that could lead the wrongdoer to stop wronging you and those you care about in the future. (For a comprehensive overview of the types of responses anger should motivate at various stages of exploitation, see Tooby, Cosmides, Sell, and Sznycer, 2008, pp. 266-269) It should be obvious that this is optimific from an impartial, maximizing consequentialist perspective.

Of course, the problem here is that even if we do place the interests of our friends and family on a par with our own, we do not equally value the interests of wrongdoers and their allies when we do not regularly interact with them. In cases in which we are asked to recommend punishment for such outgroup third parties, as in legal contexts, there is no guarantee that the psychological mechanisms underlying anger will take into account their well-being in the way that it would if they were members of one's ingroup. All else being equal, then, we would expect our intuitive responses to outgroup third parties to be overly punitive. Indeed, recall that Petersen, Sell, Tooby, and Cosmides (in press) found that participants were much more likely to recommend punishment over rehabilitation for outgroup criminals than they were to recommend punishment to ingroup criminals. Likewise, Lieberman and Linke (2007) found that, holding the crime constant, altering the social category of the criminal altered participants punishment recommendations; participants recommended harsher punishments for foreigners than they did for either family members or schoolmates, and harsher penalties for schoolmates than for family members. In a second study, Lieberman and Linke (2007) altered not the

---

about how much we should weigh each other's interests in everyday life, anger should only be elicited when people do in fact fail to place adequate weight on each other's interests. And causing people to increase the weight they place on others' interests by temporarily withholding benefits, imposing costs, or threatening to do either of these things will in fact be optimific, assuming there is no other option that is less costly to implement.

social category of the criminal, but the social category of the victim. They found that punishment was highest when the victim was a family member than when the victim was a schoolmate or a foreigner. All of this suggests that our tendency to discount the well-being of people who are not our close associates may lead us to recommend excessive punishment in legal contexts and other contexts in which the potential targets of our blame and punishment are not our close associates—especially when the victims are.

That being said, that does not mean there are no contexts outside of interpersonal relationships in which punishment could be optimific. Even if people are not motivated to promote the interests of all members of the moral community in these contexts, it may have the side effect of doing so. In his (2007) "After Incompatibilism: A Naturalistic Defense of the Reactive Attitudes", Shaun Nichols lays out such a consequentialist defense of anger and punishment. Citing studies in behavorial economics (Fehr & Gächter, 2000; 2002; Fehr & Fischbacher, 2004a; 2004b), Nichols claims (a) that anger motivates punishment of non-cooperators in public goods games, (b) that punishment pushes cooperation near ceiling in public goods games, and (c) cooperation dwindles to zero when punishment is not an option. On these grounds, Nichols thinks we ought to retain the reactive attitudes, even if we are convinced consequentialists about moral responsibility.

Unfortunately, the empirical matters here are not as clear-cut as Nichols presents them. Whether punishment in these economic games is optimific in the way he claims is in fact highly contentious. In the time since Nichols' published his paper, a massive literature has exploded around this question of just how punishment affects the outcomes of various economic games, and the studies from this literature greatly complicate

Nichols' claims about anger's optimific tendencies. For instance, Dreber, Rand, Fudenberg, and Nowak (2008) offer evidence to suggest that, even if punishment in public goods games increases cooperation, it does not on balance increase benefits to the group, and in some cases decreases it. In response to Dreber et al., Gächter, Renner, and Sefton (2008) ran another experiment in which they found that punishment does in fact increase group payoffs, but only in the long run. Still further research suggests that mere expression of the reactive attitudes rather than full-blown material punishment can increase cooperation even in the short run.[45] Needless to say, the question of when and to what extent anger and punishment actually secure cooperation in these contexts, to say nothing of maximizing benefits, is tricky. A number of other issues complicate these claims, and I have neither the space nor the empirical work necessary to settle this issue here. However, I do think there are general reasons to think that, whatever the optimific ways to secure cooperation in these kinds of large-scale contexts, the consequentialist will have no psychological or theoretical problems promoting them. The reason why is that the situations in which our normal practices are most likely to be non-optimific (e.g., material punishment of third parties) are situations in which these practices typically play out over an extended period of time (e.g., in a courtroom), allowing sufficient time for conscious reflection and deliberation. I discuss this a bit further below.

There is one further mismatch between ancestral environments and the modern environment that one might think to be relevant here: in ancestral environments, one-shot interactions would have been extremely rare, but in the modern environment, one-shot interactions are everywhere. The problem here would be that we might continue carrying

[45] For a brief overview of this literature, see Houser, Kurzban, and Xiao (2011).

out non-optimific punishment in one-shot interactions, where punishment is costly and there is no possibility for future exploitation in the first place. Therefore, in a very common context (i.e., one-shot interactions with strangers), we will be strongly inclined to punish when exploited, even though doing so would hurt both wrongdoers and us while helping no one.

The problem with this claim is twofold. First, it's not always true that punishing in one-shot interactions helps no one. If punishing in these contexts helps contribute to an effective regime of general deterrence, then doing so could be optimific from a moral perspective. Secondly, even though one-shot interactions are more prevalent in the modern environment than they were in ancestral environments, they are still relatively rare for those who do not travel frequently. Even more rare are one-shot interactions in which punishment would fail to benefit third parties by contributing to an effective regime of general deterrence. It follows that in order to know whether one is confronted with a one-shot interaction in which punishment would fail to benefit third parties, one must still gather a large amount of information. This has two important upshots. On the one hand, gathering information takes time, and the opportunity cost of gathering information would likely outweigh any marginal benefits one could produce by avoiding unnecessary punishment in the rare circumstances where it would be non-optimific. On the other hand, where gathering information is not too costly, time allows angry impulses to give way to slow, deliberative processes, so the worry that we might still be inescapably motivated to punish in these situations seems unfounded. Therefore, a general policy of punishing exploitation when one does not have time to gather information will be closer to optimific than a policy of always gathering information.

When one does have time to gather information, learning that punishment would fail to maximize beneficial consequences (because exchange with this wrongdoer is a one-time event and because punishment could not contribute to general deterrence), one will be able to demotivate punishment, both because we presumably have a consequentialist motive to engage in only optimific punishment and because the angry impulses that would have motivated punishment anyway would have become weaker over time, allowing them to be overridden by slow, conscious deliberation. Only in certain contexts in which information comes at a very low cost and seems to indicate that punishment would truly fail to maximize beneficial consequences might we be tempted to engage in non-optimific punishment. But these are presumably rare contexts.

Although a great deal of empirical work remains to be done, I have argued that we can draw two very general and very tentative conclusions about the beneficial nature of anger in the modern environment. First, we can expect that anger and punishment will typically remain optimific in everyday interpersonal relationships, both because these contexts only bear on the well-being of those involved—well-being that the anger system was designed to maximize—and because the factors that regulate anger in these contexts remain largely valid in the modern environment. Secondly, in contexts beyond everyday interpersonal relationships (e.g., legal contexts), anger is likely to lead to non-optimific, excessively punitive responses, both because we are less likely to take into account the well-being of the wrongdoer and those whose interests are tied up with his and because anger will be less sensitive to valid consequentialist considerations.[46]

[46] One might think it curious that I spent an entire section arguing that we should in principle allow punishment of the innocent and yet haven't considered any policies that might do this. Like most consequentialists, I highly doubt that explicitly intending to

How do these claims impinge on the viability of consequentialism at the normative level? This pattern of facts would only pose a problem for consequentialism at the normative level if motivational worries a la Smilansky undermined our tendency to get angry in response to everyday mistreatment, and if we were left unable to temper our anger in response to mistreatment outside of ordinary interpersonal relationships. Fortunately for the consequentialist, these conditions do not seem to be met. In fact, everyday interpersonal relationships provide exactly the contexts in which we would expect the normative impotence objection to apply, and so Smilansky's worries about moral luck could not undermine our optimific anger responses in these contexts even if these responses *did* rely on presuppositions that we came to reject. Likewise, even though our anger responses may not be optimific outside of everyday interpersonal relationships, these are *exactly* the contexts in which we would expect the normative impotence objection *not* to apply. That is because the issues that arise in these sorts of contexts typically play out across time, allowing people to temper their initial, automatic responses in favor of slower, thoughtful processes like the kinds of cost-benefit analyses the consequentialist would recommend.

Smilansky might worry that, in the event that we find out that punishment really is optimific in these legal contexts, considerations of moral luck might undermine our motivation to carry out these optimific punishments. I do not believe this worry is well

---

punish the innocent will ever be an optimific policy. That being said, part of the reason why shorter sentences might be a more effective deterrent than longer sentences is that shorter sentences are rarely appealed in court. This gives short sentences the qualities of being swift and certain, which according to Kennedy (2008) are far better predictors of deterrence than sentence length. Whether these correlations amount to causation is an open question, but if they do, this provides the consequentialist with a reason to endorse a policy that might lead to more false convictions but better consequences overall.

founded. If we were convinced consequentialists and were convinced that carrying out a particular act of punishment would maximize beneficial consequences, we would surely be able to muster the motivation to carry it out, just as we are able to muster the motivation to carry out any temporarily difficult task (e.g., going to the dentist) in hopes of obtaining greater benefits (or avoiding greater costs) in the future. Perhaps we wouldn't be able to make ourselves angry to help ourselves through it, but the kinds of punishments that would likely be optimific in these contexts would probably not be the kinds of punishments that would require us to get angry anyway. To take one of the most extreme examples possible—an example I doubt will ever in fact be justified on consequentialist grounds—even carrying out the death penalty no longer requires any particular person to get angry. Although you might need anger in order to effectively deliver a clever retort to a friend who has slighted you, you do not need any anger whatsoever in order to slip a needle into somebody's arm. Or to hire someone else to do it for you.

4.3 The Collective Deterrence Problem

Even if the psychological objections enumerated above should not worry the consequentialist, it might be that another set of worries arising from game theoretic considerations should. I will call these the 'collective deterrence problem' and the 'fairness problem'. I will consider the collective deterrence problem first.

The collective deterrence problem arises from the fact that general deterrence is a public good whose production requires us to solve a collective action problem. The problem is this: no individual seems to have any reason to punish wrongdoers, and yet everyone is worse off when no one punishes wrongdoers. Unfortunately, this is not just a

problem with moral motivation—even people who are completely morally motivated seem to have no reason to punish wrongdoers. Why? The thought is that any individual act of punishment makes no difference to whether people will be generally deterred from committing wrongs, and yet each act carries moral costs by making both the punisher and the punished worse off. When every other crime is being punished, one individual act of punishment will not deter any additional would-be criminals from committing their crimes. When no crimes are being punished, one individual act of punishment will not deter any potential criminals from committing crimes. And yet punishment is always costly. Therefore, people who are concerned to maximize consequentialist benefits will never punish. And yet when no one punishes, the public good of general deterrence disappears.

Put in the language of cost-benefit analysis, the problem is that the expected utility of punishment with an eye toward general deterrence always seems negative. The expected benefits always seem to be zero, and the expected costs always seem non-zero. If the consequentialist is to overcome this apparently insurmountable problem, she will have to provide reason to think these estimates are false.

Contra conventional wisdom, Shelly Kagan (2011) has recently provided reason to think that the relevant expected utility analyses in collective action problems of this sort in fact favor contribution over defection. In the case of punishment, this would mean that the expected utility of punishment would have to be positive. How could this be? If no individual act of punishment ever makes a difference either way, how can its expected utility ever be anything above zero? Kagan's move is simply to deny that no individual act ever makes a difference. Kagan's argument is a straightforward reductio ad

absurdum: If no individual act ever made any difference, then it would be impossible for any number of such acts to together make a difference. And yet many such acts together do make such a difference. Therefore, at least some individual acts *must* make a difference.

To see how Kagan's argument might work in practice, imagine a regime of general deterrence in which the government allows banks to foreclose on people's houses when they cannot pay their mortgage in order to incentivize other homeowners to pay their mortgage. We can assume this system is largely effective and makes it possible for more to people to take out mortgages and purchase homes than otherwise would be able to, which is presumably a good thing in the absence of a real estate bubble. In the simplest possible case, there is some threshold rate of foreclosure (say, 75 out of 100) such that the first 74 foreclosures bring no deterrent benefits to anyone whatsoever, but the 75th foreclosure brings along with it some amount of deterrent benefits and any additional punishments bring no benefits to anyone. If the overall state of affairs in which these deterrent benefits are had is preferable to the state of affairs in which there are no deterrent benefits (but also no punishments), then the expected utility of punishment must be greater than that of non-punishment. Even though the probability that your individual act of punishment will be the 75th punishment is only 1 out of 100 (assuming the principle of indifference), the state of affairs that the 75th punishment would bring about would presumably be at least 100 times better than the good outcome that you could individually bring about (with certainty) by not punishing. (Otherwise, why would you care if no one punished?) Therefore, the expected utility of punishing will be greater than the expected utility of not punishing.

In the real world, general deterrence is not so simple. Rather than there being just one threshold, there are likely multiple thresholds that each trigger greater and greater amounts of deterrence. In any case, it will be false to say that the expected utility of punishment is zero. In fact, as long as the added deterrent benefits of meeting each threshold are great enough to outweigh the low chance that any individual's act of punishment will be the act that crosses the threshold, expected utility will favor punishment. And as long as the world in which everyone punishes is better than the world in which no additional people punish, the added deterrent benefits of meeting each threshold *will* be great enough to outweigh the low chance that any individual's act of punishment will be the act that crosses that threshold. (If a world with all punishing is better than a world with none punishing and we have no knowledge of whether others will punish, the ratio of the added deterrent benefits of crossing the threshold to the probability of *your act* crossing the threshold will always be positive.)

And yet there still seems to be a problem: if deterrent effects are already more or less at ceiling, then it really does seem to be true that any individual act of punishment will stand no chance of making a difference. Consider an example given by Alan Goldman in which a judge has to decide whether to allow a bank to foreclose on a poor, elderly widow:

> The additional assets to the bank will have a negligible effect on its overall financial position, while the widow will suffer greatly if evicted. Similarly, a single court's decision on moral merits instead of law will have little effect on the stability or predictability of the legal system. The problem is that the cumulative effect of many judges reasoning only on these grounds could be disastrous to the legal and financial institutions (also: in our example, to the ability of widows to obtain loans). This special fallibility, the inability in the absence of a rule to take account of overall effects in the single case, together with the fact that morally minded judges will be tempted to bypass law on moral grounds in such cases, justifies the imposition of a rule requiring a legal decision according to law and

not unfettered moral perception. (2002, p. 43)

In this case, the rate at which poor elderly widows are foreclosed on far exceeds the threshold rate that would be needed to maintain an effective deterrent regime. Consequently, expected utility truly does not seem to favor foreclosure.

Expected utility does seem to favor foreclosure in this case, and this is because the judge can be quite confident that his decision to foreclose would have no chance of triggering greater benefits. The judge can be quite confident that his decision to foreclose would have no chance of triggering greater benefits because the judge knows that most other judges will continue to let the banks foreclose on other poor elderly widows. But one judge defecting does not a collective action problem make. In order for general deterrence to generate a collective action problem, a majority of morally motivated judges must know that a majority of morally motivated judges will cooperate (or defect). But this is impossible. While it is true that a majority of judges can know that, at the current rate of foreclosure, their decision not to allow the bank to foreclose on the widow would have no effect on the financial system, they cannot all simultaneously know that their *next* decision not to foreclose would have no effect on the financial system. This is because, by stipulation, they are all still deciding whether to cooperate or not. In other words, people's decisions are indeterminate, and the decision to foreclose again becomes a decision under uncertainty.

So what is the expected utility of foreclosure in the case in which the majority of judges are still deciding whether to foreclose? If the threshold rate of foreclosure for an effective deterrent regime is 75 out of 100 possible foreclosures, and there is no way of knowing what all of the other morally motivated judges are going to choose, then it

seems rational by the principle of indifference to assign the outcome of being the 75th foreclosure (or, more accurately, of being a member of a cohort of exactly 75 foreclosures) a probability of 1 out of 100. As long as a world with an effective deterrent regime is better than a world in which none of the 100 judges approved of foreclosure—but 100 poor, elderly widows got off the hook—expected utility favors foreclosure.[47]

If this is right, then in a situation in which 100 consequentialist judges were each making a decision under uncertainty about whether to let off one of 100 poor, elderly widows, 100 poor, elderly widows would end up being foreclosed on. This is not ideal. Since only 75 widows need to be foreclosed on in order to avoid legal and financial ruin, 25 widows are being needlessly harmed. If I am right and the judges truly are making a decision under uncertainty—that is, they have no way to coordinate their decisions—this is simply unavoidable. If, however, we are making a decision in which we have some information about the decisions of other judges, then it becomes possible to inch closer to the threshold. For instance, if you are sure that 75 of the 100 judges are strict rule-followers, then you ought to let your widow off the hook. Not to do so would be to needlessly harm her.

4.4 The Fairness Problem

I was just assuming that there are circumstances where we can know the threshold and where judges can and should act so as to approximate it. Even granting this, Goldman would object. Goldman's objection can be seen most clearly when considering

[47] Where the net benefits of an effective deterrent regime are A and the net benefits of letting the widow off the hook are B, the expected utility of foreclosure is A/100 and the expected utility of letting the widow off the hook is B (i.e., we *know* that the widow will benefit from being let off the hook, and we *know* that we can singlehandedly let her off the hook). As long as A > 100(B), expected utility will favor foreclosure.

a case like the one just discussed:

> Even if judges could identify the optimal pattern and confer on ways to achieve it, they could not in good conscience assign themselves the required different roles in regard to applying the law. This means that they could not adopt randomizing strategies to achieve the same effect either. If some impoverished people were allowed to keep their homes while others were evicted, these egregious violations of the fundamental principle not to treat cases differently without morally relevant differences between them would be as damaging to the legal system as would crossing the original threshold. The optimal pattern from a purely consequentialist viewpoint would soon be upset as citizens reacted to these considerations of (comparative) fairness. (Goldman, 2002, p. 45)

There are two things to say about Goldman's worries. First, if Goldman were right that this policy would upset citizens, thus undermining the authority of the law, then the consequentialist would have a good reason not to let any of the poor elderly widows off the hook. But that's nothing the consequentialist can't accept. Goldman's more fundamental objection is the fairness problem: it's unfair to let some people off the hook if you can't let everyone off the hook. Paul Russell voices a similar set of objections to David Hume's weak consequentialism:

> First, even if we concede that it is possible for us to 'check' our desire for retaliation in light of utilitarian considerations, such a policy seems to encourage and condone widespread hypocrisy and insincerity. That is to say, if we are constantly deciding whether or not to express our moral sentiments, in word or deed, on the basis of (forward-looking) utilitarian considerations then it seems clear that we will frequently have reason not to express our true or sincere sentiments towards one another. In these circumstances, while our moral sentiments will be conditioned by backward-looking considerations, our actual treatment of the individuals who are the objects of these sentiments will be shaped (at least in part) by forward-looking considerations. This discrepancy between, on the one hand, the way we think and feel about one another and, on the other hand, the way we treat one another – if it is livable – will require a considerable loss of spontaneity. Moreover, on the basis of such a policy we may find ourselves treating individuals in inconsistent ways. For example, in certain circumstances utilitarian considerations may dictate that we punish some individuals for relatively minor failings while other individuals, whom we more strongly disapprove of, go unpunished (because punishment would serve no further purpose in their case). Similarly, we may find that two individuals who are the objects of similar negative moral sentiments should nevertheless be treated

> differently in respect of punishment because it will be of utility to punish one but not the other. Such circumstances may be unlikely or improbable, perhaps, but it is equally clear that they could well arise. In short, it may be argued that even if we concede to Hume that it is possible to curb our retributive practices in the way that he suggests, we will, nevertheless, have to pay a high price for such a policy in terms of hypocrisy, insincerity and inconsistency. We should be reluctant, the critic may argue, to allow such a large discrepancy between our thought and action, feeling and practice, to develop. (1990, p. 561)

Throughout this paper, I have argued that holding one another moral responsibility ought to be considered a tool intended only to bring about beneficial consequences. Although I have talked primarily about well-being, I have generally tried to be as neutral as possible as to the nature of those beneficial consequences. For that reason, it is perfectly consistent for a consequentialist about moral responsibility to think that fairness, hypocrisy, insincerity, inconsistency, and unfairness are all moral evils that ought to be avoided. That being said, I do not think that allowing some widows to keep their homes while allowing others to get foreclosed on would necessarily involve anything that should be called hypocrisy, insincerity, inconsistency, or unfairness. At least, if it does involve hypocrisy, insincerity, inconsistency, or unfairness, it is hypocrisy, insincerity, inconsistency, and unfairness of a benign sort.

Why would it be unfair to allow a few widows off the hook if you could not allow all of the widows off the hook? Goldman claims that it would require us to "treat cases differently without morally relevant differences between them" (2002, p. 45). But would it? Perhaps it would. Perhaps there is no fact of the matter about *which* of the two widows should be let off the hook without foregoing any deterrent benefits, and perhaps the fact that *that one* was let off the hook rather than the other is morally arbitrary in a way that is objectionable. But the question is: is this moral cost so great as to prevent us from bringing about an increase in the well-being of either of the two widows? Surely not.

Even if the moral arbitrariness of which of the two widows gets off the hook is a morally bad feature of the case, it is certainly less bad than reducing the well-being of either of the two widows.

The situation here recalls a famous objection to egalitarian theories of distributive justice. According to the leveling down objection to egalitarianism, if we have a choice between an economic system that makes everyone unequally better off (i.e., some are even better off than others) and a system that makes everyone equally worse off, we have at least some reason to prefer the world in which everyone is equally worse off. But that seems absolutely wrong. Even if the second world (or the world in which only one widow gets off the hook) is unfair, it does not seem to be unfair in any sense that we should try to avoid. (Parfit, 1997)

The same sorts of responses can be given to Russell's worries about hypocrisy, insincerity, and inconsistency. Even if we do automatically feel one way about a particular person, it is not necessarily hypocritical or insincere of me not to express it. If I am a convinced consequentialist, it is not hypocritical or insincere of me to reject my automatic inclination to blame and punish once I realize that doing so would be non-optimific. And, again, there may be some sense in which I am being inconsistent when I let one widow off the hook but not the other, but it does not seem like this is any kind of inconsistency I should be concerned to avoid. In fact, being *consistent* here would seem to require of me a cruel and unusual disregard for other people's well-being. And I would prefer inconsistency to that any day.

To see just how wrong it would be to care about fairness given the assumption that moral luck makes us all morally equal (or helps us recognize that we are all morally

equal), consider Saul Smilansky's recent remarks on what he calls 'funishment'. According to Smilansky, recognizing the pervasiveness of moral luck in producing people's actions—good and bad—requires us not to punish wrongdoers, but to funish them. The difference between punishment and funishment is that funishment requires—out of a demand for fairness—that wrongdoers who are removed from society be compensated for their hardship:

> Funishment would resemble punishment in that criminals would be incarcerated apart from lawful society; and institutions of funishment would also need to be as secure as current prisons, to prevent criminals from escaping. But here the similarity ends. For institutions of funishment would also need to be as delightful as possible. They would need to resemble five-star hotels, where the residents are given every opportunity to enjoy life. This would go beyond material conditions: each criminal will need to be permitted considerable leeway in running his or her own personal lives, as well as a large measure of freedom of social interaction (including frequent visits from outsiders, when possible). (Smilansky, 2011, p. 355)

But if we were to funish rather than punish, we would no longer be able to deter crime:

> Criminals currently have to balance the temptations of crime with the risks of punishment: the risk that, if caught, they are likely to spend many of the best years of their lives in miserable, ugly, harsh, nasty, violent and otherwise highly unpleasant institutions. Some people nevertheless take the risk, while many others are deterred. But once funishment replaces punishment, matters change radically. The potential offender knows that, if he is not caught, he can enjoy the spoils of his crime. But even if he is caught, he faces only some time in an institution of funishment, which – apart from being separated from lawful society – will be like a fabulous holiday. Indeed, society would be committed to doing all that it can to assure that the level of well-being of the captured criminal would not be lower than life on the outside. For, we recall, according to hard determinism no one deserves to be harmed (and to be made worse off than another), whatever crime he or she has committed. So if society singles out a criminal for detention, society must fully compensate this person for the privation involved. (Smilansky, 2011, p. 359)

And the problems only get worse:

> Whether by failing to provide a sufficient disincentive for crime, or perhaps through providing even a positive motive to engage in it, following hard determinism would lead to a flood of crime. The number of people who would

> need to be kept apart from lawful society would increase enormously. Many people who would otherwise not have become involved in crime, nor ever suffer detention, would be caught up in that very life. In the meantime, the rest of us would be living in the worst possible world: suffering unprecedented crime waves while paying unimaginable sums for the upkeep of offenders in opulent institutions of funishment. (Smilansky, 2011, p. 360)

As far as I am concerned, Smilansky has given us a beautiful argument for why we should be consequentialists about moral responsibility. Smilansky, however, rejects consequentialism:

> But why cannot the hard determinist [hard incompatibilist] avoid my argument by simply opting for some sort of utilitarian-like consequentialism? Using (non-desert) consequentialist considerations would let the hard determinist avoid the reductio, since the requirement that I posited for compensating the criminals, and the concomitant motivational catastrophe, would never get off the ground. However, a hard determinist arguing in this way would betray the moral force of hard determinism. Utilitarianism has, of course, been available for a long time, but people worried about free will and moral responsibility are concerned with more than utility. This concern has been the basis for taking the moral high ground as against a solely utilitarian justification of blame, guilt, and harsh punishment, rightly saying that making people suffer guilt or punishment just because doing so would be *socially useful* is morally unacceptable. We cannot use the morally innocent in such ways even if it furthers social interests. Blame and punishment must also be just, and not only socially efficient. And in order for them to be just, they must follow upon the choices and actions of moral agents, who through their free actions have made themselves liable to blame and punishment. That is why the punishment of the innocent is the paradigm of injustice (yet it is of no principled concern to a utilitarian). But for hard determinists, everyone is morally innocent! Hard determinism as a moral position thus holds that no one deserves to be made to suffer, or to be made worse off than another, and hence that it would be unjust to do so. Hard determinism cannot turn to consequentialism for assistance in overcoming the reductio, for it would thereby completely betray itself as a distinct ethical position. (Smilansky, 2011, p. 362-363)

As far as I can tell, Smilansky's argument against going the consequentialist route is that it would be fundamentally unfair to punish criminals rather than to funish them. But, given that funishment would ultimately lead *everyone* to be worse off—albeit equally so—it is not clear to me why we should care at all about fairness here. When the costs of

being fair involve massive blows to the well-being of everyone in the community *including the few who would otherwise be punished,* favoring fairness over everything else seems absolutely despicable.

I have argued in this section that the debate about consequentialism about moral responsibility is not just academic; it very well may have important implications for how we go about holding each other morally responsible. Furthermore, I have argued that the domains in which the consequentialist will likely want to maintain or reform our current practices are exactly the domains in which attempts at maintenance or reformation will be least likely to fall victim to the self-defeat worries considered earlier. That being said, I have not given anything close to a conclusive argument that we ought to pursue any particular set of reforms (or not). Questions about what we should actually do to promote the values identifies by consequentialism about moral responsibility are best pursued not by philosophers but by psychologists, criminologists, and economists. If my arguments have succeeded, they have at least established that these questions are worth pursuing seriously.

Chapter 5: Conclusion

5.1 Summary

I have argued that we should be consequentialists about moral responsibility. In Chapter 1, I examined why previous versions of consequentialism failed, what it is a consequentialist theory of moral responsibility ought to do, and why re-envisioning the responsibility theorist's project in this way upends traditional investigations of moral responsibility. In Chapter 2, I laid out a particular consequentialist view of moral responsibility and defended it against objections. After laying out an extensive error

theory for a particularly robust non-consequentialist intuition, I argued that combining this error theory with considerations about moral luck should lead morally motivated people to care about the interests of wrongdoers. In Chapter 3, I extended these arguments to support the conclusion that there is no in-principle prohibition against punishing the innocent. In Chapter 4, I addressed various practical objections to implementing consequentialism about moral responsibility in our daily lives and legal institutions.

Acknowledgments

The ideas presented in this thesis were refined over the course of countless hours in the office of my advisor, Professor Matt Haug. I cannot thank Professor Haug enough for his patience and support throughout not only the writing of this thesis, but my entire undergraduate career. I would also like to thank Professor Neal Tognazzini, whose door was always open for quick (and not so quick) conversations about my crazy ideas, and Professor Chris Freiman, who was always eager to talk with me about error theories. Thanks are also due to Professor Laura Ekstrom, who graciously supervised my independent study on luck and human agency during the spring of my junior year.

This thesis relies heavily on empirical work, some of which was original and the vast majority of which was not. With respect to the original work, I owe a great deal to Jonathan Phillips and the staff at Experiment Month, as well as to Professor Harvey Langholtz. For introducing me to evolutionary moral psychology, Professor Lee Kirkpatrick has my deepest gratitude. I would also like to thank Dr. Michael Petersen for kindly providing me with a copy of Petersen, Sell, Tooby, & Cosmides (in press) before it had even been accepted for publication. Eric Mandelbaum was also kind enough to provide pre-publication drafts of his (2012) and (in press), which introduced me to a significant chunk of the empirical literature cited in this thesis.

Lastly, I would like to thank the Charles Center for generously funding this project. Without a William & Mary Honors Fellowship, I would have spent the summer before my senior year working a job to help pay for school rather than doing philosophy. Because I had that time to read, write, think, and conduct empirical work, this thesis is far better than it could have been.

References

Aharoni, E. (2009). *Why do we punish? Studies of lay judgments against criminal offenders*. University of California, Santa Barbara, United States. Retrieved from http://search.proquest.com/docview/304852873?accountid=15053

Arneson, R. J. (2003). The smart theory of moral responsibility and desert. In S. Olsaretti (Ed.), *Desert and justice* (pp. 233–258). Oxford: Clarendon Press.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology*, *43*, 437-449.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244. doi:10.1037/0022-3514.71.2.230

Baron, J., Gowda, R., & Kunreuther, H. (1993). Attitudes toward managing hazardous waste: What should be cleaned up and who should pay for it? *Risk Analysis*, *13*(2), 183–192. doi:10.1111/j.1539-6924.1993.tb01068.x

Baron, J., & Ritov, I. (1993). Intuitions about penalties and compensation in the context of tort law. *Journal of Risk and Uncertainty*, *7*(1), 17–33. doi:10.1007/BF01065312

Barrett, H. C. (2005). Enzymatic computation and cognitive modularity. *Mind and Language*, *20*, 259-287. doi: 10.1111/j.0268-1064.2005.00285.x

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994) Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243-267.

Bem, D. J. (1970). *Beliefs, Attitudes, and Human Affairs.* Belmont, CA: Brooks/Cole.

Bennett, C. (2002). The varieties of retributive experience. *The Philosophical Quarterly*, *52*(207), 145–163.

Bennett, J. (2008). Accountability (II). In M. McKenna, & P. Russell (Eds.), *Free will and reactive attitudes: Perspectives on P.F. Strawson's "freedom and resentment"*. Burlington, VT: Ashgate

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*(6), 1142–1163. doi:10.1037/0022-3514.70.6.1142

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828-841.

Bootzin, R. R., & Bailey, E. T. (2005). Understanding placebo, nocebo, and iatrogenic treatment effects. *Journal of Clinical Psychology*, *61*(7), 871–880. doi:10.1002/jclp.20131

Bootzin, R. R., Herman, C. P., & Nicassio, P. (1976). The power of suggestion: Another of misattribution and insomnia. *Journal of Personality and Social Psychology*, *34*(4), 673–679. doi:10.1037/0022-3514.34.4.673

Bourrat, P., Baumard, N., & McKay, R. (2011). Surveillance cues enhance moral condemnation. *Evolutionary Psychology*, *9*(2), 193–199.

Brockner, J., & Swap, W. C. (1983). Resolving the relationships between placebos, misattribution, and insomnia: An individual-differences perspective. *Journal of Personality and Social Psychology*, *45*(1), 32–42. doi:10.1037/0022-3514.45.1.32

Burnette, J. L., McCullough, M. E., Van Tongeren, D. R., & Davis, D. E. (2012). Forgiveness results from integrating information about relationship value and exploitation risk. *Personality and Social Psychology Bulletin*, *38*(3), 345 –356.

Capes, J. A. (in press). Mitigating soft compatibilism. *Philosophy and Phenomenological Research*.

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*(2), 119–137. doi:10.1007/s11211-008-0068-x

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299. doi:10.1037/0022-3514.83.2.284

Carlston, D. E., & Skowronski, J. J. (2005). Linking versus thinking: Evidence for the different associative and attributional bases of spontaneous trait transference and spontaneous trait inference. *Journal of Personality and Social Psychology*, *89*(6), 884–898. doi:10.1037/0022-3514.89.6.884

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition," *Behavioral and Brain Sciences*, *32*, 121–82.

Carruthers, P. (2010). Introspection: Divided and partly eliminated, *Philosophy and Phenomenological Research*, *80*, 76–111.

Corwin, M. (1982, May 16). Icy killer's life steeped in violence. *Los Angeles Times*, pp. 1, 3, 29–30.

Cranston-Cuebas, M. A., Barlow, D. H., Mitchell, W., & Athanasiou, R. (1993). Differential effects of a misattribution manipulation on sexually functional and dysfunctional men. *Journal of Abnormal Psychology*, *102*(4), 525–533. doi:10.1037/0021-843X.102.4.525

Crawford, M. T., Skowronski, J. J., Stiff, C., & Leonards, U. (2008). Seeing, but not thinking: Limiting the spread of spontaneous trait transference II. *Journal of Experimental Social Psychology*, *44*(3), 840–847. doi:10.1016/j.jesp.2007.08.001

Darwall, S. (2006). The second-person standpoint: Morality, respect, and accountability. Cambridge, MA: Harvard University Press.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5-18.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348–351. doi:10.1038/nature06723

Dretske, F. (1995.) *Naturalizing the Mind*. Cambridge: MIT Press.

Dretske, F. (2004.) Knowing what you think vs. knowing that you think it. In R. Schantz (Ed.), *The externalist challenge* (pp. 389-399). Berlin: Walter de Gruyter.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*, 1087-1100.

Erdley, C. A., & D'Agostino, P. R. (1988). Cognitive and affective components of automatic priming effects. *Journal of Personality and Social Psychology*, *54*(5), 741–747. doi:10.1037/0022-3514.54.5.741

Fehr, E. & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*, 187–190.

Fehr, E. & Fischbacher, U. (2004b). Third party punishment and social norms. *Evolution and Human Behavior, 25*, 63–87.

Fehr, E. &Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*, 980–994.

Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137–140.

Ferreira, M. B., Garcia-Marques, L., Hamilton, D., Ramos, T., Uleman, J. S., & Jerónimo, R. (2012). On the relation between spontaneous trait inferences and intentional inferences: An inference monitoring hypothesis. *Journal of Experimental Social Psychology*, *48*(1), 1–12. doi:10.1016/j.jesp.2011.06.013

Fischer, J. M. (2011). The zygote argument remixed. *Analysis*. doi:10.1093/analys/anr008

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Fischer, J. M., and Tognazzini, N. A. (2011). The physiognomy of responsibility. *Philosophy and Phenomenological Research*, *82*(2), 381-417.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510. doi:10.1126/science.1164744

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*(4), 708–724. doi:10.1037/0022-3514.78.4.708

Galinsky, A. D., & Moskowitz, G. B. (2006). Further ironies of suppression: Stereotype and counterstereotype accessibility. *Journal of Experimental Social Psychology*, *43*(5), 833–841. doi:10.1016/j.jesp.2006.09.001

Gendler, T. S. (2008a). Alief in action (and reaction). *Mind & Language*, *23*(5), 552–585. doi:10.1111/j.1468-0017.2008.00352.x

Gendler, T. S. (2008b). Alief and belief. *Journal of Philosophy*, *105*(10), 634–663.

Gilbert, D., Tafarodi, R., & Malone, P. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*(2), 221–33.

Goldman, A. H. (2002). *Practical Rules: When We Need Them and When We Don't*. New York: Cambridge University Press.

Gopnik, A., and A. Meltzoff. 1994. Minds, bodies and persons: Young children's understanding of the self and others as reflected in imitation and 'theory of mind' research. In S.T. Parker, R. W. Mitchell, & M. L. Boccia (Eds.), *Self-Awareness in Animals and Humans*. New York: Cambridge University Press, 166–186

Graham, S., & Lowery, B. S. (2004). Priming unconscious racial stereotypes about adolescent offenders. *Law and Human Behavior*, *28*, 483-504.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. doi:10.1037/0022-3514.74.6.1464

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. doi:10.1037/0033-295X.108.4.814

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp., 852-870). Oxford: Oxford University Press.

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method, and Point*. New York: Clarendon Press.

Harman, G. (2009). Guilt-free morality. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics: Volume four* (pp. 203-214). New York: Oxford University Press.

Heatherton, T. F., Polivy, J., & Herman, C. P. (1991). Restraint, weight loss, and variability of body weight. *Journal of Abnormal Psychology*, *100*(1), 78–83. doi:10.1037/0021-843X.100.1.78

Hume, David. (1888). *A Treatise of Human Nature*. (L.A. Selby-Bigge, Ed.). Oxford: Clarendon Press.

Higgins, E. T. (1989). Knowledge accessibility and activation: Subjectivity and suffering from unconscious sources. In J. S. Uleman (Ed.), *Unintended thought* (pp. 75–123). New York: Guilford Press.

Houser, D., Kurzban, R., & Xiao, E. (2011). Social and biological evidence on motives for punishment. In O. Vartanian & D. R. Mandel (Ed.), *Neuroscience of decision making* (pp. 243-256). New York: Psychology Press.

Jensen, N. H., & Petersen, M. B. (2011). To defer or to stand up? How offender formidability affects third party moral outrage. *Evolutionary Psychology*, *9*(1), 118–136.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420–1436. doi:10.1037/0278-7393.20.6.1420

Kagan, S. (2011). Do I make a difference? *Philosophy & Public Affairs*, *39*(2), 105–141. doi:10.1111/j.1088-4963.2011.01203.x

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*(5), 871–888. doi:10.1037/0022-3514.78.5.871

Kearns, S. (2011). Aborting the zygote argument. *Philosophical Studies*, 1–11. doi:10.1007/s11098-011-9724-3

Kellogg, R., & Baron, R. S. (1975). Attribution theory, insomnia, and the reverse placebo effect: A reversal of Storms and Nisbett's findings. *Journal of Personality and Social Psychology*, *32*(2), 231–236. doi:10.1037/0022-3514.32.2.231

Kennedy, D. M. (2008). *Deterrence and Crime Prevention: Reconsidering the Prospect of Sanction*. New York: Routledge.

Kihlstrom, J. F. (2004). Implicit methods in social psychology. In C. Sansone, C. C. Morf, and A.T. Panter (Eds.), *The SAGE handbook of methods in social psychology* (pp. 195–212), Thousand Oaks, CA: Sage.

Kleiman, M. A. R. (2009). *When brute force fails: How to have less crime and less punishment*. Princeton, NJ: Princeton University Press.

Kurzban, R. (2011). *Why everyone (else) is a hypocrite: Evolution and the modular mind*. Princeton, NJ: Princeton University Press.

Kurzban, R., & Aktipis, A. C. (2007). Modularity and the social mind. *Personality and Social Psychology Review*, *11*(2), 131 –149.

Kurzban, R., & DeScioli, P. (2009). Adaptationist punishment in humans. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1368784

Kurzban, R., DeScioli, P., & O'Brien, E. (2007a). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84. doi:10.1016/j.evolhumbehav.2006.06.001

Kurzban, R., DeScioli, P., & O'Brien, E. (2007b). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84. doi:10.1016/j.evolhumbehav.2006.06.001

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience*, *19*(1), 42–58. doi:10.1162/jocn.2007.19.1.42

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007), Understanding and using the implicit association test: IV. In B. Wittenbrink, & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59-102). New York: Guilford.

Latus, A. (2003). Constitutive luck. *Metaphilosophy*, *34*(4), 460–475. doi:10.1111/1467-9973.00285

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology*, *72*, 252-287.

Levy, N. (2012). Skepticism and sanction: The benefits of rejecting moral responsibility. *Law and Philosophy*, 1–17. doi:10.1007/s10982-012-9128-3

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology*, *5*(1), 289–305.

Lycan, W. (2008). Phenomenal intentionalities. *American Philosophical Quarterly, 45*(3): 233–52.

Lycan, W. (1986). Tacit beliefs. In R. Bogdan (Ed.), *Belief*. Oxford: Oxford University Press.

MacLin, M. K., & Herrera, V. (2006). The criminal stereotype. *North American Journal of Psychology*, *8*(2), 197–207.

Macrae, C.N., Bodenhausen, G.V., Milne, A.B., & Jetten, J. Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, *67*, 808-817.

Madriz, E. (1997). Images of criminals and victims: A study on women's fear and social control. *Gender & Society, 11*(2), 342-356.

Mandelbuam, E. (in press). Against a robust notion of alief. *Philosophical Studies*.

Mandelbaum, E. (2010). *The architecture of belief: An essay on the unbearable automaticity of believing*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3408816).

Mandelbuam, E. (2012). Thinking is believing. Unpublished manuscript.

McCullough, M. E., Rachal, K. C., Sandage, S. J., Worthington Jr., E. L., Brown, S. W., & Hight, T. L. (1998). Interpersonal forgiving in close relationships: II. Theoretical elaboration and measurement. *Journal of Personality and Social Psychology*, *75*(6), 1586–1603. doi:10.1037/0022-3514.75.6.1586

McCullough, M. E., Worthington Jr., E. L., & Rachal, K. C. (1997). Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology*, *73*(2), 321–336. doi:10.1037/0022-3514.73.2.321

McGeer, V. (2011). Co-reactive attitudes and the making of moral community. In R. Langdon, & C. MacKenzie (Eds.), *Emotions, imagination, and moral reasoning*. New York: Psychology Press.

McKenna, M. (2008). A hard-line reply to Pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research*, *77*(1), 142–159. doi:10.1111/j.1933-1592.2008.00179.x

McKenna, M. (2009). Compatibilism & desert: critical comments on *four views on free will*. *Philosophical Studies*, *144*(1), 3–13. doi:10.1007/s11098-009-9373-y

McKenna, M. S. (1998). The limits of evil and the role of moral address: A defense of Strawsonian compatibilism. *The Journal of Ethics*, *2*(2), 123–142.

Moskowitz, G. B. (2010). On the control over stereotype activation and stereotype inhibition. *Social and Personality Psychology Compass*, *4*(2), 140–158. doi:10.1111/j.1751-9004.2009.00251.x

Nadelhoffer, T., & Matveeva, T. (2009). Positive illusions, perceived control and the free will debate. *Mind & Language*, *24*(5), 495–522. doi:10.1111/j.1468-0017.2009.01372.x

Nahmias, E., & Murray, D. (2010). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff, & K. Frankish (Eds.), *New waves in philosophy of action* (pp. 189-216). New York: Palgrave-Macmillan.

Nahmias, E., Coates, D. J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies In Philosophy*, *31*(1), 214–242. doi:10.1111/j.1475-4975.2007.00158.x

Nichols, S. (2007). After incompatibilism: A naturalistic defense of the reactive attitudes. *Philosophical Perspectives*, *21*(1), 405–428. doi:10.1111/j.1520-8583.2007.00131.x

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, *41*(4), 663–685. doi:10.1111/j.1468-0068.2007.00666.x

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. doi:10.1037/0033-295X.84.3.231

Nowell-Smith, P. (1948). Freewill and Moral Responsibility. *Mind*, *57*(225), 45–61.

O'Connor, M. E. (1984). The perception of crime and criminality: The violent criminal and swindler as social types. *Deviant Behavior*, *5*, 255-274.

Parfit, D. (1997). Equality and priority. *Ratio*, *10*(3), 202–221. doi:10.1111/1467-9329.00041

Pereboom, D. (2001). *Living without free will*. New York: Cambridge University Press.

Pereboom, D. (2008). A hard-line reply to the multiple-case manipulation argument. *Philosophy and Phenomenological Research*, *77*(1), 160–170. doi:10.1111/j.1933-1592.2008.00192.x

Pereboom, D. (2009). Free will, love, and anger. *Ideas y Valores: Revista de Colombiana de Filosofia*, *141*, 5–25.

Petersen, Michael Bang. (2010). Distinct emotions, distinct domains: Anger, anxiety and perceptions of intentionality, *Journal of Politics*, *72*(2), 357-365.

Petersen, M.B., Sell, A., Tooby, J., & Cosmides, L. (2010). Evolutionary psychology and criminal justice: A recalibrational theory of punishment and reconciliation. In H. Høgh-Olesen (Ed.), *Human morality and sociality: Evolutionary and comparative perspectives* (pp. 72–131), Hampshire: Palgrave Macmillan.

Petersen, M.B., Sell, A., Tooby, J., & Cosmides, L. (in press). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior*.

Piazza, J., & Bering, J. M. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, *6*(3), 487–501.

Pizarro, J. M., Stenius, V. M. K., & Pratt, T. C. (2006). Supermax prisons. *Criminal Justice Policy Review*, *17*(1), 6 –21. doi:10.1177/0887403405275015

Prinz, J. J., & Nichols, S. (2010). Moral emotions. In J. M. Doris (Ed.), *The moral psychology handbook.* Oxford: Oxford University Press.

Reed, J. P. & Reed, R. S. (1973). Status, images, and consequence: Once a criminal always a criminal. *Sociology and Social Research. 57*(4), 460-472.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology*, *50*(4), 703–712. doi:10.1037/0022-3514.50.4.703

Russell, P. (1990). Hume on responsibility and punishment. *Canadian Journal of Philosophy*, *20*(4), 539–563.

Sarkissian, H., Catterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language*, *25*(3), 346–358. doi:10.1111/j.1468-0017.2010.01393.x

Schlick, Moritz. (1966). When is a man responsible? In B. Berofsky (Ed.), *Free Will and Determinism*. New York: Harper & Row.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, *91*(4), 531–553.

Schwitzgebel, E. (2011) Belief. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy: Winter 2011 edition*, Retrieved from http://plato.stanford.edu/archives/win2011/entries/belief/

Sell, A. (2006). *Regulating welfare tradeoff ratios: Three tests of an evolutionary-computational model of human anger.* University of California, Santa Barbara, United States. Retrieved from http://search.proquest.com/docview/305005175

Sell, A. (2011). Applying adaptationism to human anger: The recalibrational theory. In Shaver, P. R., & Mikulincer, M., *Human aggression and violence: Causes, manifestations, and consequences* (pp. 53–70). Washington, DC: American Psychological Association.

Sell, A. N. (2011). The recalibrational theory and violent anger. *Aggression and Violent Behavior*, *16*(5), 381–389. doi:10.1016/j.avb.2011.04.013

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*(35), 15073 –15078. doi:10.1073/pnas.0904312106

Shariff, A. F., Greene, J. D., and Schooler, J. W. (2011). His brain made him do it: Encouraging a mechanistic worldview reduces punishment. Unpublished manuscript.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(7075), 466–469. doi:10.1038/nature04271

Singerman, K. J., Borkovec, T. D., & Baron, R. S. (1976). Failure of a "misattribution therapy" manipulation with a clinically relevant target behavior. *Behavior Therapy*, *7*(3), 306–313. doi:10.1016/S0005-7894(76)80056-1

Smart, J. J. C. (1961). Free-will, praise and blame. *Mind*, New Series, *70*(279), 291–306.

Smilansky, S. (2003). Compatibilism: The argument from shallowness. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *115*(3), 257–282.

Smilansky, S. (2001). Free will: From nature to illusion. *Proceedings of the Aristotelian Society (Hardback)*, *101*(1), 71–95. doi:10.1111/j.0066-7372.2003.00022.x

Smilansky, S. (2011). Hard determinism and punishment: A practical *reductio*. *Law and Philosophy*, *30*(3), 353–367. doi:10.1007/s10982-011-9099-9

Smith, A. M. (2007). On being responsible and holding responsible. *The Journal of Ethics*, *11*(4), 465–484. doi:10.1007/s10892-005-7989-5

Sripada, C. S. (2011). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research*, no. doi:10.1111/j.1933-1592.2011.00527.x

Srull, T.K., & Wyer, R.S, Jr. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, *37*, 1660-1672.

Srull, T. K., & Wyer, R. S., Jr. (1980). Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgments. *Journal of Personality and Social Psychology*, *38*, 841-856.

Strawson, P. F. (1982). Freedom and resentment. In G. Watson (Ed.), *Free will*, (pp. 256-286). New York: Oxford University Press.

Taylor, R. (1992). *Metaphysics*. Englewood Cliffs, NJ: Prentice Hall.

Thibodeau, R. and Aronson, E. (1992). Taking a closer look: Reasserting the role of the self-concept in dissonance theory. *Personality and Social Psychology Bulletin*, *18*(5), 591-602.

Todd, P. (2011). A new approach to manipulation arguments. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *152*(1). doi:10.1007/s11098-009-9465-8

Todd, P. (2012). Defending (a modified version of) the zygote argument. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 1–15. doi:10.1007/s11098-011-9848-5

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology*, *39*(6), 549–562. doi:10.1016/S0022-1031(03)00059-3

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology*, *87*(4), 482–493. doi:10.1037/0022-3514.87.4.482

Tooby, J., & Cosmides, L. (1996). Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. *Proceedings of the British Academy*, *88I,* 119-43.

Tooby, J., Cosmides, L., & Barrett, H. C. (2005). Resolving the debate on innate ideas: Learnability constraints and the evolved interpenetration of motivational and conceptual functions. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and content* (pp. 305–337). New York: Oxford University Press.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In A. J. Elliot (Ed.), *Handbook of approach and avoidance motivation* (pp. 251–271). New York: Psychology Press.

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2007). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*(1), 329–360. doi:10.1146/annurev.psych.59.103006.093707

van Inwagen, P. (1986). *An essay on free will*. New York: Clarendon Press.

van Inwagen, P. (2000). Free will remains a mystery. *Noûs*, *34*, 1–19.

Van Overwalle, F., Drenth, T., & Marsman, G. (1999). Spontaneous trait interferences: Are they linked to the actor or to the action? *Personality and Social Psychology Bulletin*, *25*(4), 450–462. doi:10.1177/0146167299025004005

Vihvelin, K. (2011) Arguments for incompatibilism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy: Spring 2011 edition*, Retrieved from http://plato.stanford.edu/archives/spr2011/entries/incompatibilism-arguments/

Wallace, R. J. (1994). *Responsibility and the moral sentiments.* Cambridge, MA: Harvard University Press.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*(2), 227-248.

Watson, G. (1987). Responsibility and the limits of evil: Variations on a Strawsonian theme, In F. Schoeman (Ed.), *Responsibility, Character, and the Emotions* (pp. 256-286). New York: Cambridge University Press.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142. doi:10.1037/0033-2909.116.1.117

Wilson, T. D., Lindsey, S., & Schooler, T. T. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126.

Wood, D. (2010). Punishment: Consequentialism. *Philosophy Compass*, *5/6*, 455–469. doi: 10.1111/j.1747-9991.2010.00287.x

Wyer, R. S., Jr. & Budesheim, T. L. (1987). Person memory and judgments: The impact of information that one is told to disregard. *Journal of Personality and Social Psychology*, *53*, 14-29.

Zimmerman, A. (2007). The nature of belief. *Journal of Consciousness Studies*, *14*(11), 61–82.

Zimmerman, D. (2008). Thinking with your hypothalamus: Reflections on a cognitive role for the reactive emotions. In M. McKenna & P. Russell. (Eds.), *Free will and reactive attitudes: Perspectives on P.F. Strawson's "freedom and resentment"* (pp. 255-272). Burlington, VT: Ashgate.