AFTER INCOMPATIBILISM: A NATURALISTIC DEFENSE OF THE REACTIVE ATTITUDES[*]

Shaun Nichols
University of Arizona

> "Some philosophers have... written as if moral concepts were a timeless, limited, unchanging, determinate species of concept, necessarily having the same features throughout their history... In fact, of course, moral concepts change as social life changes" (MacIntyre 1966)

From the first time I encountered the problem of free will in college, it struck me that a clear-eyed view of free will and moral responsibility demanded some form of nihilism. Libertarianism seemed delusional, and compatibilism seemed in bad faith. Hence I threw my lot in with philosophers like Paul d'Holbach, Galen Strawson, and Derk Pereboom who conclude that no one is truly moral responsible. But after two decades of self-identifying as a nihilist, it occurred to me that I had continued to treat my friends, colleagues, and acquaintances as morally responsible. Hardly ever did I call on my philosophical views to excuse people's actions. I'm increasingly inclined to think that my practice was appropriate and that it was my philosophical view that was defective. In this essay, I defend the practice. The guiding vision is provided in the epigraph from MacIntyre. Even though we are committed to incompatibilism about responsibility, we needn't be completely hostage to this commitment. We might relinquish or ignore the commitment rather than give up our practices that depend on moral responsibility.

In this paper, I will assume that incompatibilism is intuitive. More precisely, I will assume that the folk find it intuitive that if determinism is true, moral responsibility is undermined (see Nichols & Knobe 2007; but cf. Nahmias et al. 2006 and Woolfolk et al. 2006). Further, I will assume that determinism is true.[1] The questions of interest will be subsequent to this. If people came to believe in determinism, would this lead to major changes in our everyday lives? *Should* it lead to major changes? Nihilists tend to answer *yes* to both questions. Determinism would and should lead to major changes in our everyday interactions.[2] I'll maintain that the answer to both questions is likely *no*. In the

case of the descriptive question, the evidence is fragmentary, but it suggests that people will pretty much stick with the status quo. I'll then turn to the prescriptive question, and for the bulk of the paper, I will draw on work in emotion theory to argue that there are good reasons to resist the cries for a revolution in our everyday lives. In the final section, I'll briefly consider the major positive case for a revolution—the argument from fairness—and I'll argue that the argument is currently too underdeveloped to be at all persuasive.

1. The Descriptive Question

If people come to accept determinism, what will happen? Opinions on this question differ radically. Some maintain that this would usher in a badly needed revolution in our practices. Others worry that the recognition of determinism would lead to catastrophe. I have a more humdrum guess—if people come to accept determinism, things will remain pretty much the same.

It's natural for incompatibilists to assume that a broad-scale recognition that we lack free will would have dramatic implications. For incompatibilists think that our everyday practices of praise, blame, as well as related emotions, presuppose that we have libertarian free will. As a result, if we come to reject such free will, it would seem that the undergirding of the practices has been ripped out. This leads Joshua Greene and Jonathan Cohen to predict a major shift:

> As more and more scientific facts come in, providing increasingly vivid illustrations of what the human mind is really like, more and more people will develop moral intuitions that are at odds with our current moral practices (Greene & Cohen 2004, 28).

In particular, they say, we will stop thinking that people are responsible and that the guilty deserve punishment. "The law will continue to punish misdeeds, as it must for practical reasons, but the idea of distinguishing the truly, deeply guilty from those who are merely victims of neuronal circumstances will. . . seem pointless" (30). In this bright new future, we will dispense with our barbaric views that depend on libertarianism.

Others have a very different vision of what will happen if people come to realize that they don't have libertarian free will. Saul Smilansky worries that if people come to realize the absence of libertarian free will this will very likely harm "our fundamental values, practices, and attitudes, such as abhorrence about the 'punishment' of the innocent, the inherent value we put on 'equality of opportunity', belief in our potential for blameworthiness" (Smilansky 2000, 189). The guilt-response will also come under threat, according to Smilansky: "compunction seems conceptually problematic and psychological dubious when it concerns matters that, it is understood, ultimately one could not in fact help doing. But such genuine feelings of responsibility. . . are crucial to being

responsible selves!" (Smilansky 2002, 500). More generally Smilansky fears that if people recognize that they lack free will, they "might succumb to ... an unprincipled nihilism" (2000, 189). As a result of this, Smilansky counsels that we not disabuse people of their mistaken belief in free will. Better to let them live in the illusion of free will than to risk catastrophe.

To evaluate the descriptive question, we should be explicit about the playing field. If people come to accept determinism, this will generate a kind of competition between theoretical incompatibilist considerations and standing practices that impute responsibility. There are two primary possibilities of interest for how this competition plays out:

  (i) people will recognize that no one is really responsible and hence grant blanket exemptions

  (ii) people will persist pretty much as before in their responsibility attributions and practices.

The latter option might come about in different ways—people might renounce their commitment to incompatibilism, or they might simply neglect this aspect of their thought. For present purposes, we can remain neutral about this.

General considerations about cultural evolution provide some reason to favor (ii). For we know that emotions and motivation are powerful factors in cultural evolution. Of special relevance for our purposes, norms that resonate with our emotions are more likely to persist than norms that don't resonate (Nichols 2004). And our practices of attributing responsibility come with strong emotional backing. Consider, for instance, the retributive norm that the guilty deserve to be punished. This plausibly gets support from the fact that anger is a powerful and pervasive response to wrongdoing (see section 2.1).

There is a modicum of historical evidence that suggests that anger does play a sustaining role in punishment norms. In the history of ancient law, one central theme is that early law arose by institutionalizing vengeance. Oliver Wendell Holmes writes, "It is commonly known that the early forms of legal procedure were grounded in vengeance. Modern writers have thought that the Roman law started from the blood feud, and all the authorities agree that the German law begun in that way." (Holmes 1881). If we look at the history of criminal law in England, we find that laws against actions likely to provoke anger (e.g. assault, rape, murder, theft) all get preserved once they are entered. Most statutes do not get revoked, but the ones that did get revoked were not closely tied to actions intrinsically likely to provoke anger. Rather, here are central examples of the kinds of actions for which statutes were repealed (Stephen 1883):

Denying the trinity

Skipping church

Convincing others to skip church

Exportation of wool

Witchcraft

Taking more than 10% interest

Buying up goods to raise prices

Buying goods wholesale & then selling them wholesale

Vagrancy

Non-elite hunting game on own land

These laws lacked the kind of direct emotional support that attends actions like assault, murder, theft, and rape.

Given the emotional support that underpins our retributivist norms and responsibility practices, the general considerations about cultural evolution suggest that those norms and practices will persist. Still, that's a weak prediction base, and there might be emotional factors that push in the other direction. So we would do well to consult some evidence on the matter.

There are, as it happens, some major cultural groups that have embraced a kind of determinism—religious predestation. Such groups are found in Jewish, Islamic, and Christian traditions. The Essenes were a Jewish sect that existed around the time of Jesus, and are presumed to be responsible for the Dead Sea Scrolls. Josephus characterizes their view as follows: "the sect of the Essenes affirm, that fate governs all things, and that nothing befalls men but what is according to its determination" (Josephus *Antiquities* 13: 172). In early Islam, the Jabarites believed that God determines everything (Thompson 1950, 215; Khadduri 1984, 41). And, most familiarly, Calvinists maintain that our fates are all predestined. Calvin puts it thus: "By predestination we mean the eternal decree of God, by which he determined with himself whatever he wished to happen with regard to every man. All are not created on equal terms, but some are preordained to eternal life, others to eternal damnation" (Calvin, *Institutes,* Vol. 3, Chapter 21, section 5).

So, what are the attitudes about responsibility and justice among these groups that embrace religious determinism? Let's start with the Essenes. Josephus says that the Essenes "allot to bad souls a dark and tempestuous den, full of never-ceasing punishments" (Josephus, *War* 2: 157). And in the Dead Sea Scrolls, we find a description of an initiation rite that includes the following curse on evildoers: "Be cursed because of all your guilty wickedness! May He deliver you up for torture at the hands of the vengeful Avengers! May He visit you with destruction by the hand of all the Wreakers of Revenge! Be cursed without mercy because of the darkness of your deeds! Be damned in the shadowy place of everlasting fire!" (1 Qumram Scroll, II, 5–9). In Islam, although there is variation on whether predestination is true, all parties agreed that "Divine Justice is perfect, eternal, and ideal" (Khadduri 1984, 41). Finally we have Calvin, who espouses

an explicitly compatibilist view. He writes "none perish without deserving it" and "the reprobate suffer nothing which is not accordant with the most perfect justice" (Calvin, *Institutes,* Vol. 3, Chapter 24, sections 12 & 14).

Thus, in these examples of major cultural groups who embrace a kind of determinism, we find that the groups do not renounce responsibility. To be sure, the Essenes, Jabarites, and Calvinists have not endured so well to this day. But the key point for our purposes is that their de facto compatibilist views were not displaced by a rejection of moral responsibility. Rather, their views were displaced by *libertarian* views (see, e.g. Slone 2004, 93–96). Perhaps, then, Libertarianism culturally trumps the kind of compatibilist views we find in Calvin, but Hard Determinism apparently trumps neither.

Now, one might well complain that these cases are compromised by the fact that they derive from religious views. So let's consult the history of philosophy. Again, our question is what will happen if people come to embrace determinism. What we find in early modern is that philosophers who accept determinism are *at least* as likely to be compatibilists as they are to be hard determinists. Of the major philosophers in the period, we traditionally count Locke, Leibniz, Hobbes, Hume, Pascal, Wolff, & Condillac as compatibilists and Spinoza, d'Holbach, Diderot, Lessing, and Voltaire as hard determinists. So even among those we take to be most reflective and philosophically sensitive, we find at least as many embrace compatibilism as hard determinism.

Finally, there is a modicum of recent experimental evidence that bears on this issue. In a large cross-cultural study, we found that people who responded as determinists were more likely to respond as compatibilists about responsibility (Sarkissian et al. in prep). Using a different strategy, Adina Roskies and I conducted an experiment in which determinism was presented either as something true in an alternate universe or as something true in our own world. In the experiment, subjects read non-technical descriptions of determinism. In one condition, subjects were told to imagine another universe that was deterministic, and in the other condition, subjects were told to imagine that *our* universe was deterministic. In the "other universe" condition, subjects tended to say that it's impossible for agents to be fully morally responsible in that universe, however in the "our universe" condition, subjects tended to say that it would still be possible for agents to be fully morally responsible if our universe is deterministic (Roskies & Nichols forthcoming). This suggests that people's views about responsibility and determinism become more compatibilist when they are forced to consider determinism about their own world.

The foregoing evidence is obviously fragmentary. But every bit of it—from cultural and intellectual history to experimental evidence—points in the same direction. What *would* happen if people embraced determinism is apparently neither revolution nor catastrophe. Rather, what would happen seems to be *status quo.* Of course, we might exert effort to dislodge the status quo if we came to think that morality demanded it. So we need to ask whether we should start a revolution.

2. The Prescriptive Question

Even if a belief in determinism is unlikely to change people's behavior, many incompatibilists maintain that it *should* do so. That's part of what makes hard determinism so bracing. It typically counsels revolution. Some incompatibilists who deny free will do not promote revolution (e.g. Smilansky 2000). As a result, it will be helpful to introduce a new label. *Revolutionaries* maintain that since we lack libertarian free will, we need to radically revise our everyday practices (e.g. Pereboom 2001, 2007; Sommers 2005, 2007; Strawson 1994; Waller 1990).[3] In this section, I will offer broadly Strawsonian considerations that weigh against revolution.[4]

In the wake of P. F. Strawson's "Freedom and Resentment", perhaps the major battle between revolutionaries and counterrevolutionaries has been fought over the "reactive attitudes"—moral sentiments like resentment, indignation, guilt and gratitude which are tied up with blaming and praising others. These sentiments directly impact our day to day lives; by contrast, theoretical reflections occupy us only fleetingly and with sustained effort.[5] Revolutionaries want to extirpate the practices and attitudes that depend on the mistaken belief in more responsibility, and they are rightly sensitive to the importance of the reactive attitudes in our everyday lives (e.g., Pereboom 2007; Sommers 2007; Strawson 1986; Waller 1990).

I want to explore two Strawsonian considerations against a revolution. The first consideration is an appeal to insulationism about our normal reactive attitudes. This move might not suffice to beat back the revolutionary, but it is instructive nonetheless. The second consideration, the appeal to "gains and losses to human life," does, I argue, provide a powerful reason to oppose the revolution.

2.1. Insulationism

Strawson maintains that it would be preposterous to revise our reactive attitudes in response to considerations about determinism because the thesis of determinism is entirely external to the reactive attitudes: "our natural human commitment to ordinary inter-personal attitudes. . . is part of the general framework of human life, not something that can come up for review as particular cases can come up for review within this general framework." (1962, 83). This basic insulationist move would overturn the entire revolutionary cause because it would show that our actual commitments as revealed in our reactive attitudes are not threatened by determinism. The argument suggests both that learning about determinism won't change our reactive attitudes and that it *shouldn't*. Determinism is simply irrelevant to the appropriateness of resentment, indignation, and guilt. Such attitudes are no more hostage to determinism than they are to special relativity.

Enshrinement

Perhaps the most famous rejection of this insulationism comes from Galen Strawson. He writes, "the roots of the incompatibilist intuition lie deep in the very reactive attitudes that are invoked in order to undercut it. The reactive attitudes enshrine the incompatibilist intuition...[It] seems very difficult for us not to acknowledge that the truth of determinism... brings the propriety of the reactive attitudes seriously into doubt" (Strawson 1986, 89). The incompatibilist intuition is part of the reactive attitudes.

Pereboom also maintains that the reactive attitudes are sensitive to incompatibilist concerns. To illustrate, Pereboom draws on Gary Watson's discussion of the ruthless murderer Robert Harris. We first hear how Harris abducted two 16-year old boys from a fast food restaurant, drove them to a remote area and killed them. Then he ate one of their hamburgers. This provokes a strong sense of moral outrage. But we subsequently learn that Harris had an appalling upbringing. His father frequently beat him. His mother apparently hated the boy. His sister reported that Robert craved some physical contact from his mother: "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" (Watson 1987, 273). Pereboom notes that after you hear about Harris' terrible childhood,

> your retributive attitude diminishes, and perhaps disappears... Arguably the best explanation for this change is that your retributive attitude presupposed the belief that the killer deserved, in the basic sense, to be the object of this attitude, and because you no longer have this belief, the attitude is deprived of the presupposition that sustained it (Pereboom 2007, 202).

Robert Kane reports a similar case from his own experience, triggered by a trial of

> a young man who had raped and murdered a sixteen year old girl... My initial thoughts of the young man were filled with anger and resentment. But as I listened daily to the testimony of how he came to have the mean character and perverse motives he did have—a sordid story of parental neglect, child abuse, bad role models, and so on—some of my resentment toward him decreased and was directed toward other persons who abused and influenced him.... In such manner, the changes in reactive attitudes... are related to beliefs about ultimate responsibility (Kane 1996, 84).

Thus, Pereboom and Kane suggest, the reactive attitudes are not so isolated as Strawson had suggested (see also Nagel 1986, 125; Pereboom 2001, 95 and 99; Sommers 2005).

There are various ways to challenge this claim. We might deny that the attenuation of anger is connected with incompatibilist intuitions. But I won't press this, for I'm interested in what follows if we grant the incompatibilist the

case. Alternatively, we might challenge the claim that people's reactive attitudes about murderers are affected at all by considering their histories. But I am not inclined to make this charge either since I share the reactions that Pereboom and Kane report—my retributive emotions become less pronounced when I learn about the sad history of these murderers. Even if we grant all this, there remains an important question about how the diminishment transpires. Does the fact that our reactions are diminished vindicate the claim that "the roots of the incompatibilist intuition lie deep in the very reactive attitudes that are invoked in order to undercut it" (G. Strawson 1986)? How deep does it really go?

Narrow vs. wide psychological profiles

To proceed, we need a distinction between two ways of construing the profile of an emotion, what I'll call "narrow" and "wide" psychological profiles. I'll begin with the narrow profile. In contemporary emotion theory, specific emotions are typically characterized in terms of *local judgments* or *appraisals* and *local tendencies for action*. Despite important differences, most emotion theorists (e.g., Ekman 1992; Griffiths 1997; Haidt 2003; Lazarus 1991; Prinz 2004; Prinz & Nichols forthcoming) focus on such relatively proximal inputs and outputs of the emotion system. The local appraisals are the characteristic triggers for the emotion, and the local action tendencies are the immediate behavioral inclinations generated by the emotion.[6] To take an example that is relatively well understood, fear is caused by an appraisal that amounts to a recognition that there is an immediate danger (e.g. Lazarus 1991, 235–238; Griffiths 1997, 92, 98). This appraisal itself can be triggered fairly automatically upon learning that there is a bear in the camp or by the perceptual representation of a barn spider. The local action tendencies generated by fear is to avoid or escape the danger (e.g., Lazarus 1991, 238; Griffiths 1997, 80). (See Figure 1.)

When we're engaged in the project of characterizing a given emotion, we often want to identify the narrow psychological profile. But we might instead be interested in the *wide* psychological profile of an emotion, i.e., how a given
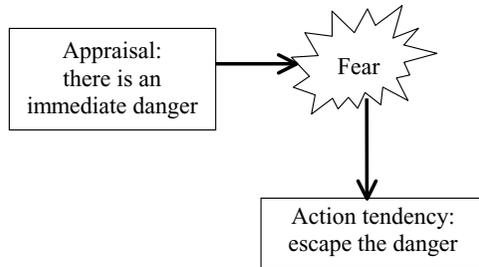

Figure 1. Narrow psychological profile of fear.

emotion presents in the overall context of the psychological system. This kind of wide psychological profile can accommodate more intelligent responses. As noted above, the percept of a barn spider triggers fear, but once a person realizes that the barn spider is harmless to humans, this can affect the wide response. If we want to characterize the emotion of fear in the overall psychological environment, we must include this kind of information.

We get a more informative picture of the wide psychological profile of an emotion if we think of the wide profile as embedding the narrow profile. On the narrow psychological profile of fear, fear is a normal, appropriate response to the perception of a barn spider. But this can be regulated at two junctures—upstream and downstream of the narrow system.[7] One way to modify the response is by altering what the fear-system gets as input. So, for instance, if a person is distracted (perhaps intentionally) from the spider-representation, then it might not end up activating the fear response. This phenomenon is familiar enough. One way to cope with unpleasant emotions is to avoid thinking about the things that trigger the emotion. This kind of modulation occurs upstream of the appraisal. Emotional reaction can also be modified by affecting what happens downstream of the narrow response. So, for instance, on recognizing that in fact the barn spider poses no threat, one might still have an unpleasant sensation, but resist the natural inclination to avoid the stimulus.

Moral anger

With this distinction between narrow and wide psychological profiles, we can now turn to the reactive attitudes themselves. I will focus on the reactive attitude *moral anger* (Keltner & Haidt 2001; see also Haidt 2003; Fessler & Haley 2003), which includes resentment and indignation (Pereboom 2007, 123). I focus on moral anger partly because there is a good deal of research on this emotion, but also because this is one of the key reactive attitudes targeted by revolutionaries.

We begin with the narrow psychological profile of moral anger. Moral anger is triggered by perceived injustices. In a large cross-cultural work, *unjust* is the highest rated factor subjects use in characterizing the trigger of an angry episode (Scherer 1997, 905). Such judgments of injustices do not require anything philosophically fancy—no Rawlsian procedure or Kantian reasoning is involved. Rather, anger is triggered by judgments of disloyalty or failure to reciprocate. Lazarus puts it in a nutshell—the appraisal for anger is the recognition of "a demeaning offense against me and mine" (Lazarus 1991, 222). The action tendency generated by moral anger is retaliation against the offender (Izard, 1977; Shaver et al., 1987; Keltner et al. 1993). Jonathan Haidt writes that the behavioral inclination is "to attack, humiliate, or otherwise get back at the person who is perceived as acting unfairly or immorally" (2003, 856). This gives us the narrow psychological profile of anger depicted in Figure 2. It's worth noting that, at least in our culture, the motivation that accompanies moral anger is not
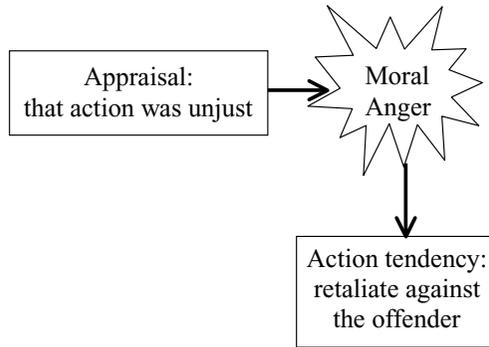
Figure 2. Narrow psychological profile of moral anger.

to rehabilitate the offender or to improve society. Rather, the motivation seems retributive in nature. In a neat study, Haidt & Sabini (forthcoming) showed subjects film clips of injustices and then asked them to rate which endings they liked best. Subjects tended to be unsatisfied by endings in which the offender was forgiven and grew as a result. Rather, subjects preferred the ending in which the offender was made to suffer, and suffer in a way that fit the original offense.

This general profile for moral anger has been corroborated by studies on how people behave in economic games. In one of these games, the ultimatum game, one subject (the "proposer") is given some money and told that he needs to offer a split with another subject (the "responder"). If the responder accepts then the money is divided accordingly; if the responder rejects the offer, then neither subject gets any money. The common finding is that when the proposer offers a highly inequitable split, the responder tends to decline. More importantly for our purposes, subjects who receive low offers often report perceiving the action as unfair and feeling angry, and the subjects who report perceived unfairness and anger are more likely to reject the offer than other subjects (Pillutla and Murninghan 1996). Using different sorts of economic games, other researchers have confirmed that inequitable offers generate anger, which in turn generates retaliation (e.g. Bosman & van Winden 2002, 159; Hopfensitz & Reuben forthcoming, MS, 13–14).

Moral anger and enshrined incompatibilism

Now let's return to the issue at hand. We have conceded that the incompatibilist intuition is in some way enshrined in reactive attitudes like moral anger. Drawing on the distinction between narrow and wide profiles, we can note that at a minimum, our concession requires that we allow that the incompatibilist intuition plays a role in the *wide psychological profile* of moral anger. For when we consider the upbringing of Robert Harris, our moral anger abates. But now,

to see how deeply enshrined this is, we need to ask whether the incompatibilist intuition is enshrined in the *narrow* psychological profile of moral anger.

On the narrow profile, an appraisal of injustice or a demeaning offence will trigger the emotion of moral anger. If an agent comes to believe that determinism is true, is moral anger still the appropriate response to a demeaning offense? By the narrow lights of the moral-anger system, the answer is presumably *yes*. It is implausible that the moral-anger system is sensitive to high-level theoretical concepts like *determinism*. The incompatibilist intuition probably doesn't go *that* deep. Just as it's narrowly appropriate for a perception of a barn spider to trigger fear, it's narrowly appropriate for a demeaning offense to trigger moral anger. It's important to keep in mind that narrow emotional systems tend not to be sensitive to a very wide range of inputs. Griffiths likens emotion systems to Fodorean modules (1997, 93–96). In the case of moral anger, the system almost certainly co-opts the phylogenetically ancient system for anger generally, and there is evidence that the response is based in phylogenetically old neural mechanisms (Sanfey et al. 2003).

Although it's likely that the narrow profile of moral anger does not enshrine the incompatibilist intuition, we do have the resources to explain why our moral anger at Robert Harris is attenuated. As noted above, sometimes our natural emotional response is regulated by processes that occur "outside" the narrow system. In the case of the barn spider, we can distract ourselves from the image, we can remind ourselves that the spider poses no threat. We can reappraise the situation by recognizing that the spider is catching pesky insects. This kind of emotion regulation is common. It is plausibly implicated in the Harris case as well. When Watson recounts Harris' childhood, he distracts the moral anger system—he leads us to attend to *something else*. Distraction alone is enough to attenuate affective responses (cf. Tracy et al. 2002). In addition, we are encouraged to *reappraise* the situation—instead of thinking about Harris' victims, we now think about Harris *as* victim. Again, reappraisal serves to attenuate or alter emotional response. We might direct some of our anger at Harris' parents. Moreover, in the case at hand, thinking of Harris as victim likely generates a new emotional response—sympathy. This new emotional response would presumably compete with moral anger, thus further suppressing the moral anger response.[8]

Let's return to the barn spider. When I perceive a barn spider, if the narrow fear system generates fear, it is responding normally. This is an appropriate narrow response to the perception of a barn spider. Similarly, when I judge that someone treated me badly, if the narrow moral anger system generates anger, it is responding normally. That is a perfectly appropriate narrow response to the judgment of an offense against me. In both cases, the wide emotional profile is much more complicated. Knowing that the spider isn't dangerous can affect my wide fear response, and knowing that the agent was determined can affect my wide moral-anger response. But there is a further question about what the *right* response is. So far we've been considering the emotion systems, both narrow and wide, from a purely descriptive standpoint. We have tried to characterize how

the emotion systems actually work, at narrow and wide levels of analysis. But we can also ask a properly evaluative question—what is the right response? In the case of the barn spider, presumably the right response is try to override the fear reaction whenever there is the slightest reason to do so. The spider poses no danger, so on reflection, it makes sense to regulate the fear response. In the case of determinism and moral anger, we have allowed that determinism does, as a matter of fact, modify the wide moral-anger response via emotion regulation and emotion competition. But again it's a further question whether we *should* respond this way. Even in the case of the barn spider, it takes effort to regulate the emotion. We have to change our focus, remind ourselves it's harmless, attempt to reappraise the situation, and so forth. Similarly, it would take significant effort to engage in the systematic regulatory override of our moral anger-system. Dislodging these reactive attitudes will not be easy.[9] The question now is, should we make the effort?

2.2. Gains and Losses to Human Life

After promoting the insulationist point, Strawson (1962) offers a second argument. He says that if we want to decide whether it's rational to retain the reactive attitudes if determinism is true, "we could choose rationally only in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment" (1962, 83). On one interpretation, Strawson is arguing that the gains and losses are the *only* thing we need consult, and since determinism is in principle irrelevant to the gains and losses to human life, determinism poses no threat whatsoever (see, e.g., Wallace 1994, 99ff.). We don't need to take such a strong stance here. For our purposes, it will suffice to treat the gains and losses as important considerations, but not as in principle unassailable. The practical considerations might get overturned by an overwhelming moral reason. We will take up that concern in section 3. For now, I will be satisfied to argue that a revolution in the reactive attitudes would generate significant losses to human life.

There are a couple of important preliminary observations before we get to the main affair of gains and losses. First, one might claim that gains and losses are beside the point. The question is whether people really are morally responsible. It might be that, given our current conception of morally responsibility, people aren't responsible. But that by itself does not mean that we should instigate a revolution in our *practices*, which is what the revolutionary is urging. Moreover, with MacIntyre, we need to recognize that even if our current notion of moral responsibility is incompatibilist, our moral notions are not immutable. In the future, our notion of moral responsibility might have different features than it does now. The second preliminary point is that all the considerations about gains and losses raised here might be construed in consequentialist terms. That doesn't mean that gains and losses *must* be interpreted in consequentialist terms. It might

be that some of the losses that we would suffer would be *deontological* losses. Although it sometimes seems like consequentialism is bound to swallow us all into its compelling maw, I am not yet prepared to succumb. The best normative ethics might be one that recognizes a plurality of incommensurable moral values or principles (Gill & Nichols forthcoming). Nonetheless, it is convenient to focus on consequences since virtually everyone agrees that consequences do matter ethically, even if they aren't all that matters.

The Strawsonian claim of interest is that our lives would be greatly diminished if we uprooted our reactive attitudes. Hard incompatibilists, who maintain that we lack free will even if determinism is false, have recently labored to assuage these worries (e.g. Pereboom 2001, 2007; Sommers 2007; Waller 1990). Part of the response maintains that many of the reactive attitudes, e.g., love, are largely unscathed by hard incompatibilism (Pereboom 2007, 121–2; Sommers 2007, 11–13). There are, however, reactive attitudes that are seriously threatened by hard incompatibilism. Pereboom writes, "Moral resentment, indignation, and guilt would likely be irrational for a hard incompatibilist, since these attitudes would have presuppositions believed to be false" (Pereboom 2007, 122; also Sommers 2007, 15). These are the attitudes targeted by the revolution. Pereboom has a two-pronged strategy for making the revolution palatable. He claims that either the targeted attitudes aren't necessary for good interpersonal lives or there are analogues that can serve in their stead (Pereboom 2007, 122). Since we've already considered in some detail the character of moral anger, we'll continue to explore the role of this reactive attitude.

Benefits of moral anger

Moral anger seems, from a rational choice perspective, completely irrational. It generates a motivation for retribution even when there is no immediate benefit for *anyone*. People who are motivated to retaliate often foresee no tangible benefit at all. They just want the offender to pay. Although people often can't see any benefits for their anger and retaliation, recent work suggests that there are indeed tangible benefits.

Let's first consider the benefits for self. In Robert Frank's excellent *Passions within Reason*, he argues that emotions generally serve to counteract immediate gratification. We tend to go for the quick pleasure, and emotions play the role of giving us short term motivation for actions that are only in our long term best interests. So, from the short-term perspective emotions like anger look irrational because they generate costly behaviors. But when we take a longer view, it becomes clear that anger is good for us. In particular, it signals intolerance for mistreatment (Haidt 2003, 856; Tavris 1989, 285), and this plausibly discourages mistreatment. It's an important feature of Frank's account that the individual is generally unaware that anger is in his long-term best interests. It does good that we don't even recognize.

Moral anger also has broader benefits. The most impressive recent empirical work comes from experimental economics. In series of ingenious studies, Ernst Fehr and colleagues have examined reactions to punishment in public goods games. A typical public goods game involves four subjects, playing anonymously on computers. Each subject is given an allotment of monetary units, and each is allowed to invest whatever portion he chooses into a common fund. For every 1 monetary unit an individual invests, 1.6 units go into the common fund, which is a net benefit for the group, but a net loss for the individual (since he only gets 40% of his investment back). Obviously it's optimal for the group if everyone invests in the common fund, but for each individual, it's selfishly better not to invest. Fehr & Gächter (2002) had subjects play a series of such games in which subjects were told (truly) that they would never interact with any player in more than one game. After each game, subjects were given an opportunity to pay to "punish" people in the group that they just played with; for each 1 monetary unit the punisher pays, 3 monetary units are deducted from the punishee's allotment. Remember that the subjects know that they will not play another game with any of these particular players, so punishing apparently has no future benefit for the subject. Nonetheless, punishment was common, and it was typically directed at defectors (i.e., individuals who contributed less than average) (Fehr & Gächter 2002, p. 137).

In a striking extension of this work, Fehr & Fischbacher (2004b) explored whether external observers, "third-parties", would be willing to punish players. The third party observed two subjects in an economic game. Fehr & Fischbacher found that about half of the third-party participants paid to punish players who violated norms of cooperation. In this study, as well as Fehr & Gächter (2002), the motivation for punishment does not seem to be anything like explicit considerations about material gains—the punisher only loses money, and they have little reason to think the punishment will materially improve the situation of the other players.

Why would subjects spend their money to punish even when it's obviously not in their material self-interest? *Anger*, say Fehr & Gächter. To investigate this hypothesis, they asked their subjects to report how they would feel under the following circumstances: "You decide to invest 16 francs to the project. The second group member invests 14 and the third 18 francs. Suppose the fourth member invests 2 francs to the project. You now accidentally meet this member. Please indicate your feeling towards this person." (Fehr & Gächter 2002, 139). Subjects reported that they would be quite angry at this person. This of course fits with the results on the ultimatum game mentioned above, in which subjects who receive low offers report high anger (Pillutla & Murninghan 1996). Anger seems to provide the critical motivation to punish defectors.

Now, what are the consequences of this anger-driven punishment? As it happens, punishment dramatically affects behavior in these games. Fehr and colleagues have consistently found increased cooperation when punishment is an available option. Perhaps the most impressive illustration comes from

experiments in which players first engage in several games in which punishment is not available. Fehr & Gächter (2000) conducted such an experiment in which subjects played for 10 rounds with no option of punishment; by the 10th round the level of contribution was quite low (below 20%). Then punishment was introduced as an option for the 11th round. Immediately, the contributions leapt to over 60% and within a few more rounds with punishment, the level of contributions was at 90%! Fehr & Gächter maintain that such punishment is driven by anger, and if they're right, then anger is a potent force for motivating cooperation (2002, 139).

There are three important facts about the relationship between punishment and cooperation in these studies. First, cooperation deteriorates without punishment. Many people start out contributing a significant amount of their endowment, but in the absence of punishment, this drops off precipitously. Second, mere knowledge that punishment is available increases cooperation. Fehr & Gächter maintain that this is because people anticipate that if they defect, the others will be angry and punish them. (A supplementary explanation is that people might be more willing to contribute if they know that they have some recourse to punish those who are taking advantage of them.) The final point is the most obvious—punishment pushes cooperation near ceiling.[10]

More generally, it's worth emphasizing that people *expect* retributive punishment if they treat others unfairly. We find this in the Fehr & Gächter study in which they anticipate that others will be angry if they fail to contribute. In their 3rd party punishment studies, Fehr and Fischbacher found that participants playing the game predict fairly well the conditions under which 3rd parties will punish (2004b, 70 [Figure 2]). It's likely that retributive punishment is effective partly because we are so receptive to it.

So, what does moral anger do for us? It's plausible that moral anger historically played a critical role in securing social norms including conventions of cooperation. The norms were probably partly fixed through a regime of punishment for defectors (see Richerson & Boyd 2005, 199–201). The experimental results suggest that moral anger *still* plays an important role in securing cooperative behavior.[11] There remain open questions about the extent to which the results in experimental economics are representative of real life. But the studies certainly suggest that moral anger and its punitive expressions work to discourage cheating, defection, and violations of reciprocity (Keltner & Haidt 2001). In Keltner and Haidt's taxonomy of social emotions, they make the plausible suggestion that social emotions like moral anger function to solve problems in the social domain. In the case of moral anger, they maintain that the function is to "motivate other to repair transgression" (2001, Table 1).[12]

Analogue for moral anger

The reactive attitude of moral anger apparently does make quite significant contributions to human life. Moral anger has benefits both at the individual level

and the group level. Of course, moral anger sometimes has bad consequences—feuds, seething resentment—and much more work needs to be done to understand the costs of moral anger. But at a minimum, it seems that we would want to be very cautious about fomenting a revolution if it means we lose the benefits of moral anger charted above. The revolutionary still has an important response, though. Recall that the claim was that for the reactive attitudes that are inconsistent with determinism, either we are better off without them *or* there are analogues that will do the needed work. We now need to consider the proposed analogues.

Once again, Derk Pereboom has provided the most extensive discussion of the matter. Returning to the case of Robert Harris, Pereboom has us recall how one's attitude changes upon learning of Harris' past:

> Indignation gradually gives way to a kind of moral sadness—a sadness not only about his past but also for his character and his horrible actions. This kind of moral sadness is a type of attitude that would not be undermined by a belief in determinism. Furthermore, I suspect that it can play much of the role that resentment and indignation more typically have in human relationships (2001, 97–98; see also 2007, 120, 203).[13]

Thus, on Pereboom's view while moral anger is theoretically irrational, moral sadness can largely serve in its stead.

Can sadness do the requisite work for moral anger, then? Well, to answer this we need to know more about sadness itself. The most important question concerns how sadness affects our behavior. For we know that moral anger produces behavior that discourages cheating, defecting, and mistreatment. What kind of behavior does sadness tend to produce? *None*, according to emotion theorists. Lazarus writes, "In sadness there seems to be no clear action tendency—except *inaction,* or withdrawal into oneself" (Lazarus 1991, 251). This is illustrated in infancy research. Infants show individual differences in their propensities to feel sad or angry when blocked from attaining a desired end—some babies are more likely to feel sad, others to feel angry. Researchers have found that when infants show sadness as their predominant emotion, this is associated with giving up (Lewis & Ramsey 2005, 518), and it seems to be akin to learned helplessness (Abramson et al. 1978). By contrast, infants who respond with anger are more likely to try to overcome the obstacle (Lewis & Ramsey 2005, 518). As a result, sadness seems too behaviorally weak to do the work of anger.

I've focused here on Pereboom, because he has offered the most explicit story about analogue emotions. But I think that the case illustrates a more general worry for revolutionaries. We cannot lightly assume that emotions are sufficiently fungible that we can co-opt emotions that are consistent with hard incompatibilism to do the work of the reactive attitudes that are under threat.

Similar considerations might apply to revolutionary treatments of guilt, though the situation is more complicated because the discussion by revolutionaries has been less clear about replacements. Emotion theorists have been fairly clear though. The basic appraisal that triggers guilt for an agent is roughly the recognition that *I have harmed someone that matters to me* (Baumeister et al. 1994; Haidt 2003, 861; Prinz & Nichols forthcoming). The action tendency for guilt is to repair the relationship, for example, by apologizing or making amends (Lazarus 1991, 243; Baumeister et al. 1994; Haidt 2003, 861). As a result, guilt is an important emotion for guiding the reparation of damaged relationships. It is probably also an important motivator for *not* mistreating people you care about, because you anticipate the guilt (Prinz & Nichols forthcoming). All of this suggests that guilt is *good*. Suffice to say that when you're looking for a mate, you probably want to avoid the ones who don't feel guilt.

What does the revolution have in store for guilt? Pereboom and Waller propose that regret can do substitute service for guilt. Pereboom writes, "suppose that you behave immorally, but because you endorse hard incompatibilism, you deny that you are blameworthy. Instead, you acknowledge that you have done wrong, you feel sad that you were the agent of wrongdoing, and you deeply regret what you have done" (Pereboom 2007, 120; see also Waller 1990, 165–7). Unfortunately, regret is not well-defined. Indeed, some theorists assimilate it to guilt (e.g. Storm & Storm 1987). Lazarus claims that the term is simply too ambiguous to be usefully compared to other emotions (1991, 244–5). Without a more detailed description of the emotion, it's hard to say whether it is likely to do the good services that guilt provides. Tamler Sommers takes a different approach. He tries to identify which aspect of guilt must be abandoned by the hard determinist. But the only aspect of guilt he identifies as incompatible with determinism "is the kind of morbid hand-wringing that keeps us awake all night thinking about what might have been" (2007, 14). That kind of guilt is probably to be rejected independently of incompatibilism, so it's not clear that Sommers' treatment of guilt here requires a revolution.

Pereboom brings in a further element that might be thought to address the shortcomings of the analogue emotions: resolve. In addition to moral sadness, we might be committed to opposing wrongdoing, and this "would allow for a resolve to resist abuse, discrimination, and oppression" (Pereboom 2007, 124). In his discussion of guilt, he says something similar: "because you have a commitment to doing what is right, and to personal moral progress, you might resolve not to perform an immoral action of this kind again, and seek out therapeutic procedures to help treat one's character problems" (Pereboom 2001, 205). People no doubt differ in the strength of their resolve. But it strikes me as unlikely that resolve will provide sufficient motivation for the bulk of the population. After all, many teenagers think that they risk going to hell if they have sex, yet this often provides insufficient motivation for abstinence. Or consider the Marxist thought that working hard will generate benefits for the state which will in turn benefit everyone. This turns out to be motivationally feeble. Marxism seems incredibly

rational, which is why it is so attractive to us intellectuals. But it turns out to be naively optimistic about the plasticity of human motivation. I suspect the same is true of the revolutionary's hope for replacing problematic reactive attitudes.

3. The Fairness Argument

Thus far I've argued that there would be significant losses to human life if we made a concerted effort to eradicate moral anger from our lives. This is a prima facie argument in favor of retaining moral anger. However, we sometimes think it appropriate to accept significant losses because of an overwhelming moral concern. And there has been a moral concern looming in the background for this entire paper: fairness. Hard determinists maintain that treating people as if they were morally responsible is morally objectionable because it's unfair. Thus, Waller writes, "It is unfair to... praise Rachel and blame Sarah... when their character and behavioral differences are products of earlier environments for which neither is morally responsible" (Waller 1990, 129; also Pereboom 2001, 156; Smilansky 2002, 493). The fairness argument is a common theme among hard determinists, though it's never articulated very fully. This is not the place for a lengthy discussion of the issue, but I will distinguish a few different interpretations of the argument, and I'll suggest that none of them is convincing without further elaboration.[14]

The following provides a basic structure for the fairness argument:

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

Before going into further detail, note that *even if* the argument succeeds in showing that it's unfair to hold people morally responsible, that still leaves open the all-in question of moral permissibility. In some cases, we regard a forbidden action like murdering an innocent as morally permissible in light of powerful countervailing factors. It's possible that we would conclude that in light of the gains to human life, it is morally permissible to treat people as morally responsible even though it's unfair to do so. But it's not yet clear that we should accept the conclusion of the fairness argument.

We've been assuming the second premise throughout the paper, so we don't want to challenge that premise. That means everything hangs on the first premise. What grounds that premise? Again, advocates of the fairness argument haven't said enough on this matter, but I presume that the answer is *intuition*. It's just intuitively plausible that it's unfair to hold determined agents responsible. I am quite willing to grant that. Indeed I fully share the intuition. But now the question is, *what is the import of that intuition?* Advocates of the fairness

argument have not been explicit about this, so let's consider three broad kinds of answers:

1. Objectivist: The intuition reflects an objective truth about fairness and blame.
2. Folk theoretic: The intuition reflects a plank in the folk theory of fairness.
3. Reflective equilibrium: The intuition is one of the starting points for a process of reflective equilibrium.

Let's consider each of these in turn.

The objectivist path is a hard row. For it depends on the controversial assumption that moral objectivism is true (cf. Mackie 1977). It further depends on the assumption that this particular intuition about fairness ("It's unfair to hold people in a deterministic universe morally responsible for wrongdoing") is objectively right. In the absence of a compelling argument for objectivism and for the objective truth of the fairness claim, I am not inclined to worry much about this version of the fairness argument.

The folk theoretic version is much less fraught. I think it's plausible that there is a folk intuition that it's unfair to hold people in a determinist universe responsible (*pace* Wallace 1994). If we take this interpretation of the import of the intuition, then we need to be more precise about the nature of the first premise. In particular, we should insert the folk qualification into the premise, which becomes:

1′. On the folk notion of fairness, it's unfair to hold people in a deterministic universe responsible.

But if this is our premise, it has dramatic consequences for the character of the argument. For the argument will no longer be a simple deduction to the conclusion "It's unfair to hold people responsible". Rather, the conclusion is something like:

3′. The folk view has the consequence that it's unfair to hold people responsible for their actions.

This conclusion is hardly trivial. It often makes sense to follow the folk. But in this case, we need to pause to recognize that the folk also embrace the view that people are often responsible for their actions. So, what we have is better cast as a dilemma. Assuming that our decisions are entirely a product of determinism, we need to give up one of our folk views:

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

It will require a substantial argument to show that we should give up the first view. And we have yet to see such an argument. Furthermore, it is worth keeping in mind that the folk view of fairness is rich and multifaceted. The incompatibilist intuition is hardly the only plank in the folk theory of fairness. We might well want to remove the incompatibilist plank while retaining most of the rest of the folk notion of fairness.

The third option is that the intuition that grounds the first premise figures into a process of reflective equilibrium. It is hard to evaluate this view without a more detailed model. But at a minimum, the reflective equilibrium enthusiast must reckon with the fact that we do find it fair to punish wrongdoers. So even if we have a strong incompatibilist intuition, we also have a strong intuition that it's fair to punish wrongdoers. It's far from obvious that under ideal reflection, we'll relinquish the latter intuition.

Of course this has been a very breezy tour of the fairness argument. There are many ways to develop the argument, and it remains to be seen how incompatibilists will proceed. My aim has been to show that the fairness argument remains too underdeveloped to justify the revolution.

4. Conclusion

This paper has promoted a counterrevolutionary agenda. Although incompatibilism might somehow be enshrined in reactive attitudes, it is probably not implicated in the narrow psychological profile—the local inputs and outputs—of the reactive attitudes. As a result, it's not surprising that the available evidence suggests that even if people come to believe in determinism, they will continue to hold each other responsible. This leaves us with the difficult question whether we should undertake a concerted effort to alter these responses. I have argued in favor of retaining moral anger and other targeted reactive attitudes. But I do not pretend to have resolved this question. A more fully developed argument from fairness might provide an overwhelming moral reason to suppress the reactive attitudes. Furthermore, the serious costs of moral anger might ultimately outweigh the benefits. My goal has been to emphasize some powerful and largely unnoticed benefits of moral anger. Pending a much fuller story about the adequacy of modern life without moral anger, I am inclined to embrace the status quo. Instead of trying to subdue moral anger in deference to the incompatibilist intuition, we are right to ignore or relinquish the commitment to incompatibilism.

Notes

was anticipated by Ted Honderich (2004), who has been an important voice for changing the way we think about the problem of free will. Earlier versions of this paper were presented at the University of San Francisco, the Southern Society for Philosophy and Psychology, and the moral psychology seminar at the University of North Carolina. Thanks to all the audiences for very useful comments.

1. The broader assumption is really that we lack libertarian free will, and various *indeterminist* theses are also incompatible with free will. I focus on determinism merely to ease exposition.

2. The focus in this paper concerns our private lives and personal interactions, not the penal system or other matters of public policy. Those issues merit a separate discussion.

3. In his important work on revisionism, Manuel Vargas distinguishes moderate revisionism from strong revisionism (Vargas 2005). Vargas characterizes strong revisionists as follows: "strong revisionism maintains that our concept, practices, or attitudes themselves are in need of elimination... Strong revisionism argues that we must dispose of some or all of the main elements addressed by a theory of responsibility" (2005, 408). On this taxonomy, revolutionaries would count as strongly revisionist. Vargas' own moderate revisionism (2007) is consistent with everything I say here.

4. Strawson concedes less to the revolutionary than I do, for he doesn't seem to acknowledge the incompatibilist intuition at all (see Vargas 2004).

5. While Strawson brought this issue to prominence, important elements of the view are anticipated by Hume (*Enquiry* Section VII), as noted by Russell (1995, 71–84) and Pereboom (2001, 90–2).

6. On some theories (e.g. Prinz 2004), the emotions can also be triggered directly by perceptions. For instance, the perception of a snarling dog can cause fear without any appraisal.

7. The phenomenon of emotion regulation is considerably more complicated than my sketch (see e.g., Ochsner & Gross 2005; Gross 2006), but the additional complexities don't change the point here.

8. The competition between anger and sympathy plays out in public debate on punishment. One group trots out the victims, in an attempt to trigger anger at the offender; the other group tries to generate sympathy for the offender by personalizing him and his plight.

9. This is entirely consonant with the views of at least some of the revolutionaries (see e.g., Sommers 2007, 20; Waller 1990, 161).

    Although it would be difficult to repress completely moral anger, it may not be beyond human capacity either. According to Briggs (1970), the Utku, an Eskimo group, largely repress anger behavior. This is plausibly because any serious social disruption could be devastating given the perils of their arctic life and the tiny size of their society, which consisted of 35 people at the time of Briggs' research. Perhaps then we could, with sustained effort, suppress our problematic reactive attitudes.

10. By contrast, reward is much less effective than punishment at securing cooperation (e.g. Andreoni et al. 2003).

11. There's an irony here for the moral anti-retributivist. Punishment secures cooperation, but punishment is costly. This leads researchers to worry that the

system itself is fragile (e.g. Henrich & Boyd 2001), for the best overall strategy for an individual is to always cooperate and never punish. Such an individual gets all the benefits of cooperation and bears none of the costs of punishing. But he is a second-order free rider. He's taking a free ride on the costly punishments of others. As it happens, 2nd-order free riding might not be much of a problem, for it might be that our anger system provides a sufficiently powerful motivation that the additional cash pales in comparison. But notice that the person who renounces retributive punishment falls in with the 2nd-order free riders. He gets the benefits of our retributivism without paying the costs.

12. Note that this function—motivate the other person to make amends—can be implemented by a mechanism for which incompatibilism is entirely irrelevant. What matters for moral anger is exhibiting your willingness to punish for no apparent material benefit (Frank 1988).

13. Relatedly, Tamler Sommers says that moral anger at, say, the murder of a family member, is irrational, but it can be "succeeded by a lasting grief. And grief, no matter how passionate or intense, is perfectly consistent with the objective attitude" (Sommers 2007, 9).

14. The treatment of the fairness argument here is indebted to discussions with Manuel Vargas. Elsewhere, Vargas and I present considerations similar to those raised here (Nichols & Vargas forthcoming), and a fuller treatment of these issues is in preparation.

References

Abramson, L., Seligman, M., & Teasdale, J. 1978. Learned helplessness in humans: critique and reformulation. *Journal of Abnormal Psychology* 87: 49–74.

Andreoni, J., Harbaugh, W., Vesterlund, L. 2003. The Carrot or the stick: Rewards, punishments, and cooperation. *The American Economic Review* 893 sq.

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

Briggs, J. 1971. *Never in Anger*. Harvard University Press.

Bosman, R. & van Winden, F. 2002. Emotional hazard in a power to take experiment. *The Economic Journal* 112: 147–169.

Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion* 6: 169–200.

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.

Fessler, D. M. T. & Haley, K. J. 2003. The strategy of affect: Emotions in human cooperation. In P. Hammerstein (ed.) *The Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press, 7–36.

Frank, R. 1988. *Passions within Reason*. Norton.

Gill, M. & Nichols, S. forthcoming. Sentimentalist pluralism: Moral psychology and philosophical ethics. *Philosophical Issues*.

Greene, J. & Cohen, J. 2004. For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London B* 359: 1775–1785.

Griffiths, P. 1997. *What Emotions* Really *Are*. Chicago University Press.

Gross, J. (ed). 2006. *Handbook of Emotion Regulation*. Guilford Press.

Haidt, J. 2003. The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (eds.) *Handbook of Affective Sciences*. Oxford: Oxford University Press, 852–870.

Haidt, J. & Sabini, J. forthcoming. What exactly makes revenge sweet?

Henrich, J. & Boyd, R. 2001. Why people punish defectors—Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology* 208: 79–89.

Holmes, O. 1881. *The Common Law*. Boston: Little, Brown.

Honderich, T. 2004. After compatibilism and incompatibilism. In J. K. Campbell, M. O'Rourke, & D. Shier (eds.) *Freedom and Determinism*. MIT Press.

Hopfensitz, A. & Reuben, E. forthcoming. The importance of emotions for the effectiveness of social punishment. Downloaded September 20, 2005 from http://www.fee.uva.nl/creed/pdffiles/HR05.pdf.

Hume, D. 1955 [1743]. *An Enquiry concerning Human Understanding*. L. Selby-Bigge (ed.). Oxford: Clarendon Press.

Izard, C. E. 1977. *Human Emotions*. New York: Plenum Press.

Kane, R. 1996. *The Significance of Free Will*. New York: Oxford University Press.

Keltner, D., Ellsworth, P. C., & Edwards, K. 1993. Beyond simple pessimism: Effects of sadness and anger on social perception. *Journal of Personality and Social Psychology* 64: 740–752.

Keltner, D. & Haidt, J. 2001. Social functions of emotions. In T. Mayne & G. A. Bonanno (eds.) *Emotions: Current Issues and Future Directions*. New York: Guilford Press, 192–213.

Khadduri, M. 1984. *The Islamic Conception of Justice*. Johns Hopkins University Press.

Lazarus, R. 1991. *Emotion and Adaptation*. Oxford University Press.

Lewis, M. & Ramsey, D. 2005. Infant emotional and cortisol responses to goal blockage. *Child Development* 76: 518–530.

MacIntyre, A. 1966. *A Short History of Ethics*. NY: Macmillan.

Mackie, J. 1977. *Inventing Ethics*. Penguin.

Mikula G. 1986. The experience of injustice: toward a better understanding of its phenomenology. In H. W. Bierhoff, R. L. Cohen, & J. Greenberg (ed.) *Justice in Social Relations*. New York: Plenum, 103–124.

Nagel, T. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. 2006. Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73: 28–53.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press.

Nichols, S. 2006. Folk intuitions about free will and responsibility. *Journal of Cognition and Culture* 6: 57–86.

Nichols, S. & Knobe, J. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous* 41: 663–685.

Nichols, S. & Vargas, M. forthcoming. How to be fair to psychopaths. *Philosophy, Psychiatry, and Psychology*.

Ochsner, K. & Gross, J. 2005. The cognitive control of emotion. *Trends in Cognitive Sciences* 9: 242–249.

Pereboom, D. 2001. *Living without Free Will*. Cambridge University Press.

Pereboom, D. 2007. Hard incompatibilism and response to Fischer, Kane, and Vargas. In J. Fischer, R. Kane, D. Pereboom, & M. Vargas, *Four Views on Free Will*. Blackwell.

Pillutla, M. & Murnighan, J. 1996. Unfairness, anger and spite. *Organizational Behavior and Human Decision Processes* 68: 208–224.

Prinz, J. 2004. *Gut Reactions*. Oxford University Press.

Prinz, J. & Nichols, S. forthcoming. Moral emotions. In J. Doris & S. Stich (eds.) *Oxford Handbook of Moral Psychology*. Oxford University Press.

Richerson, P. & Boyd, R. 2005. *Not by Genes Alone*. University of Chicago Press.

Roskies, A. & Nichols, S. forthcoming. Bringing responsibility down to earth.

Russell, P. 1995. *Freedom and Moral Sentiment: Hume's Way of Naturalizing Responsibility*. Oxford University Press.

Sanfey, A. G., Rilling, J. K., Aronson J. A., Nystrom L. E. & Cohen, J. D. 2003. The neural basis of economic decision making in the Ultimatum Game. *Science* 300: 1755–1758.

Sarkissian, H., Knobe, J. & Nichols, S. in prep. Is belief in free will a cultural universal?

Scherer, K. R. 1997. The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology* 73: 902–922.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology* 52: 1061–1086.

Slone, D. 2004. *Theological Incorrectness*. New York: Oxford University Press.

Smilansky, S. 2000. *Free Will and Illusion*. Oxford: Oxford University Press.

Smilansky, S. 2002. Free will, fundamental dualism, and the centrality of illusion. In R. Kane (ed.) *The Oxford Handbook of Free Will*. Oxford University Press.

Sommers, T. 2005. *Beyond Freedom and Resentment: an Error Theory of Free Will and Moral Responsibility*. PhD Dissertation, Duke University.

Sommers, T. 2007. The objective attitude. *Philosophical Quarterly*. 1–21.

Stephen, J. 1883. *History of the Criminal Law of England*. London: Macmillan.

Storm, C. & Storm, T. 1987. A taxonomic study of the vocabulary of emotions. *Journal of Personality and Social Psychology* 53: 805–816.

Strawson, G. 1986, *Freedom and Belief*. New York: Oxford University Press.

Strawson, G. 1994, The impossibility of moral responsibility. *Philosophical Studies* 75: 5–24.

Strawson, P. F. 1962, Freedom and resentment. *Proceedings of the British Academy* 48: 1–25. Reprinted in G. Watson (ed.) *Free Will*, second edition. Oxford University Press (2003), 72–93. Page references are to the reprinted version.

Tavris, C. 1989. *Anger: The Misunderstood Emotion*. New York: Simon & Schuster.

Thompson, W. 1950. Free will and predestination in early Islam: A critique and appreciation. *Muslim World* 40: 207–216.

Tracey, I., Ploghaus, A., Gati, J., Clare, S., Smith, S., Menon, R., & Matthews, P. 2002. Imaging attentional modulation of pain in the periaqueductal gray in humans. *The Journal of Neuroscience* 22: 2748–2752.

Vargas, M. 2004. Responsibility and the aims of theory: Strawson and revisionism. *Pacific Philosophical Quarterly* 85(2): 218–241.

Vargas, M. 2005. The revisionist's guide to moral responsibility. *Philosophical Studies* 125: 399–429.

Vargas, M. 2007. Revisionism. In J. Fischer, R. Kane, D. Pereboom, & M. Vargas, *Four Views on Free Will*. Blackwell.

Wallace, J. 1994. *Responsibility and the Moral Sentiments*. Harvard University Press.

Waller, B. 1990. *Freedom without Responsibility*. Temple University Press.

Watson, G. 1987. Responsibility and the limits of evil. In F. Schoeman (ed.) *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press, 256–286.

Woolfolk, R., Doris, J., & Darley, J. 2006. Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition* 100: 283–301.