Free Will Skepticism and Criminal Punishment

in *The Future of Punishment*, Thomas Nadelhoffer, ed., New York: Oxford University Press, 2013, pp. 49-78.

Derk Pereboom, Cornell University

Penultimate draft

Perhaps the most frequently and urgently voiced criticism of free will skepticism, the view according to which we lack the sort of free will required for moral responsibility (Pereboom 1995, 2001; cf. Smilansky 2000; Sommers 2012; Spinoza 1677/1985; Strawson 1986; Waller 1990), is that the responses to criminal behavior it would permit as justified are insufficient for acceptable social policy.[1] Indeed, some of the most prominent justifications for punishing criminals, such as retribution theory, are incompatible with free will skepticism. And alternative justifications that are not ruled out by the skeptical view per se face significant independent moral objections. But free will skepticism leaves intact other ways to respond to criminal behavior, in particular preventative detention, rehabilitation, and alteration of relevant social conditions, and I contend that these methods are morally justifiable and sufficient for good social policy.

---

[1] Thanks to Dana Nelkin, Thomas Nadelhoffer, Brian Leiter, my free will seminar at Cornell University in the Spring of 2011, and an anonymous reviewer for very helpful comments. I also wish to thank William Lee, whose work as a research assistant was highly valuable and much appreciated.

PRELIMINARIES

We human beings participate in a practice of holding ourselves and each other morally responsible. When an agent behaves in a way that appears to be immoral, we may ask him to explain or excuse the action; when the explanation or excuse is unsatisfactory, we may reprimand, ask for apology or restitution, punish, or isolate. Our treatment of suspected criminals is part of this practice, but it is a part that has, throughout the history of civilization, undergone continuous reform in response to challenges based on considerations of fairness, equality, and humanity. Such challenges are fueled by advances in moral reflection and in empirical science, and crucially, they form a core component of the practice of holding responsible.

Stephen Morse (2004, 2007) distinguishes between internal and external challenges to this practice, a distinction that stems from Ludwig Wittgenstein (1953) and is famously applied to the practice of holding morally responsible by P. F. Strawson (1962). An external challenge is based on considerations that are not features of the practice, or are at least general enough not to be specific to it. An internal challenge, by contrast, is based on considerations that are specific to the practice. Morse, like Strawson, characterizes the challenge that free will skeptics pose to the practice and to its criminological component in particular, as external, and rejects it for this reason; the skeptical view "generates only an external critique of responsibility that is unrelated to forensic practice and would obliterate the possibility of responsibility altogether" (Morse 2007: 104). I have argued, and will do so again in what follows, that this type of challenge is internal and thus cannot be dismissed for the type of reason Morse proposes (Pereboom 2001: 123). The considerations on the basis of which skepticism about free will

challenges the practice of holding responsible are in fact specific to the practice itself. Moreover, contrary to what Morse contends, the changes it aims to justify not only leave the practice intact but also can in fact be recommended on independent moral grounds.

A plausible example of an external challenge is David Hume's argument that our inductive practice is unjustified. Empirical induction, he points out, relies on the assumption that the future resembles the past, or, more broadly, that the unobserved resembles the observed, but we have no a priori or noncircular empirical justification for an assumption of this sort (1739/1978, 1748/2000). This skeptical challenge is external because is based on highly general standards for adequacy of justification that are not specific to inductive practice. Attempts to respond to Hume have not met with widespread acceptance, and thus the challenge remains in place. To my mind, opponents of external challenges have not made clear what, precisely, is defective about them. But we can add that in this example, no human being has ever opted out of the inductive practice despite the arguably unanswered external challenge, and Hume plausibly contended that it is psychologically impossible for us to do so. Thus, the skeptical challenge to induction is not only external, but also it is merely academic in the sense that it can have no serious effect on how we behave. One might similarly think that the free will skeptic's contention is merely academic in this sense. Strawson has made this, or a closely related, claim (1962). I will argue that this is not so, that the challenge from skepticism about free will can have an effect on our practice of holding responsible and, moreover, that this effect might well be salutary for general moral reasons.

Note first that the free will skeptic does not challenge every element of the practice of holding morally responsible. As a case in point, in everyday life, when we encounter

presumptively immoral behavior, we consider it legitimate to ask the agent, "Why did you decide to do that?" or, "Do you think it was the right thing to do?" If the reasons given in response to such questions are morally unsatisfactory, we regard it as justified to invite the agent to evaluate critically what her actions indicate about her intentions and character, to demand apology, or to request reform. Engaging in such interactions is reasonable in light of how they contribute to our understanding of our moral commitments and to our moral improvement. The moral responsibility invoked here has been called the *moral answerability* or the *fittingness of providing a moral explanation* sense, and it is the variety of moral responsibility that is most thoroughly ingrained in our practice and least controversial (Bok 1998; Fischer 2009; McKenna 2012; Arthur Kuflik, in conversation; Scanlon 1998). It may well characterize human interactions across all cultures. Our capacity to be sensitive to, to act on, and to evaluate behavior on the basis of moral reasons is essential to our being morally responsible in this way. The main thread of the historical free will debate does not pose determinism as a challenge to moral responsibility as answerability, and free will skeptics accept that we are morally responsible in this sense.

Rather, in this debate, skeptics, and incompatibilists more generally, argue that determinism is incompatible with moral responsibility—and the free will required for it—of a distinct kind. The variety of moral responsibility that incompatibilists claim not to be compatible with determinism is set apart by the notion of *basic desert*. For an agent to be morally responsible for an action in this sense is for it to be hers in such a way that she would deserve to be the recipient of an expression of moral resentment or indignation if she understood that it was morally wrong, and she would deserve be the recipient of an expression of gratitude or

praise if she understood that it was morally exemplary. The desert at issue here is basic because the agent, to be morally responsible, would deserve to be the recipient of the expression of such an attitude just because she has performed the action, given sensitivity to its moral status, and not, for example, by virtue of consequentialist or contractualist considerations. As I understand it, one might be the recipient of the expression of such an attitude when this expression is either covert or overt. Moral responsibility in this sense is presupposed by our attitudes of resentment, indignation, gratitude, and moral praise because having such an attitude essentially involves the supposition that the agent basically deserves to be the recipient of its expression, and it is thus the variety of moral responsibility that Strawson brings to the fore (1962).[2]

This sense of moral responsibility is backward looking in a distinctive way because it licenses retrospective assignments of basic desert. But moral responsibility in the basic desert sense does not have a monopoly on backward-looking considerations. The answerability sense on its own authorizes retrospective assessments, such as moral disapproval based on failure to act in accord with moral reasons, which do not invoke basic desert. This last notion can appropriately be regarded as a variety of blame, so one need not appeal to basic desert to secure at least one kind of blame.

Thus, it is not the whole practice of holding responsible—not the whole practice of blaming and praising—that is targeted by free will skepticism, but rather a component that relies on the assumption of basic desert. Likewise, it is not the entire practice of holding

---

[2] Benjamin Vilhauer (forthcoming) is a free will skeptic who endorses a theory of punishment founded on a contractualist theory of (nonbasic) desert.

criminals responsible that is placed under threat by this skeptical view, but only the justification of criminal punishment that depends on basic desert. The retributivist justification of punishment, according to which the punishment of a criminal is justified on the ground that he deserves it just because he has knowingly committed a serious offense, is the one that most intimately invokes the basic desert sense of moral responsibility, together with the freedom it demands. This is evident in Immanuel Kant's retributivist account, and also in those of Morse and Michael Moore (Kant 1797/1963; Moore 1987, 1998; Morse 2004). The connection is simply this: a judgment that an agent is blameworthy in the basic desert sense for a serious offense involves the supposition that he also deserves, in the basic sense, to be punished.[3] As we shall see, however, the main alternatives to retributivism for justifying criminal punishment either rely on retributivism at crucial junctures, and are thus also subject to the free will skeptic's challenge, or else are vulnerable to serious moral objections of other kinds. These other objections are driven by features of the practice of holding morally responsible, and thus

[3] Thanks to Dana Nelkin for discussion of this point. For minor offenses, perhaps a judgment that an agent is blameworthy in the basic desert sense also involves the supposition that he also deserves, in the basic sense, to be punished because here judging blameworthy involves supposing that the agent deserves, in the basic sense, to be the recipient of the expression of resentment or indignation, and this can count as punishment because targeting someone as a such a recipient involves an intention to inflict psychological pain, which therefore counts as punishment. See Wallace (1994: 51-83) and Nelkin (2011: 31-50) for discussions of related issues.

elements of the practice itself already threaten to dislodge these attempts to justify punishment.

There is an intuitively legitimate theory for prevention of dangerous crime that is undercut neither by free will skepticism nor by other moral considerations. This theory draws an analogy between treatment of dangerous criminals and treatment of carriers of dangerous diseases (Schoeman 1979). The free will skeptic claims that criminals are not morally responsible for their actions in the basic desert sense. Plainly, many carriers of dangerous diseases are not responsible in this or in any sense for have contracted these diseases. We generally agree that it is sometimes permissible to quarantine them nevertheless. But then, even if a dangerous criminal is not morally responsible for his crimes in the basic desert sense (say, because no one is ever in this way morally responsible), it could be as legitimate to detain him as to quarantine the nonresponsible carrier of a serious communicable disease. Furthermore, just as less dangerous diseases may justify only preventative measures less restrictive than quarantine, so, too, moderately serious criminal tendencies might permit only more moderate restraints. In addition, an account based on this analogy would demand a degree of concern for the rehabilitation and well-being of the criminal that would alter much of current practice. For just as fairness dictates that we must seek to cure the diseased we quarantine, so, too, fairness would require that we attempt to rehabilitate the criminals we detain. And if a criminal cannot be rehabilitated, and our safety requires his indefinite confinement, this account theory provides no justification for making his life more miserable than needed to guard against the danger he poses. Finally, there are measures for preventing

crime more generally, such as providing for adequate education and mental health care, that the free will skeptic can readily endorse.

This is the account I defend (Pereboom 2001, chapter 6). Although it is revisionist relative to widespread and historical punitive practice, similar views have frequently been advocated on independent grounds (e.g., Boonin 2008; Menninger 1968; Montague 1995), and it is sufficiently continuous with traditional and current policies to count as part of our general moral practice of holding morally responsible. Let me add that Morse is right to argue that free will skepticism should not be raised as a relevant issue in specific criminal cases and thus should not serve to alter this component of the practice (Morse 2007). Rather, this position is relevant to which general sorts of treatment of criminals can be justified, and thus for which laws concerning criminal punishment, detention, and rehabilitation are legitimate.

PROBLEMS FOR LIBERTARIANISM

Advocates of the claim that we do have the sort of free will required for moral responsibility in the basic desert sense are either proponents of libertarianism, the incompatibilist conception of this sort of free will, or of a compatibilist view. Recent times have witnessed the explicit differentiation of two widely held versions of libertarianism: the event-causal and the agent-causal types. In event-causal libertarianism, actions are caused solely by way of events, standardly conceived as objects having properties at times, and some type of indeterminacy in the production of actions by appropriate events is held to be a decisive requirement for moral responsibility (Balaguer 2010 2009; Ekstrom 2000; Kane 1996). This position has an ancestor in the Epicurean view according to which a free decision is an uncaused swerve in the otherwise downward path of an atom (Lucretius 50 B.C.E.). According to agent-causal libertarianism, free

will of the sort required for moral responsibility is accounted for by the existence of agents who possess a causal power to make choices without being determined to do so (Chisholm 1964, 1976; Clarke 1993; 2003; Griffiths 2010; Kant 1781/1787/1987; O'Connor 2000; Reid 1788/1983; Taylor 1966, 1974). In this view, it is essential that the causation involved in an agent's making a free choice is not reducible to causation among events involving the agent, but is rather irreducibly an instance of the agent-as-substance causing a choice not by way of events. The agent, fundamentally as a substance, has the causal power to cause choices without being determined to do so.

Robert Kane (1996) has developed a much-discussed version of event-causal libertarianism. In his proposal, the paradigm example of an action for which an agent is responsible is one of moral or prudential struggle, in which there are reasons for and against performing the action in question. The sequence that produces the action begins with the agent's character and motives and proceeds through the agent's making an effort of will to act, which results in the choice for a particular action. The effort of will is a struggle to choose in one way in a situation in which there are countervailing pressures. In the case of a freely willed choice, this effort of will is *indeterminate*, and as a result, the choice produced by the effort is *undetermined*. Kane explains this last specification by drawing an analogy between an effort of will and a quantum event (on one conception; Kane 1996, 128). The effort of will is indeterminate in the sense that its causal potential does not become determinate until the choice occurs. Before this pivotal interaction, there are different ways in which this causal potential can be resolved, and thus when it is resolved, the resulting choice will be undetermined. Significantly, Kane cautions against construing his view in such a way that the

indeterminacy occurs after the effort is made: "One must think of the effort and the indeterminism as fused; the effort is indeterminate and the indeterminism is a property of the effort, not something that occurs before or after the effort." If an agent is morally responsible for a choice, it must either be free in this sense or there must be some such free choice that is (or has a key role in) its sufficient ground, cause, or explanation (1999, 232). Kane embellishes his account by endeavoring to show how the particle analogy for free choice might actually work in the functioning of the brain's neural networks (1996, 128-130).

Critics of libertarianism have argued that if actions are undetermined, agents cannot be morally responsible for them. A classical presentation of this objection is found in Hume's *Treatise of Human Nature* (Hume 1739/1978, 411-412; cf. Mele 2006). In Hume's version, the concern highlighted is that if an action is uncaused, it will not have sufficient connection with the agent for her to be morally responsible for it. This idea might profitably be explicated as follows. For an agent to be morally responsible for a decision, she must exercise a certain type and degree of control in making that decision. In an event-causal libertarian picture, the relevant causal conditions antecedent to a decision—agent-involving events—would leave it open whether this decision will occur, and the agent has no further causal role in determining whether it does. With the causal role of these antecedent conditions already given, it remains open whether the decision occurs, and whether it does is not settled by anything about the agent. So whether the decision occurs or not is in this sense a matter of luck, and, intuitively, the agent lacks the control required for taking moral responsibility for the decision.

Libertarians agree that an action's resulting from a deterministic sequence of causes that traces back to factors beyond the agent's control would rule out her moral responsibility

for it. The deeper point of this objection is that if this sort of causal determination rules out moral responsibility, then it is no remedy simply to provide slack in the causal net by making the causal history of actions indeterministic. Such a move would yield a requirement for moral responsibility—the absence of causal determinism for decision and action—but it would not supply another—sufficiently enhanced control (Clarke 1997, 2003). In particular, it would not provide the capacity for an agent to be the source of one's decisions and actions that, according to many incompatibilists, is unavailable in a deterministic framework.

The agent-causal libertarian's solution to this problem is to specify a way in which the agent could have this enhanced control, which involves the power to settle which of the antecedently possible decisions actually occurs. The proposed solution is to reintroduce the agent as a cause, this time not merely as involved in events, but rather fundamentally as a substance (for detractors, see Haji 1998 Mele 2006). The agent-causal libertarian maintains that we possess a distinctive causal power—a power for an agent, fundamentally as a substance, to cause a decision without being causally determined to do so. This position might well be incipient in the medieval originators of modern libertarianism, but the Humean objection to indeterministic free will occasioned a more precise formulation. Thomas Reid provides an agent-causal account in direct response to Hume (1788/1983), and arguably, Kant does as well in his account of transcendental freedom (Clarke 2003; Chisholm 1964, 1976; Griffiths 2010; Kant 1781/1787/1987, cf. Watkins 2005; O'Connor 2000; Reid 1788/1983).

One traditional objection to the agent-causal picture is that we have no evidence that we are substances of the requisite sort. Kant expresses two further concerns for the agent-causal view, which in his account calls for our endorsement on practical, but not on evidential,

grounds. One is that despite our inability to discern an internal incoherence in our conception of transcendental freedom, our having this power might nonetheless not be really possible. In Kant's view, we can discern something to be really possible only through experience, and we do not experience ourselves as agent causes. The second concern is that agent causation might not be reconcilable with what we would expect given our best empirical theories. Kant himself believed that the physical world, as part of the world of appearance, is governed by deterministic laws, whereas the transcendentally free agent-cause would exist not as an appearance, but as a thing in itself. In this agent-causal picture, when an agent makes a free decision, she causes the decision without being causally determined to do so. On the path to action that results from this undetermined decision, alterations in the physical world, for example, in her brain or some other part of her body, are produced. But it would seem that we would at this point encounter divergences from the deterministic laws. For the alterations in the physical world that result from the undetermined decision would themselves not be causally determined, and they would thus not be governed by deterministic laws. One might object that it is possible that the physical alterations that result from every free decision just happen to dovetail with what could in principle be predicted on the basis of the deterministic laws, so nothing actually occurs that diverges from these laws. However, this proposal would seem to involve coincidences too wild to be credible. For this reason, it seems that agent-causal libertarianism is not reconcilable with the physical world's being governed by deterministic laws. Kant thought that more needed to be said. In one text, he appears to claim that when an agent makes a transcendentally free decision, he, as atemporal noumenal subject, also freely produces everything in the past that causally determines his free actions (1788/1996, Ak 97-

98). But this sort of atemporalist line is at best insignificantly more credible than an overt contradiction (Pereboom 2006; Wood 1984).

More recent expositors of the agent-causal view, such as Randolph Clarke (1993, 2003) and Timothy O'Connor (2000, 2009), suggest that quantum indeterminacy can help with the reconciliation project. On one interpretation of quantum mechanics, the physical world is not in fact deterministic, but is rather governed by laws that are fundamentally merely probabilistic or statistical. Suppose, as is controversial, that significant quantum indeterminacy percolates up to the level of human action. Then it might seem that agent-causal libertarianism can be reconciled with the claim that the laws of physics govern at least the physical components of human actions. However, it appears that wild coincidences would also arise on this suggestion (Pereboom 1995, 2001). Consider the class of possible actions, each of which has a physical component whose antecedent probability of occurring is approximately 0.32. It would not violate the statistical laws in the sense of being logically incompatible with them if, for a large number of instances, the physical components in this class were not actually realized close to 32% of the time. Rather, the force of the statistical law is that for a large number of instances, it is correct to *expect* physical components in this class to be realized close to 32% of the time. Are free choices on the agent-causal libertarian model compatible with what the statistical law would lead us to expect about them? If they were, then for a large enough number of instances, the possible actions in our class would almost certainly be freely chosen nearly 32% of the time. But if the occurrence of these physical components were settled by the choices of agent-causes, then their actually being chosen close to 32% of the time would also amount to a wild coincidence. The proposal that agent-caused free choices do not diverge from what the

statistical laws predict for the physical components of our actions would be so sharply opposed to what we would expect as to make it incredible.

At this point, the agent-causal libertarian might propose that exercises of agent-causal libertarian freedom do result in divergences from what we would expect given our best assessments of the physical laws. Roderick Chisholm (1964) proposes such a view. Divergences from the probabilities that we would expect absent agent-causes do in fact occur whenever we act freely, and these divergences are located at the interface between the agent-cause and that part of the physical world that it directly affects, likely to be found in the brain. There are different ways in which agent-caused free choices might diverge from what the laws would predict. One way is by not being subject to laws at all. Another is by being subject to different statistical laws, an option on which the agent-cause would be governed by probabilistic laws that are its own because they emerge only in the right sorts of agential contexts, and not to those that generally govern the physical events of the universe (O'Connor 2008). A concern for these kinds of claims is that we currently have no evidence that they are true.

THE FREE WILL SKEPTIC AND COMPATIBILISM

We human beings typically assume that we can be, and often are, morally responsible for our actions in the basic desert sense, and this assumption is manifest in the uncompromised sense of legitimacy that often accompanies the basic desert involving reactive attitudes. But in everyday life, we seldom, if ever, bring to bear on this assumption any theory about the general causal nature of the universe that might threaten its rationality, even if we know that such a theory is likely to be true. For example, in ordinary life, most of us do not question the rationality of this assumption on the basis of the theory that every event, including our choices

and actions, results from deterministic causal processes that trace back to a time before we existed; or on the basis of the theory that if this sort of determinism is false, only event-causal indeterminism is likely to be true. Spinoza held that the fact that the complete causal story of our actions is hidden from us accounts for our belief that we are free; "men think themselves free, because they are conscious of their volitions and their appetite, and do not think, even in their dreams, of the causes by which they are disposed to wanting and willing, because they are ignorant of [those causes]" (Spinoza 1677/1985, 440). But he contends that full rationality demands a change in view, the rational integration of the justified claim that all of our actions are deterministically caused requires that we relinquish the belief that we are free in the sense at issue. By contrast, compatibilists, such as Morse, Strawson, and Fischer, hold that the rational response is to retain our everyday assumption that we are free in the sense required for basic desert, together with the attitudes and practices that are founded on this assumption.

How might this disagreement between the free will skeptic (and incompatibilists more generally) and the compatibilist be mediated? I think that the best way for the skeptic to try to do so begins with the intuition that if an agent is causally determined to act by, for example, scientists who manipulate her brain, then she is not morally responsible for that action, even if she meets compatibilist conditions on moral responsibility (Ginet 1990; Kane 1996; Mele 1995, 2006; Pereboom 1995, 2001; Taylor 1974; van Inwagen 1983). The subsequent step is that there are no differences between the manipulated agents and their ordinary deterministic counterparts that can justify the claim that the manipulated agents are not morally responsible while the determined agents are.

My multiple-case version of such an argument first of all develops examples of an action that results from an appropriate sort of manipulation and in which the prominent compatibilist conditions on moral responsibility are satisfied (Pereboom 1995, 2001, 2011). In the setup, in each of the four cases the agent commits a crime, say a homicide, for the sake of some personal advantage. The cases are designed so that the crime conforms to the prominent compatibilist conditions. This action meets certain conditions advocated by Hume: the action is not out of character because for the agent, it is generally true that selfish reasons typically weigh heavily—too heavily when considered from the moral point of view; and in addition, the desire that motivates him to act is nevertheless not irresistible for him, and in this sense he is not constrained to act (Hume 1739/1978). The action fits the condition proposed by Harry Frankfurt (1971), that is, his effective desire (i.e., his will) to commit the crime conforms appropriately to his second-order desires for which effective desires he will have. That is, he wills to kill the victim, and he wants to will to do so, and he wills this act of murder because he wants to will to do so. The action also satisfies the reasons-responsiveness condition advocated by John Fischer and Mark Ravizza (1998): our agent's desires can be modified by, and some of them arise from, his rational consideration of the reasons he has, and if he knew that the bad consequences for himself that would result from the crime would be much more severe than they are actually likely to be, he would have refrained from committing the crime for that reason. This action meets the condition advanced by Jay Wallace (1994): the agent has the general ability to grasp, apply, and regulate his actions by moral reasons. For instance, when egoistic reasons that count against acting morally are weak, he will typically regulate his behavior by moral reasons instead (further advocates of views that privilege reasons and

rationality include Bok 1998; Nelkin 2008, 2011; and Wolf 1990). This ability also provides him with the capacity to revise and develop his moral character over time, a condition that Alfred Mele emphasizes (1995, 2006).

These manipulation examples, taken separately, indicate that it is possible for an agent not to be morally responsible in the basic desert sense even if the compatibilist conditions are met and that, as a result, these conditions are insufficient. But the argument has additional force by virtue of setting out three such cases, each of which is progressively more like a fourth, in which the action is causally determined in a natural way. The first case involves manipulation that is local and determining, and hence very likely to elicit the nonresponsibility intuition. The second is like the first, except it restricts manipulation to a remote position at the beginning of the agent's life. The third is similar, except the manipulation results from strict community upbringing; and the fourth, again, is the ordinary deterministic case. The aim is to formulate these examples so that it is not possible to draw a principled line between any two adjacent cases that would explain why the agent would not be morally responsible in the basic desert sense in the first but would be in the second. The conclusion is that the agent is not morally responsible in this sense in all four cases, and the best explanation for this must be that he is causally determined by factors beyond his control in each, and this result conflicts with the compatibilist's central claim.

Strawson contends that features of the practice of holding morally responsible insulate attributions of moral responsibility from scientific or metaphysical challenges such as the one based on causal determinism (1962). And Morse, although he acknowledges that the skeptical view "can produce an internally coherent, forward-looking consequential system that treats

human action specially and that might possibly encourage good behavior and discourage bad," claims that it "cannot explain or justify our present blame and punishment practices, which are essentially retrospectively evaluative," and in particular, it "cannot explain or justify the relation between action and desert" (cf. Alexander, Ferzan, and Morse 2009: 13-15; Morse 2004: 433). In the last analysis, these compatibilists want to claim that free will skepticism is to be rejected because it does not cohere with our practice of holding responsible. In my view, the best sort of argument against this position involves what Wallace calls a generalization strategy—arguing from generally accepted excuses or exemptions to the conclusion that causal determinism rules out moral responsibility (Wallace 1994). The excuses and exemptions that form the basis of this sort of argument would have to be generally accepted (but perhaps not uncontested) so that they are plausibly features internal to the practice of holding people morally responsible. The kinds of exemptions that I exploit in the four-case argument are due to deterministic manipulation, and it is a feature of our practice that we exempt agents from moral responsibility in the basic desert sense when they are manipulated in this way, as in the local and remote manipulation examples. It is also a feature of our practice that if no morally relevant difference can be found between agents in two situations, then if one agent is legitimately exempted from moral responsibility in this sense, so is the other. No morally relevant difference can be found between agents in the manipulation examples and agents in ordinary deterministic situations such as the final case. Thus, it is the practice itself—in particular, key rules governing the practice—that renders moral responsibility in the sense at issue vulnerable to causal determinism. Each of these rules is a component of our notion of

fairness, and accordingly I dispute Morse's claim that "we currently adhere to no principle of fairness that is related to the truth or falsity of determinism" (2004: 442).

Morse follows Fischer and Wallace in making responsiveness to reasons, and to moral reasons in particular, the core requirement for moral responsibility (Morse 2004). But what it is about reasons-responsive action that justifies the ascription of basic desert and, as Fischer and Wallace maintain, justifies the reactive attitudes that presuppose this notion of responsibility (Fischer 2007; Wallace 1994 2004)? Although it may be natural for us to respond to reasons-responsive actions with reactive attitudes, and for us to refrain from or withdraw such responses when actions fail to meet this standard, it is a further task to show that the reasons responsiveness of an action legitimates the attribution of basic desert moral responsibility and the attitudes that presuppose it. By contrast, the "answerability" or "fittingness of providing an explanation" notion of moral responsibility, which the free will skeptic can endorse, is transparently connected to the notion of reasons responsiveness.

RETRIBUTIVISM

According to the retributivist position, punishment of a wrongdoer is justified for the reason that he deserves something bad to happen to him—pain, deprivation, or death, for example—just because he has done wrong (Kant 1797/1963; Moore 1987, 1998). Hence, a wrongdoer's deserving to be harmed is not reducible to a component of a scheme justified solely on the basis of its consequences. This claim is typically subjected to qualifications such as that the agent had to have committed the wrong knowingly. But what is crucial to the free will skeptic's challenge is that according to the retributivist, it is the basic desert attached to the criminal's immoral action alone that provides the justification for punishment. The retribution theory

does not appeal to a good such as the safety of society, or the moral improvement of the criminal in justifying punishment. Rather, the good by means of which retributivism justifies punishment, or the principle of right action that justifies punishment, is that an agent receive what he deserves just because of his having knowingly done wrong.

This position would be undermined if the free will skeptic were right because if agents do not deserve blame just because they have knowingly done wrong, neither do they deserve punishment just because they have knowingly done wrong. Retributivism justifies punishment solely on the grounds of basic desert, and the skeptical position is incompatible with retributivism for the reason that it claims this notion does not apply to us. Free will skepticism thus recommends that the retributivist justification for punishment be abandoned.[4]

Morse argues that libertarianism is not plausible, and of the remaining options, we should reject free will skepticism and accept compatibilism because of these only compatibilism is consonant with our practice. Significantly, he does so without offering replies to the objections that have been raised against this position. Morse is right to assert that given the rejection of libertarianism, only compatibilism would underwrite the uncompromised legitimacy of the reactive attitudes and retributivistically justified punishment. But can

---

[4] Erin Kelly (2009: 446) expresses a similar idea without endorsing free will skepticism: "Retributivism, as I understand it, is the view that justice requires the punishment of criminal wrongdoers, apart from the (further) social benefits a system of punishment might bring. The case for this notion of justice is built on reactive attitudes that presuppose a wrongdoer's moral capacity to have acted as morality demands. If we drop the assumption that offenders always have this capacity, we must reevaluate the aims of punishment."

retributivism be adequately grounded in compatibilism in this way absent satisfying responses to the objections that compatibilism faces? Consider, for example, the last murderer remaining in prison in Kant's imagined island society that is about to dissolve itself (in the Metaphysical Elements of Justice, a section of Kant's *Metaphysics of Morals*). Kant strenuously advocates that he should be executed, just because of the crime he has committed, that is, for reasons of retributive desert alone (1797/1963, Ak VI 331-333). But imagine the offender protesting that he was determined by natural causes to act as he did, and this plausibly undercuts the claim that he is morally responsible in the basic desert sense. Would the following reply count as morally acceptable? "Our practice of punishing criminals for retributivist reasons requires the assumption that we are morally responsible in the basic desert sense, and we need to believe that our being free in the way required for our having this kind of responsibility is compatible with determinism to in order to rationally maintain this assumption." It would not. If the retributivist justification of punishment featured by our actual practice requires the rationality of the belief that compatibilism is true, while at the same time there are serious and unanswered objections to this position, we cannot legitimately respond to a challenge to this part of the practice just by saying that it is supported by compatibilism. Punishment inflicts harm, and in general, justification for harm must meet a high epistemic standard. If it is significantly probable that one's justification for harming another is unsound, then prima facie that behavior is seriously wrong, and one must refrain from engaging in it. A strong and credible response to the objections to compatibilism is required to meet this standard.

One might suppose that setting aside the free will skeptic's challenge, retributivism would provide a resilient and powerful justification for punishment. However, there are

substantial arguments for the claim that retributivism turns out to be unacceptable even disregarding the skeptic's considerations (e.g., Braithwaite and Pettit, 1990, 156-201; Montague 1995, 11-23, 80-90; Ten 1990, 38-65). Perhaps the deepest problem for this theory derives from the hypothesis that the intuitions that drive the retributivist position are at root fueled by vengeful desires, and that therefore retribution has little more plausibility than vengeance as a morally sound policy for action (Moore 1987). Acting on vengeful desires might be wrong for the following sort of reason. Although acting on such desires can bring about pleasure or satisfaction, no more of a moral case can be made for of acting on them than can be made for acting on sadistic desires, for example. Acting on sadistic desires can bring about pleasure, but in both cases, acting on the desire aims at the harm of the one to whom the action is directed, and in neither case does acting on the desire essentially aim at any good other than the pleasure of its satisfaction. But then, according to the objection, because retributivist intuitions have their source in vengeful desires, acting for the sake of retribution is also morally wrong.

One counterstrategy involves pointing out salient differences between vengeance and retribution. For example, by contrast with vengeance, retribution is in principle limited in its severity, and although vengeance often engenders further vengeance, retribution brings about closure. But one might reply that because what fuels retributivist intuitions are vengeful desires, retributivistically justified punishment would at root amount to a form of controlled vengeance. Another retributivist response is that in central cases, the sentiment of vengeance is an emotional expression of the sense of retributivist justice.[5] In such cases, the sense of retributivist justice is explanatorily before the sentiment of vengeance, and thus if retribution

[5] George Sher once made this suggestion in conversation.

has an independent justification, then it might well not be threatened by the objection from vengeance. But these responses on behalf of the retributivist are not sufficiently convincing for the view to attain the high epistemic standard that a theory of punishment must meet.

In addition, it is not clear that there are strong positive reasons for maintaining that retributivist justification for punishment reflects a genuine moral good, or, on a view in which the right is prior to the good, that it reflects an authentic principle of right. When Kant introduces his retributivist justification of punishment in the Metaphysical Elements of Justice (1797/1963), no apparent link is forged with the comprehensive moral theory expressed in the various formulations of the categorical imperative. Does treating humanity in persons as an end in itself require that we consider each other as subject to punishment retributivistically justified? Will legislation for a kingdom of ends by a community of rational beings invoke retributivist justification? Far from obviously so. Add to this the fact that purely consequentialist or contractualist views will not accommodate this sort of justification and at best can only secure a weaker facsimile.

Lastly, supposing that the requisite capacity for control is in place and that basic desert could be secured as good or right, we can ask whether the state has the right to invoke it in justifying punishment. The legitimate functions of the state are generally held to include protecting its citizens from significant harm and providing a framework for human interaction to proceed without significant impairment. These roles arguably underwrite justification that in the first instance appeals to prevention of crime. But these roles have no immediate connection to the aim of apportioning punishment in accord with basic desert. The concern can be made vivid by considering the proposal that the state set up a well-funded set of institutions designed

to comprehensively and fairly distribute rewards on the grounds of basic desert (cf. Husak 2000: 973-974). Political theories that would demand that the state provide such institutions are not widely held. In addition, as Douglas Husak (2000: 974-975) argues against Moore (1998), a state institution designed to give criminals what they deserve is extremely expensive, its benefits do not obviously outweigh its costs, and these resources might be spent on more valuable programs. And as human history demonstrates, such an institution is subject to grave and frequent error or abuse, which gives rise to a tremendous amount of unmerited suffering.

If the skeptic's challenge were a good one, then the retributivist justification for punishment would be undermined. But there are independent moral and political grounds for rejecting a retributivist theory of punishment. These other arguments against retributivism adduce moral and political considerations uncontroversially internal to the practice of holding morally responsible. We have already seen that despite initial appearances, close examination indicates that the considerations on which the free will skeptic's case against compatibilism is based are internal to the practice. Now it is also clear that the justification of punishment that the skeptic's challenge targets was already under threat by arguments that are uncontroversially internal. Thus, in taking aim at retributivism, the free will skeptic is not opposing a view that is otherwise free from uncontroversially internal opposition.

MORAL EDUCATION THEORY

There are further ways of justifying criminal punishment that do not appeal to a notion of basic desert, and thus they may not run afoul of free will skepticism specifically. Moral education theories typically draw an analogy to the justification of the punishment of children. Children are not typically punished to exact retribution, but rather to educate them morally. Because

moral education is a generally acceptable goal, a justification for criminal punishment based on this analogy is one the free will skeptic can potentially accept.

In developing their views, advocates of the moral education theory have proposed specific ways in which punishment or threat of punishment could serve to educate a child. Herbert Morris argues that punishment educates a child morally by apprising her of the consequences of her wrongdoing for herself and for others. To use his example, if a child cheats, a parent might exclude her from playing the game for a time, thereby making vivid to her the possible game-undermining consequences of cheating for both herself and others (Morris 1981: 160). Similarly, Jean Hampton proposes that a wrongdoer might be "made to endure an unpleasant experience designed, in some sense, to "represent" the pain suffered by her victim(s)" (Hampton 1984: 227). Hampton also contends that punishment can morally educate a child by conveying to him the degree of seriousness of the wrongdoing (1984: 225-226). Different levels of severity can communicate distinct levels of seriousness of wrongdoing and, and partly in consequence of this, they can convey important moral boundaries. A further way, which Morris also points out, is that punishment can indicate the strength of the parents' attachment to the moral rules that have been breached. Finally, one might argue that coercing a child into behaving in accord with morality could serve to acquaint him with the benefits of moral virtue, which he might subsequently come to value for its own sake.

A serious concern for this type of theory is that it is far from evident that punishing adult criminals is likely to result in these sorts of moral improvement. Children and adult criminals differ in several salient respects. Adult criminals, unlike children, typically understand the moral code accepted in their society. Contrary to Hampton's view in particular, one could not justify

punishment on the ground that it would convey to these criminals that their actions are morally wrong, and to communicate to them "that there is a barrier of a very special sort against these kinds of actions." There are indeed adult criminals who do not comprehend that their actions are morally wrong, but we are disposed not to punish them for reasons of incompetence. In fact, in many contemporary jurisdictions, it is specifically when a criminal does not know that his actions are morally wrong that he is judged not liable to punishment.

One might instead claim that punishment is nonetheless likely to significantly motivate criminals to improve morally. But relevant differences between children and adult criminals also threaten this proposal. Children are generally more psychologically malleable than adult criminals. For a more contingent difference, punishment that has the best prospect for morally educating children would transpire in a caring environment. Punishment outside of such a context is much more likely to engender resentful attitudes and behavior rather than motivation to moral improvement. Here, the analogy between children and adult criminals is again weak in salient respects, and as a consequence, the claim that punishment can result in such motivation in the case of criminals is insufficiently plausible.

Because criminals differ in significant ways respects from the sorts of agents for whom the success of punishment in producing moral education might be reasonably thought to have been established, strong empirical evidence that criminals could be similarly educated would be required. Without such evidence, it would be morally wrong to punish criminals for the reason that it can attain such an outcome. In general, if one proposes to harm someone to achieve a salutary result, one must have strong evidence that harming her can have this effect.

Hampton proposes specific ways in which punishment might realize the moral education of a criminal:

> One way the moral education theorist can set punishments for crimes is to think about "fit." Irrespective of how severe a particular crime is, there will sometimes be a punishments that seems naturally suited to it; for example, giving a certain youth charged with burglarizing and stealing money from a neighbor's house the punishment of supervised compulsory service to this neighbor for a period of time, or giving a doctor charged with cheating a government medical insurance program the punishment of compulsory unremunerated service in a state medical institution. (Hampton 1984: 227-228)

But service requirements of the kind Hampton suggests are not paradigmatic kinds of punishment. One might reasonably question whether they should be classified as punishment at all, and not instead as programs for moral rehabilitation. Hampton might respond that these kinds of service include a punitive aspect—the restriction of a criminal's freedom (1984: 224). And such a program might often involve psychological discomfort or even anguish. Still, compulsory rehabilitative programs also involve restriction of freedom and will often feature psychological pain.

Moreover, what might it be about a service requirement that would plausibly yield moral improvement? Could it be the punitive aspect—the restriction of freedom or the pain of carrying out the service? More likely, what would produce salutary moral change is involvement with the people or the kinds of people the criminal has harmed. The restriction of freedom and the pain of service that such a program involves are best seen as a side effect of

the method, and not its goal or its means. This is a further reason for viewing such approaches as programs for moral rehabilitation rather than as cases in which punishment morally educates.

Suppose we had strong evidence that punishment can morally educate criminals. Even then, if there are nonpunitive methods for achieving comparable moral education, they should be preferred, all else being equal. If, for example, a criminal can be effectively morally educated through a humane rehabilitative program, then that method should be favored over a punitive alternative that would realize the same result. By analogy, if a neurophysiological problem such as insufficient serotonin production explains an agent's lack of moral motivation, then effective drug therapy should be preferred to a regimen of punishment that would achieve the same goals. All else being equal, if two methods achieve the same goal for an agent, but one harms an agent whereas the other does not, the one that does not harm the agent should be preferred.

DETERRENCE THEORIES

One objective that societies have in punishing criminals is to prevent them as well as other prospective criminals from committing crimes. On deterrence theories it is the prevention of criminal wrongdoing that serves as the good by means of which punishment is justified. Initially, it would seem that there is no feature of skepticism about free will that makes deterrence theories less acceptable to it than to libertarianism or to compatibilism. As we shall see, however, at least some deterrence theories are not clearly immune to the skeptic's challenge because they presuppose a retributive principle. Furthermore, like retributivism, deterrence

justifications of paradigmatic sorts of punishment face difficult objections that are independent of skepticism about free will.

The classic deterrence theory is Jeremy Bentham's. In his conception, the state's policy toward criminal behavior should aim at maximizing utility, and punishment should be administered if and only if it does so. The pain or unhappiness produced by punishment results from the restriction on freedom that ensues from the threat of punishment, the anticipation of punishment by the person who has been sentenced, the pain of actual punishment, and the sympathetic pain felt by others such as the friends and family of the criminal (Bentham 1823/1948). The most significant pleasure or happiness that results from punishment derives from the security of those who benefit from its capacity to deter both the criminal and other potential criminals. No feature of the free will skeptic's position, specifically, challenges this view.

Several more general objections, however, have been raised against the utilitarian deterrence theory (Montague 1995, 6-11; Ten 1987, 7-37). Three of these objections are especially threatening. The first is that this approach will justify punishments that are intuitively too severe. For it would seem that in certain cases, extremely harsh severe punishments would be more effective deterrents than much milder forms, whereas such punishments are intuitively too severe to be fair. For example, if society is threatened by a crime wave, administering penalties of this sort might well maximize utility. The utilitarian could reply that if we are careful to include the pain of punishment in the calculation, the resulting severity will typically or always be intuitively unacceptable. He might also claim that in certain uncommonly dangerous situations, extremely severe penalties might indeed be justified. But nevertheless,

one might reasonably fear that utilitarian recommendations will often fail to conform to our intuitions about fairness.

Second, the theory would seem to justify punishing the innocent (McCloskey 1965). If after a series of horrible crimes, the actual perpetrator is not caught, potential criminals might come to believe that they can get away with serious wrongdoing. Under such circumstances, it might maximize utility to frame and punish an innocent person. Utilitarians might reply that the probability of such a scheme being discovered is always significant and that, as a result, punishing the innocent is unlikely to maximize utility in any situation. However, it is far from obvious that this response is convincing. John Rawls offers a different response on behalf of the utilitarian. There exist good utilitarian reasons for a punishment policy to be general, stable, and public, in short, to be institutionalized. Consequently, Given this fact, there will be solid utilitarian reasons against deceptively punishing the innocent (Rawls 1955).[6] As a matter of

---

[6] Rawls writes about an institution, "telishment," that would allow for punishing the innocent. Once one realizes that one is involved in setting up an *institution*, one sees that the hazards are very great. For example, what check is there on the officials? How is one to tell whether their actions are authorized? How is one to limit the risks involved in allowing such systematic deception? How is one to avoid giving anything short of complete discretion to the authorities to telish anyone they like? In addition to these considerations, it is obvious that people will come to have a very different attitude toward their penal system when telishment is adjoined to it. They will be uncertain as to whether a convicted man has been punished or telished. They will wonder whether or not they should feel sorry for him. They will wonder whether the same fate won't at any time fall on them. If one pictures how such an institution would actually work,

,practical fact, it is doubtful that a general, stable, and public scheme that would allow deceptively punishing the innocent could achieve the envisioned maximization of utility. RatherIn fact, an institution of this sort could easily engender massive disutility in society. A worry about Rawls's reply is that this practice would seem to be more deeply wrong than can be accounted for by the utilitarian reasons he presents.

A further serious misgiving raised against utilitarian deterrence theory is the "use" objection. A general problem for utilitarianism is that it allows people to be harmed severely, without their consent, in order to benefit others, and this is often intuitively wrong. Punishing criminals for the security of society would appear to be just such a practice. Even if this problem fails to undermine utilitarian deterrence theory decisively, it should challenge one's confidence in this approach. Again, in assessing justifications for punishment, it is crucial that for a theory to be legitimately applicable in practice, we must be reasonably confident that it can withstand the objections that have been raised against it. Criminal punishment involves treating people severely—often it has very harmful short- and long-term consequences for the person being punished. If we were only mildly confident about the justification for such punishment, it would be morally wrong to administer it. Thus, we have solid reasons not to employ the classic utilitarian deterrence theory in justifying actual punishment policy—whether or not the free will skeptic is right.

At this point, one might combine retributivism and utilitarianism and argue that there is a retributivist justification for the legitimacy of using criminals for the good of society more

---

and the enormous risks involved in it, it seems clear that it would serve no useful purpose. A utilitarian justification for this institution is most unlikely (Rawls 1955, 3-32).

generally. I suspect that many who think that the main point of punishment is deterrence are relying on a retributivist assumption of this sort. But if the skeptic is right, such a view is unavailable, and, as we have seen, there are other moral and political reasons to question retributivism.

PUNISHMENT JUSTIFIED BY THE RIGHT TO HARM IN SELF-DEFENSE

In my view, the free will skeptic should invoke the right to self defense and defense of others in justifying its policy for treatment of criminals. I don't think that the skeptic should aim to justify punishment by appealing to this right, but rather, for the most dangerous criminals, should aim for a policy of detention modeled on quarantine. Several theorists, however, have argued that it is indeed punishment that might be justified on such grounds. Daniel Farrell offers one of the finest developments of a theory of this sort (Farrell 1985; Kelly 2009; cf. Quinn 1985). Farrell's theory is impressive if only because it justifies punishment on grounds that are widely accepted.[7] Because the free will skeptic can also endorse the right to harm in self-defense and defense of others, even she may aspire to accepting a justification of punishment of the kind

[7] Warren Quinn's theory is similar, but differs from Farrell's in the following way (Quinn 1985). Although Farrell contends that it is legitimate to threaten to harm an aggressor in certain circumstances because one may harm him in those circumstances, Quinn argues that one may harm him just because one may threaten to harm him. Accordingly, Quinn aims to establish that punishing criminals is legitimate because threatening to punish them is legitimate. Kelly (2009: 447-448) endorses Farrell's justification.

that Farrell develops. But as we shall see, again, there are reasons independent of the skeptic's position to doubt the soundness of this type of view.

Farrell's account highlights the distinction between special deterrence—punishment aimed at preventing the criminal from engaging in criminal behavior, and general deterrence—punishment aimed at preventing agents other than the targeted criminal from doing so. In his view, special deterrence is significantly easier to ground in the right to harm in self-defense or defense of others than is general deterrence. Farrell also distinguishes between the right of *direct* self-defense, your right to harm an unjust aggressor to prevent him from harming you or someone else, and the right of *indirect* self-defense, your right to threaten an unjust aggressor with a reasonable amount of harm to prevent him from harming you or someone else.

In broad outline, Farrell's justification of punishment as special deterrence is this. Each of us has the right of direct self-defense, and each of us also has the right of indirect self-defense. Because we have the right of direct self-defense, we have the right to inflict a reasonable amount of harm on a potential unjust aggressor to prevent him from harming us. Because we have the right of indirect self-defense, we also have the right to threaten to inflict this amount of harm. Our right of direct self-defense permits us carry out this threat against him once the condition has been violated. But also, because we have these rights, the state, acting as proxy for us, may issue appropriate general threats to harm unjust aggressors, and may also carry out such threats once their conditions have been violated. In this way, the right to self-defense can ground a legitimate state institution of punishment as special deterrence.

This special deterrence theory avoids some of the key objections to its utilitarian counterpart. On the concern for justifying punishment that is intuitively too severe, one may

not, on grounds of indirect self-defense, issue a threat to inflict harm more severe than the minimum required to effectively deter the target crime. So, if a threat of 1 year in prison would be sufficient to deter auto theft, the state may not issue a threat of a 10-year term. On the concern for punishing the innocent, the right to self-defense justifies harming only unjust aggressors themselves. For instance, the right does not justify harming an unjust aggressor's innocent children even if this would deter him.

At the same time, harming an unjust aggressor in self-defense or defense of others does involve harming him, without his consent, for the benefit of persons other than himself, and this arguably would count as an instance of using him as a means to the benefit of others. Perhaps this is a legitimate kind of use because its target brings it upon himself by his unjust aggression. But this suggestion proposal might seem to invoke the notion of basic desert. Significantly, Farrell argues that the right of self-defense assumes a type of retributivism, albeit a weak form. In his view, underlying the right to direct and indirect self-defense, and thus also special deterrence on his account, is a "weakly retributive" principle of distributive justice:

> If an aggressor forces one to make a choice between harming the aggressor or allowing him/herself or others to be harmed, then one may harm the aggressor to the degree that preventing the harm to oneself or others requires (Farrell 1985).

(Farrell plausibly contends that this principle applies only within bounds—if the harm threatened is minor, but killing the aggressor would be required to prevent it, then killing the aggressor is wrong.) If this principle is in fact retributive, and thereby presupposes basic desert, and it does indeed underlie the right to self-defense, it would appear that the arguments for

the skeptical position imperil this right and a theory of punishment on which it is based. However, it is not correct, I think, to call this principle "retributive" if in doing so basic desert is invoked. For this principle and the right to harm in self-defense more generally very plausibly apply to aggressors who are not morally responsible in the basic desert sense, such as people who have been brainwashed, those who are significantly mentally impaired, and animals.

Now the “use” problem arises again—isn’t the aggressor who threatens to seriously injure me whom I then harm in self-defense being used merely as a means to secure my own safety? I think the answer is affirmative, but that in such cases, if the harm inflicted is the minimal amount reasonably required to prevent the serious injury, the force of the “use” objection is outweighed by the right to harm in self-defense. Farrell contends that the type of theory he proposes will not extend to full-fledged general deterrence, for this would involve harming someone to prevent not just his aggression but also the potential aggression of others, and this gives rise to a convincing "use" objection. Yet he argues that some general deterrence can be justified on the basis of his principle of distributive justice. When an agent wrongs you in such a way as to make you more vulnerable than you would otherwise be to the aggression of others, then you are justified in countering just this degree of added vulnerability by harming him. My sense is that, in such cases, the force of the “use” concern is plausibly outweighed by the right to harm in self-defense, by contrast with a practice of full-fledged general deterrence.

Thus, so far it seems that Farrell has proposed a justification for criminal punishment that the free will skeptic can endorse. And I do accept it in its essentials. The concern I have is whether for dangerous violent criminals, Farrell’s line of reasoning can justify punishment, by contrast with preventative detention. What makes it appear as if punishment can be justified in

this way is, I think, the model of an unjust aggressor in a situation in which state law enforcement and criminal justice agencies have no role—let's call it a "state of nature" situation. A state of nature situation in which an aggressor poses an immediate danger is very different from the circumstances of criminals in our society when state punishment is carried out. They are then in the custody of the law. And the kinds of harms that the right of self-defense justifies in the case of aggressors in the state of nature differ in kind from those that this right would justify for those in the custody of the law.

An analogy will help show why it is wrong to carry out a threat against a criminal in custody that would legitimately be carried out in circumstances in which he poses an immediate danger. Threats that can legitimately be carried out against an immediately dangerous potential aggressor specify what one would reasonably believe to be the minimum harm required to prevent his aggression. Suppose that someone clearly aims to kill you, and that to prevent his doing so you may knock him out with a baseball bat. It would then be permissible for you to threaten him with this amount of harm. Suppose he does attempt to kill you, but in the process he trips over the toys on the floor, and this allows you to pin him to the ground and tie him up. At this point is it still legitimate for you to knock him out with the bat? To do so would clearly be wrong, and not justified by the right to harm in self-defense. For this right justifies only what one would reasonably believe to be the minimum harm required to prevent what the aggressor threatens to inflict. Or suppose the aggressor clearly aims to kill your companion, and that to protect her it is legitimate for you to knock him out with the bat and threaten to do so. Suppose that despite your efforts, he kills her, but that subsequently he trips and you tie him up. Is it then legitimate for you to knock him out with the bat? Not on the

basis of the right to harm in self-defense and defense of others because he no longer poses an immediate threat. You retain the right to protect yourself and others against him, but not by carrying out the threat designed to prevent a harm that has already occurred.

So, then, a threat that one could justifiably make and carry out to protect against an aggressor in a state-of-nature situation cannot legitimately be carried out in a situation in which the aggressor is in custody. The reason for this is clear. The minimum harm required to protect ourselves from someone who is immediately dangerous in a state-of-nature situation is typically much more severe than the minimum harm required for protection against a criminal in custody. If our justification is the right to harm in self-defense, what we can legitimately do to a criminal in custody to protect ourselves against him is determined by the minimum required to protect ourselves against him in his actual situation. If one proposed to harm him more severely, for instance, to provide credibility for a system of threats, the right to harm in self-defense would not supply the requisite justification, and one would again be in danger of endorsing a position subject to the "use" objection.

What is the minimum harm required to protect ourselves from a violent and dangerous criminal in custody? It seems evident that nothing more severe would be required than isolating him from those to whom he poses a threat. Thus, it would appear that Farrell's reasoning cannot justify *punishment* of criminals, exactly, supposing that punishment involves the intentional infliction of significant harm, such as death or severe physical or psychological suffering. Rather in the case of violent and dangerous criminals, this reasoning would at best justify only preventative detention.

AN INCAPACITATION ACCOUNT THE QUARANTINE ANALOGY

A more resilient proposal for treatment of dangerous criminals than either the moral education or deterrence theories, and one that is compatible with free will skepticism, invokes our right to defend ourselves and to secure our safety, but employs the analogy to quarantine. Ferdinand Schoeman has argued that if protection of society justifies quarantining carriers of severe communicable diseases, then it also justifies isolation of the criminally dangerous (Schoeman 1979). Suppose an agent poses a danger to society by a demonstrably strong tendency to commit murder. Even if he is not in general a morally responsible agent in the basic desert sense, the state would nevertheless seem to have as much right to isolate him as it does to quarantine a carrier of a deadly communicable disease who is not responsible in this sense for being a carrier. The concern about using people merely as means has force in this context, and this together with the weight of the general right to liberty should restrict preventative detention to especially dangerous cases. Crucially, these countervailing factors count more heavily against punishment policy justified on consequentialist grounds than it does against preventative detention based on the quarantine analogy. For on the quarantine view, just as it is morally wrong to treat carriers of a disease more severely than is necessary to keep them from being dangerous to society, it will be wrong to treat those with violent criminal tendencies more harshly than is required to keep them from being a danger to society. Furthermore, the less dangerous the disease, the less invasive the legitimate prevention methods would be; and similarly, the less dangerous the criminal, the less invasive the legitimate prevention methods of incapacitation would be. For certain minor crimes, perhaps only some degree of monitoring could be defended.

Schoeman explores the acceptability of preventative detention for those who have not yet committed crimes, reflection on which occasions the following concern for this incapacitation account uarantine view. If justification of detention by this analogy is tenable, must it not then be legitimate to detain those who have not committed a violent crime, if by some means it has been ascertained that they are likely to do so? Schoeman contends, and I agree, that the right to liberty must carry weight in this context, as should the concern about using people merely as means. In addition, the risk posed by a state policy that allows for preventative detention of nonoffenders needs to be taken into serious consideration. In a broad range of societies, allowing the state this option stands to result in much more harm than good because misuse would be likely. Schoeman also argues that although the kinds of testing required to determine whether someone is a carrier of a communicable disease may often not be unacceptably invasive, the type of screening necessary for determining whether an agent has violent criminal tendencies might well be invasive in ways that raise serious moral difficulties. Moreover, available psychiatric methods for discerning whether an agent is likely to be a violent criminal are not especially reliable, and as Morse points out, detaining someone on the basis of a screening method that frequently yields false positives is seriously objectionable (Morse 1999; Nadelhoffer et.al. forthcoming)

However, there is reason to think that impressive neural methods of testing for violent tendencies are being developed (Nadelhoffer et al. forthcoming). The time may come soon when we are able to determine with reasonable accuracy on the basis of neural factors that an agent is likely to commit violent crimes in his normal environment. Wouldn't an account based on the quarantine analogy endorse preventative detention even if the agent had not yet

manifested such violence, supposing the violence would be very serious and highly likely in his normal environment, and less invasive measures such as effective monitoring or drug therapy were unavailable? Perhaps it would. But this should not count as a strong objection to this incapacitation account quarantine view because virtually everyone should agree that preventative detention of nonoffenders would be legitimate under some possible conditions. Imagine that an agent has involuntarily been given a drug that makes it virtually certain he will brutally murder at least one person while he is under its influence, which is for a period of 1 week. There is no known antidote, and because he is especially strong, mere monitoring would be ineffective. I daresay almost everyone would affirm that it would be at least prima facie permissible to preventatively detain him for the week. If you agree, then in principle you accept that preventatively detaining nonoffenders is legitimate under some possible circumstances. Now suppose that reliable neural screening reveals that an agent, if left in his normal environment, is virtually certain to murder at least one person within the next 5 years. There is no known viable drug therapy, and mere monitoring would be ineffective. Whether detaining this agent is also prima facie legitimate might be judged by how similar his case is to that of the drug victim. And it might well be similar enough. Here, it is important to understand that the advocate of the incapacitation account quarantine view will specify that the circumstances of such detention should not be harsh, and that allowing the agent to be reasonably comfortable and to pursue fulfilling projects must be given high priority. But at the same time, in many societies, the danger of a state policy that allows for preventative detention of even highly

dangerous nonoffenders is a grave concern that stands to outweigh the value of the safety provided by such a policy.[8]

When a person with cholera is quarantined, she is typically made to experience deprivation she does not deserve. Society benefits by this deprivation. It is plausibly a matter of fairness that society should do what it can, within reasonable bounds, to make the victim safe for release as quickly as possible. If we quarantined cholera victims but were unwilling to provide medical care for them because it would require a modest increase in taxation, then we would be acting unfairly. Similarly, when a dangerous agent, whether or not he has already committed crimes, is preventatively detained, then supposing that the free will skeptic is right, he is also made to experience a deprivation he does not fundamentally deserve, and from which society benefits. By analogy with the cholera case, here also it is a matter of fairness for us to do what we can, within reasonable bounds, to make him safe for release. For a society or state to oppose programs for rehabilitation because it is unwilling to fund them would involve serious unfairness.

Policies for making an agent safe for release would address conditions that underlie actual or potential criminal behavior. These conditions include psychological illness, but also problems that are not plausibly classified as illness, such as insufficient sympathy for other people, or a strong tendency to assign blame to others for whatever goes wrong. What binds these policies together is not that they treat the agent as mentally ill and therefore in need of

[8] For additional discussion of objections to the quarantine analogy, see Pereboom 2001: 174-177.

psychiatric treatment. Rather, they are policies that attempt to bring about moral change by non-punitively addressing conditions that underlie actual or potential criminal behavior.

It is often argued that rehabilitative views are objectionable because they treat an offender as suffering from an illness and not as a morally responsible agent. Morris contends that the problem for typical forms of therapy proposed for altering criminal dispositions is that they circumvent rather than address the qualities in human beings that confer dignity, in particular our abilities to regulate actions autonomously and rationally (Morris 1968). However, an offender may be in need of therapy and may correctly be regarded as morally responsible in the sense of answerability or of the fittingness of providing a moral explanation, and this sense of moral responsibility is grounded precisely in these capacities for autonomy and rationality in action. Moreover, in the case of many criminals, these capacities are impaired, but then regard for their humanity would recommend therapeutic measures to restore them. All of this the free will skeptic can unequivocally endorse.[9]

FINAL WORDS

Thus, if the free will skeptic is right, criminal punishment for retributive reasons is ruled out. We would then need to relinquish one of the most prominent ways for justifying criminal punishment, although there are independent objections to this position. Moreover, deterrence theories of punishment, although typically not overtly retributivist, are objectionable on independent grounds, and some may need to avail themselves of retributivist justifications at crucial junctures. But a theory of crime prevention that would be acceptable whether or not the

[9] For further discussion of objections to rehabilitation programs, see Pereboom 2001: 178-185.

skeptic is right can be developed by analogy with our rationale for quarantining carriers of dangerous diseases. Such a theory would not justify the sort of criminal punishment whose legitimacy is most dubious, such as death or confinement in the most common kinds of prisons in our society. More than this, it demands a certain level of care and attention to the well-being of criminals, which would change much of current procedure. The free will skeptic would also endorse measures for reducing crime that aim at altering social conditions, such as improving education, increasing opportunities for fulfilling employment, and enhancing care for the mentally ill (Kleiman 2009, 117-163; Slobogin 2006). These proposals would not extinguish our practice of holding morally responsible, not even its criminological component. Rather, they amount to revisions to the practice that its own rules would recommend. Thus, the challenge that the free will skeptic proposes is not external, but rather thoroughly internal to our practice of holding morally responsible.

References

Alexander, Larry, and Kimberly Kessler Ferzan, with Stephen Morse (2009). *Crime and Culpability*, Cambridge, UK: Cambridge University Press.

Balaguer, Mark (2010). *Free Will as an Open Scientific Problem*, Cambridge MA; MIT Press.

Bentham, Jeremy (1823/1948). *Introduction to the Principles of Morals and Legislation*, New York: Macmillan.

Bok, Hilary (1998). *Freedom and Responsibility*, Princeton NJ: Princeton University Press.

Boonin, Daniel (2008). The Problem of Punishment, Cambridge, UK: Cambridge University Press.

Braithwaite, John, and Philip Pettit (1990). *Not Just Deserts*, Oxford: Oxford University Press.

Chisholm, Roderick (1964). “Human Freedom and the Self,” The Lindley Lecture, Department of Philosophy, University of Kansas. Reprinted in *Free Will*, ed. Derk Pereboom, Indianapolis: Hackett, 2009.

Chisholm, Roderick (1976). *Person and Object*, La Salle IL: Open Court.

Clarke, Randolph (2003). *Libertarian Theories of Free Will*, New York: Oxford University Press.

Ekstrom, Laura W. (2000). *Free Will, A Philosophical Study,* Boulder CO: Westview.

Farrell, Daniel M. (1985). “The Justification of General Deterrence,” *Philosophical Review* 104, 367–394.

Fischer, John Martin, and Mark Ravizza (1998). *Responsibility and Control, A Theory of Moral Responsibility*, New York: Cambridge University Press.

Fischer, John Martin (2007), “Compatibilism,” in Fischer, John Martin, Robert Kane, Derk Pereboom, and Manuel Vargas. *Four Views on Free Will*, Oxford: Blackwell Publishers.

Frankfurt, Harry G. (1971). “Freedom of the Will and the Concept of a Person,” *Journal of Philosophy* 68, 5–20.

Ginet, Carl (1990). *On Action*, Cambridge, UK: Cambridge University Press.

Griffiths, Meghan (2010). “Why Agent-Caused Actions Are Not Lucky,” *American Philosophical Quarterly* 47, 43–56.

Hampton, Jean (1984). “The Moral Education Theory of Punishment,” *Philosophy and Public Affairs* 13, 208–238.

Haji, Ishtiyaque (1998). *Moral Accountability*, New York: Oxford University Press.

Honderich, Ted (1988). *A Theory of Determinism,* Oxford: Oxford University Press.

Hume, David. (1739/1978). *A Treatise of Human Nature*, Oxford: Oxford University Press.

Hume, David. (1748/2000). *An Enquiry Concerning Human Understanding*, Oxford: Oxford University Press.

Husak, Douglas (2000). "Retribution in Criminal Theory," *San Diego Law Review* 37, 959–986.

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

Kant, Immanuel (1797/1963). *The Metaphysical Elements of Justice*, New York: Bobbs-Merrill, tr. John Ladd, 99–107.

Kelly, Erin (2009). "Criminal Justice without Retribution," *Journal of Philosophy* 106, 440–462.

Kleiman, Mark (2009). *When Brute Force Fails: How to Have Less Crime and Less Punishment*, Princeton, NJ: Princeton University Press.

McCloskey, H. J. (1965). "A Non-utilitarian Approach to Punishment," *Inquiry* 8, 239–255.

McKenna, M. (2012). *Conversation and Responsibility*, New York: Oxford University Press.

Mele, Alfred (1995). *Autonomous Agents*, New York: Oxford University Press.

Mele, Alfred (2006). *Free Will and Luck*, New York: Oxford University Press.

Menninger, Karl (1968). *The Crime of Punishment*, New York: Penguin.

Montague, Philip (1995). *Punishment as Societal Defense*, Lanham, MD: Rowman and Littlefield.

Moore, Michael (1987). "The Moral Worth of Retribution," in *Responsibility, Character, and the Emotions*, Ferdinand Schoeman, ed., Cambridge, UK: Cambridge University Press, pp. 179–219. Reprinted in Punishment and Rehabilitation, 3rd edition, Jeffrie G. Murphy, ed., Belmont, CA, Wadsworth Publishing Company, 1995, pp. 94–130.

Moore, Michael (1998). *Placing Blame*. Oxford: Oxford University Press.

Morris, Herbert (1968)."Persons and Punishment," *The Monist* 52, 475–501

Morris, Herbert (1981). "A Paternalistic Theory of Punishment," *American Philosophical Quarterly* 18. Reprinted in *Punishment and Rehabilitation*, 3rd edition, Jeffrie G. Murphy, ed., Belmont, CA: Wadsworth Publishing Company, 1995, page numbers are from the latter edition.

Morse, Stephen (1999). "Neither Desert nor Disease," *Legal Theory* 5, 265–309.

Morse, Stephen (2004). "Reasons, Results, and Criminal Responsibility," *University of Illinois Law Review* (2004), 363–444.

Morse, Stephen (2007). "The Non-Problem of Free Will in Forensic Psychiatry and Psychology," Scholarship at Penn Law, paper 157. Available at: http,//lsr.nellco.org/upenn_wps/157.

Nadelhoffer, Thomas, Stephanos Bibas, Scott Grafton, Kent A. Kiehl, Andrew Mansfield, Walter Sinnott-Armstrong, and Michael Gazzaniga (forthcoming), "Neuroprediction, Violence, and the Law: Setting the Stage," *Neuroethics*.

Nelkin, Dana (2008). "Responsibility and Rational Abilities, Defending an Asymmetrical View," *Pacific Philosophical Quarterly* 89, 497–515.

Nelkin, Dana (2011). *Making Sense of Freedom and Responsibility*, Oxford: Oxford University Press.

Nichols, Shaun (2007). "After Compatibilism, a Naturalistic Defense of the Reactive Attitudes," *Philosophical Perspectives* 21, 405–428.

O'Connor, Timothy (2000). *Persons and Causes,* New York: Oxford University Press.

O'Connor, Timothy (2008). "Agent-Causal Power," in *Dispositions and Causes*, Toby Handfield, ed., Oxford: Clarendon Press, pp. 189–214.

Pereboom, Derk (1995). "Determinism Al Dente," *Noûs* 29, 21–45.

Pereboom, Derk (2001). *Living Without Free Will*, Cambridge, UK: Cambridge University Press.

Pereboom, Derk (2004). "Is Our Conception of Agent Causation Incoherent?" *Philosophical Topics.* 32, 275–286.

Pereboom, Derk (2005). "Defending Hard Incompatibilism," *Midwest Studies in Philosophy* 29, 228–247.

Pereboom, Derk (2006). "Kant on Transcendental Freedom," *Philosophy and Phenomenological Research* 73, 537–567.

Pereboom, Derk (2007). "Hard Incompatibilism," and "Response to Kane, Fischer, and Vargas," in *Four Views on Free Will*, Robert Kane, John Martin Fischer, Derk Pereboom, and Manuel Vargas, eds., Oxford UK: Blackwell, pp. 85–125, 191–203.

Quinn, Warren (1985)."The Right to Threaten and the Right to Punish," *Philosophy and Public Affairs* 14, 327–373.

Rawls, John (1955). "Two Concepts of Rules," *The Philosophical Review* 64, 3–32.

Scanlon, Thomas (1998). *What We Owe to Each Other*, Cambridge, UK: Harvard University Press.

Schoeman, Ferdinand D. (1979). "On Incapacitating the Dangerous," *American Philosophical Quarterly* 16, 27–35.

Slobogin, Christopher (2006). *Minding Justice*, Cambridge, MA: Harvard University Press.

Smilansky, Saul (2000). *Free Will and Illusion*, Oxford: Oxford University Press.

Sommers, Tamler (2007). "The Objective Attitude," *Philosophical Quarterly* 57, 321–41.

Sommers, Tamler (2012). *Relative Justice: Cultural Diversity, Free Will, and Moral Responsibility*, Princeton, NJ: Princeton University Press.

Spinoza, Baruch (1677/1985). Ethics, *in The Collected Works of Spinoza*, Edwin Curley, ed. and tr., Vol. 1, Princeton, NJ: Princeton University Press.

Strawson, Galen (1986). *Freedom and Belief*, Oxford: Oxford University Press.

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

Taylor, Richard (1966). *Action and Purpose*, Englewood Cliffs, NJ: Prentice-Hall.

Taylor, Richard (1974). *Metaphysics,* 4th ed., Englewood Cliffs, NJ: Prentice-Hall.

Ten, C. L. (1987). *Crime, Guilt, and Punishment*, Oxford: Oxford University Press.

van Inwagen, Peter (1983). *An Essay on Free Will*, Oxford: Oxford University Press.

Vilhauer, Benjamin (forthcoming). "Persons, Punishment, and Free Will Skepticism," *Philosophical Studies,* Online First, DOI 10.1007/s11098-011-9752-z, forthcoming.

Wallace, R. Jay (1994). *Responsibility and the Moral Sentiments*, Cambridge MA: Harvard University Press.

Waller, Bruce (1990). *Freedom Without Responsibility,* Philadelphia, Temple University Press.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*, tr. G. E. M. Anscombe, Oxford: Basil Blackwell.

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

Wood, Allen (1984). "Kant's Compatibilism," in *Self and Nature in Kant's Philosophy,* Ithaca NY: Cornell University Press, pp. 73–101.