Chapter 6:

Deontology and Deterrence for Free Will Deniers

Benjamin Vilhauer

In this paper[1] I outline what I take to be a solution to a problem about free will denial and the justification of punishment pointed out by Saul Smilansky (2011). Smilansky argues that free will deniers must acknowledge that some institution of punishment is necessary to maintain law and order, but since criminals do not deserve to be punished, it is unjust to punish them, and we therefore have a duty to compensate them. Since this is a great injustice, we must compensate them very heavily—in fact so heavily that the institution of punishment will cease to deter, and will instead become an incentive to commit crime. Previous responses to Smilansky's "practical *reductio*" argument by Neil Levy (2012) and Derk Pereboom (2014) have emphasized consequentialist moral reasons. I advocate a deontological social contract approach to punishment which draws on Kantian and Rawlsian notions of treating criminals as ends by respecting their rational consent to punishment (Vilhauer 2013)[2]. In the course of explaining how my approach provides a response to Smilansky's challenge, I will also respond to some objections to it from Pereboom.

1. Introduction

Let me begin with a few terms and a bit of background. By "free will skepticism" I mean the view that we do not know whether or not we have free will, and by "free will denial" I mean the view that we lack free will. I endorse skepticism rather than denial. I think that it is possible that we have free will, and that this possibility is sufficient to justify appeals to free will in some contexts (which include taking up some of the positive reactive attitudes, and supporting the "ought implies can" principle) but *not* in the context of retributively justifying

[1] Acknowledgements: Thanks to Elizabeth Shaw for organizing the conference for which this paper was first developed, and to Elizabeth as well as Gregg Caruso and Derk Pereboom for their editorial work on this volume. Thanks also to the conference participants for their valuable suggestions.

[2] I draw on remarks in my 2013 paper in some sections of this paper.

serious harm of the sort involved in punishment. People deserve the benefit of the doubt. The upshot is that the constraints placed on the ethics of punishment by free will skepticism of this sort are the same as those placed on it by free will denial.[3]

According to some theorists, an approach to criminal justice which dispenses with the idea of retributive desert should not be called an approach to *punishment*, because such desert is part of the concept of punishment. I am not convinced that this is correct, since we do speak of the injustice of punishing the innocent. I would nonetheless be happy to conform to the usage recommended by these theorists, and call the approach I recommend an approach to criminal justice rather than punishment, were it not for a danger of euphemism here which is greater than the danger of infringing a norm of usage. Calling a practice *punishment* emphasizes the need for justification in a way that calling it *criminal justice* does not, and free will skeptics must be especially sensitive to this need.

As I read Smilansky's practical *reductio* argument, it has two important stages. First, free will deniers acknowledge that violent criminals do not deserve to be incarcerated based on their crimes. But violent criminals must be incarcerated to maintain enough law and order to avoid a return to the state of nature. We must therefore compensate them for the wrong we are doing to them by incarcerating them. Since this is a very great wrong, their compensation must equivalently great—we must make life in prison very pleasant. We must make it so pleasant that not only will the deterrent effect of incarceration disappear, but also we will create an incentive to commit violent crimes as a means to the end of becoming incarcerated, an incentive which will inevitably lead some people to commit violent crimes which they

[3] This part of my position is in significant agreement with Pereboom's position. He too endorses what I am calling skepticism rather than denial, and holds that the possibility of free will is not sufficient to justify appeals to free will in retributive justifications of harm. He argues that the "reasonable doubt" standard in criminal law should also be applied to metaphysics of moral responsibility (Pereboom 2001: 161). I agree with this, but I argue that people deserve the benefit of the doubt in the context of praise as well as blame, and that it is justified to offer harmless praise (praise that benefits some without harming others) so long as can be reasonably doubted that the candidate for praise did *not* act with free will (Vilhauer 2015).

would not have committed in the absence of that incentive. The institution of incarceration will become a facilitator of violent crime rather than an inhibitor. We will have replaced punishment with “funishment,” with the consequence of destroying the law and order society needs to survive.

The second stage of the argument is Smilansky’s claim that a morally deep solution to this problem must be deontological. He thinks that while it is easy to solve this problem with utilitarianism, if free will deniers solve it in this way, or with any other form of consequentialism that does not recognize giving people what they deserve based on their actions among the valued consequences, they make the distinctive moral intuitions proper to free will denial inexplicable, and they cease to be able to understand free will denial as an ethical position which is distinct from utilitarianism. That is, free will deniers typically direct their attention to punishment when prompted by the intuition that there is a grave injustice in the way our society punishes criminals, because they do not deserve to suffer for their crimes. However, according to the sort of consequentialism that easily solves this problem, there need be no injustice here at all, so long as our punitive practices lead to the best consequences. Further, if we endorse this sort of consequentialism, then we must accept that the entire debate about whether we have free will and moral responsibility never really mattered for ethics in the first place, because even if we do have free will and moral responsibility, we have no obligation to treat each other in the ways in which we deserve to be treated based on our actions, because such treatment is at best contingently related to the production of valuable consequences.

I think Smilansky is correct that the justification of deterrence is a pressing problem for free will skeptics, and that it is valuable to offer a deontological justification. However, while Smilansky claims that free will deniers do not have a deontological justification available, I think they do, as I will argue shortly.

I do not think that a utilitarian approach is as fundamentally flawed as Smilansky argues. As Neil Levy points out, utilitarians can sensibly advocate free will denial as a way of winning new converts to utilitarianism, since many strong objections to utilitarianism derive from intuitions about action-based desert which free will deniers reject (Levy 2012: 484). This would seem to imply that the intuitions which flow from denying action-based desert claims are not necessarily at odds with utilitarianism in the way Smilansky suggests.

There are, however, reasons for all philosophers to worry about utilitarianism which need not have anything to do with intuitions about action-based desert. For example, if our sole justification for punishment is that it reduces overall suffering in society, then we are using the people punished as mere means that we manipulate in order to benefit others. Further, suppose it turns out that we get the most deterrence for the least punitive pain with practices that strike our pretheoretical moral intuitions as abhorrent, such as imposing maximally painful punishments, weakening or violating due process, and framing non-criminals when it is more effective than punishing real criminals. As many have noted, if these practices turn out to give us the lowest ratio of suffering-caused-to-suffering-prevented, then utilitarianism must endorse them.

Our sense of the abhorrence of these practices does not have to be explained with reference to intuitions about action-based desert. As a free will skeptic, I think it is true that no criminal deserves a maximally painful punishment based on his criminal actions, but I think this would be true *even if we did* have free will and moral responsibility. As I see it, and as I think many others do, there is nothing anyone *could possibly do* to deserve a maximally painful punishment, even if we did have free will and moral responsibility. If this is true, then stating that no criminal should be given a maximally painful punishment because he does not deserve it based on his criminal actions does not *explain* the wrongness of maximally painful punishments. So objections to utilitarianism about punishment need not derive from intuitions about action-based desert. This gives free will deniers not already committed to utilitarianism a good reason to seek alternative ethical approaches to punishment.

Deontology is of course not the only alternative to simple utilitarianism. Pereboom's response to Smilansky's challenge is to recommend a consequentialism "where morally fundamental rights being honored and not violated count among the good consequences" (Pereboom 2014: 172). Let us refer to this approach as "rights-consequentialism." If we assume that we have fundamental rights to be free from maximally painful punishment, violations of due process, and framing, then according to rights-consequentialism, we should sometimes prefer an outcome which offers a level of protection of these rights L over an outcome that protects these rights less than L but achieves greater overall happiness. However, if we can maximize the overall protection of rights by violating rights in a few cases, then it would seem that rights-consequentialists must endorse violating rights in a few cases. Suppose that by falsely accusing a few criminal court judges of violations of due process, and then framing and punishing them, we could generate a great blast of deterrence that would dramatically reduce the overall violation of due process by other participants in the legal system. Should we do it? It looks like rights-consequentialists have to say "yes". Now, I do not want to claim that this is obviously wrong, and I lack space here to argue that it is wrong. I think it is fair to acknowledge that this is a hard problem. But it also seems fair to claim that framing and punishing even a few people clashes with our pretheoretical intuitions about the absolute status of our rights to due process, and it should be of interest to free will deniers whether a theory that makes these rights absolute is available to them. Perhaps we can get absolute rights from some version of rule-consequentialism, but I have never been able to get over the often-noted stumbling block for rule-consequentialism that it would seem to give us sufficient reason to break the rules whenever the consequences of breaking them are better than the consequences of following them. So I think it is worth exploring a deontological approach.

2. Funishment and Circustine

Before I set out the deontological approach in more detail, I would like to briefly consider another aspect of Pereboom's approach, which he uses to set a limit on the harshness of the treatment we can impose on criminals, giving it a response to the objection to rights-consequentialism just made. Pereboom emphasizes an analogy between quarantining carriers of dangerous illnesses and incarcerating violent criminals. I think this analogy is valuable for shifting the paradigm we use to think about punishment, and I think the approach I advocate could serve as a sort of deontological scaffolding for the quarantine justification, since I think they prescribe the same sort of treatment for prisoners in the end.[4]

Pereboom argues for the quarantine justification by pointing out that there is broad agreement that quarantine can be justified, despite the fact that carriers of disease have done nothing to deserve to be quarantined, and arguing by analogy for a justification of punishment. Society must have as much right to separate a violent criminal from the society he threatens as it does to quarantine a disease carrier. Pereboom also argues that constraints on how we can treat prisoners follow from this. On his view, we could only be justified in imposing unpleasant treatment on people in quarantine insofar as it is necessary to protect society from them. The same follows for incarceration when justified by way of this analogy—we cannot impose unpleasant treatment on people in prison beyond what is necessary to keep society safe from them (Pereboom 2001: 177). This limitation on the harm we can legitimately impose on criminals places an important side-constraint on his rights-consequentialism, by (for example) ruling out violations of the right to be free from cruel and unusual punishment of a few people even in cases in which this would maximize society's

[4] Gregg Caruso (2019, this volume) also defends a justification of punishment based on the quarantine analogy.

access to this right in the aggregate. If I understand the view correctly, Pereboom holds that publicizing the policy of imprisoning violent criminals is likely to have a deterrent effect despite the fact that our justification for incarceration is drawn from the quarantine analogy (Pereboom 2001: 177), but the fact that our justification is drawn from the quarantine analogy implies that we are justified in coercively separating violent offenders from the rest of the population but *not* treating them in ways that generate deterrence *except* insofar as deterrence is a fortunate by-product of the coercive separation (Pereboom 2014: 171). In other words, on Pereboom's model, deterrence that may result from quarantine is a fortunate contingency, but is not something that we may, as it were, calibrate the conditions of incarceration to achieve.

As I understand it, it is central to Pereboom's argument that in quarantine, we find a social practice which has a justification that is clear and uncontroversial, and which is also entirely independent of questions of deterrence. However, I think that philosophical reflection on the notion of effective quarantine shows that it depends upon deterrence, and that a justification of effective quarantine therefore demands a justification of deterrence. We can imagine having reasons of morality or even social policy to want the conditions of quarantine to be pleasant enough that we would risk creating an incentive for people to voluntarily expose themselves to a disease as a means to the end of entering quarantine. So the basic structure of the argument Smilansky makes about funishment applies here as well. Since people exposed to a serious illness have done nothing to deserve the isolation of quarantine, we would seem to have a duty to make quarantine as pleasant as possible, and if we do a good enough job in fulfilling this duty, we create an incentive to enter quarantine. We might aim at making quarantine as entertaining as the circus, so let us name this problematically pleasant quarantine "circustine" to emphasize the parallel to funishment.

Given the fear of suffering and death provoked by most of the diseases against which we would employ quarantine, it may seem absurd to suggest that anyone might have an incentive to expose herself to disease to enter quarantine. But imagine an extremely contagious and incurable philosomania that made the afflicted so pathologically excited about reading and discussing philosophy that they were unable to do anything else except eat and sleep. Without a strict and permanent quarantine this disease would mean the end of civilization. Given the injustice involved in separating the afflicted from their families and communities, we would have strong moral reasons to make this quarantine as enjoyable as possible, and allow the people confined to go on about their philosophical activities. But given that philosomania would not be intrinsically unpleasant, especially when in the company of its other victims, it is not hard to imagine that a substantial number of people might be inclined to expose themselves to it in order to enter the quarantine, thereby creating the conundrum of circustine.

For a less empirically implausible example, consider the 2014 Ebola epidemic, which experts worried might lead to a global pandemic. At its heights, some commentators proposed easing the complaints of people asked to remain in quarantine by paying them a substantial sum for their trouble. They seemed to be motivated by moral concerns about infringement on autonomy, by pragmatic concerns about getting people to stay in quarantine, and, interestingly enough, by a desire to create an incentive for some people to expose themselves to Ebola. Now, the only people they sought to motivate to expose themselves to Ebola were medical personnel, more of whom they wanted to travel to the hardest-hit countries. However, if high enough quarantine payments were authorized, other people might have been motivated to voluntarily expose themselves to Ebola too, for example, the very poor, or very elderly people who wanted to leave something to their great-grandchildren. This may give us a more realistic path to circustine.

These examples demonstrate that we must ensure a minimum level of unpleasantness, or a maximum level of pleasantness, depending on how you look at it, for quarantine to achieve its goal, and it seems clear that we need a moral justification to explain why we are entitled to calibrate quarantine conditions to this level. It seems to me correct to describe this as calibrating quarantine conditions to achieve deterrence. If this is right, then the justification of quarantine turns out to be inherently entangled with the justification of deterrence, in the sense that it is necessary to have a justification of whatever degree of deterrence is necessary to avoid circustine if one is to have a general justification of effective quarantine. Certainly in most cases the fear of the disease itself and its symptoms will provide the necessary deterrent, but the examples show that this is a contingent matter rather than something entailed by the concepts of disease or quarantine.

This line of thought also highlights an important difference in the degrees of deterrence needed to avoid circustine and funishment. Since most of the diseases against which we would institute quarantine are more like Ebola than philosomania, an incentive sufficient to prompt people to enter quarantine would have to be powerful enough to overcome not just the disincentive of confinement and isolation from mainstream society, but also the disincentive of exposure to the disease and the harm of possibly getting sick. An incentive sufficient to prompt potential criminals to enter the sort of humane prisons contemplated by most free will deniers would only have to be strong enough to overcome the disincentive of confinement and isolation from mainstream society, assuming that the potential criminals would not find it distressing to commit the violent crimes they would have to commit to be admitted.

I take this to demonstrate two things that reinforce Smilansky’s worry: first, in both quarantine and incarceration, these institutions will not fulfill their intended function of making society safer without some degree of deterrence, and second, we should assume that the conditions necessary to ensure deterrence in incarceration are more unpleasant than the conditions needed in quarantine. If more unpleasant conditions are needed to ensure deterrence in incarceration, we must have an account of why we are justified in imposing them, and how unpleasant they can be. As I explained above, I think that it is valuable for free will deniers to have a deontological justification available.

3. A Deontological Approach to Punishment for Free Will Deniers

The deontological approach to be defended here draws on a Rawlsian version of the Kantian idea of refraining from treating people as mere means to ends.[5] The Kantian idea is that we refrain from treating others as means by refraining from coercing, deceiving, or otherwise manipulating other people into serving as means to our ends by causing them to do things they would not rationally consent to do. Crucially, the “others as ends” principle does not require us to avoid treating each other as means in all cases, because it is sometimes rational to consent to serving as a means to another’s end, for example, when he is reciprocally serving as a means to one's own end. For example, if I have the end of teaching some philosophy, then the students I teach are among the means to my end. If those students have the end of learning philosophy, then I, as their teacher, am among the means to their end. So the students can be a means to my end and I can reciprocally be a means to their end. We are treating each other as ends as well as means, because we can rationally consent to this interaction, in light of our complementary ends.

Social contract theory can model rational consent. If it would be rational to choose to join in a social contract with a particular institutional structure, then we can view that institutional structure as one to which we would rationally consent. I think that Rawls’ approach to social contract theory should be of special interest to non-consequentialist free will skeptics, because we can use original position deliberation to capture an underlying moral distinction between the action-based kind of desert typically at issue in the free will literature, on the one hand, and personhood-based desert, on the other.

[5] Sharon Dolovich (2004) proposes a similar approach, but not in the context of free will skepticism. My own approach develops out of ideas in Vilhauer (2009).

A "desert base" is whatever grounds a desert claim. Many commonplace desert claims are based on actions. Examples of action-based desert claims are claims about praise and blame, Lockean claims about coming to own property by mixing our labor with matter, and claims about why criminals deserve retribution. Actions can be desert bases only if agents are morally responsible for actions, so free will deniers must hold that action-based desert claims are never legitimate. Many philosophers assume that all desert claims are action-based, and as a result it comes naturally to many free will deniers to hold that there are no legitimate desert claims. But I think there are desert claims which are not action-based. Some are based on personhood. Claims to deserve respect, not to be used as mere means, and to access our human rights are not based on our actions but are instead based on the mere fact that we are persons. Free will denial does not undermine personhood-based desert in the way it undermines action-based desert, so free will deniers can endorse personhood-based desert claims.

Consider the presumption of innocence. As a conceptual matter, we regard the presumption of innocence to be such that nobody could act in such a way as to deserve to have it withdrawn. Someone who was not treated as innocent until proven guilty could legitimately claim to deserve rectifying measures. But she could not explain this claim by pointing out that she had not done anything to deserve not to be treated as innocent until proven guilty, because there is nothing anyone *could* do to deserve not to be treated as innocent until proven guilty. On my view, her claims to deserve to be treated as innocent until proven guilty, and to deserve rectifying measures if she is not, are personhood-based. I think all our claims to due process of law, and equal treatment before the law, have this structure.

On my view, free will deniers should think that the kind of rational consent that matters is rational consent in light of personhood- but not action-based desert. In other words, what should matter for free will deniers is what it would be rational for us to consent to if all we had in view was the mere fact that we are persons. The Rawlsian social contract fits well here. As is of course well-known, on this view, we choose the basic principles of society from a deliberative standpoint called "the original position," behind a "veil of ignorance." Rawls' moral purpose in designing the original position is to create a standpoint in which deliberators' reasons of self-interest result in the selection of basic principles of society which follow the lines of broadly Kantian reasons of fairness when they are implemented. As an original position deliberator, one has broad empirical knowledge about the human condition, including psychological, sociological, and economic knowledge, but one lacks knowledge of particular features of oneself such as whether one is rich or poor, or what one's religion, ethnicity, or sex is. One also cannot know what patterns of action one exhibits, for example, whether one is industrious or lazy. Rawls thinks it is just to demand ignorance about these features of ourselves because they are morally irrelevant to choosing the basic principles of society, and he thinks this at least in part because he thinks we do not deserve to have these particular features. In other words, Rawls is motivated at least in part by a kind of skepticism about desert, and this makes him a natural ally for deontologically-inclined free will deniers.

When it comes to human rights and distributive justice, I think that free will deniers can take on board Rawls' view of original position deliberation in its entirety. Rawls thinks that original position deliberators will insist that everyone's basic needs are met, and that there is equality of rights and political liberties. He also thinks they will choose the "difference principle," the principle that economic inequalities are just if and only if they improve the conditions of the worst-off members of society. Part of what makes it rational for us to choose the difference principle in the original position is the veiling of our degree of industry in the original position.

Rawls himself does not apply original position deliberation to penal justice. Rawlsian "ideal theory" requires us to assume that we will be able to follow the laws we choose in the original position. But this presupposes a kind of control over our actions which opens the door to the justification of retribution, which is off-limits for free will deniers. So free will deniers adopting original position deliberation about punishment must assume ignorance about whether we will follow the law. In other words, in the original position, I cannot assume that I can avoid liability to punishment by avoiding crime—call this the "avoidability of crime assumption." If I assume that I can control my liability to punishment by avoiding crime, my fear that I will suffer punishment myself will diminish, and if I assume that punishment deters I will prefer more severe punishments than I would otherwise have accepted. But accepting the avoidability of crime assumption entails accepting ideas about control that free will deniers must reject.

Rejecting the avoidability of crime assumption is a substantial departure not just from Rawls, but from much of the social contract tradition, since criminals are often represented as "contract breakers" who have placed themselves back in the state of nature with respect to the rest of society. On my view, any contract we could rationally enter must include specifications for how we will be treated if we violate the laws—it must be, in this sense, an unbreakable contract.

How would original position deliberation apply to punishment? Rawls' general idea about original position deliberation is that deliberators will use maximin reasoning: that is, they will focus on the lot of the worst-off, and choose principles that make their lot as good as it can be. Rawls at points argues for maximin reasoning as a kind of risk-aversion which would be rational under conditions of uncertainty, and this has been disputed by many. But he also argues that we should stipulate that original position deliberators reason in this way because it ensures fairness, by procedurally implementing a broadly Kantian idea about fair cooperation and equality among rational beings, and free will deniers interested in exploring alternatives to consequentialism have no reason to dispute this. Let me explain this point in more detail. A variety of critics including Harsanyi (see e.g. Harsanyi 1975) hold that that Rawls fails to show that the concept of *choosing behind the veil of ignorance* implies the maximin principle on its own, and that Rawls must build non-consequentialist moral premises into his theory to arrive at the maximin principle. The concern is that without stipulating that rational choice is non-consequentialist choice, there is no way to rule out the possibility that original position deliberators might choose consequentialist principles that (for example) maximize overall well-being and set no limits on the wretchedness of the circumstances of the worst-off, with the deliberators simply betting that they will not be among the worst-off. Now, this concern is a reason to reject Rawlsian methodology if one's goal is to deduce moral principles solely from the mere concept of rational choice behind the veil of ignorance. But that is not my goal. I am proceeding from the assumption that it is valuable for free will skeptics to have a non-consequentialist approach to punishment, so my approach is already encumbered with a general sort of non-consequentialist assumption similar to the one critics attribute to Rawls. My goal is to borrow Rawlsian methodology precisely *because* it is non-consequentialist, but nonetheless filters out action-based desert-claims. I think Rawls is correct that his conception of original position deliberation models moral commitments about fairness and equality at the core of the Kantian approach to ethics, and for present purposes I am happy to assume the correctness of these commitments.

However, despite the moral reasons in favor of maximin reasoning, there are disanalogies between distributive justice and penal justice which create obstacles to applying maximin reasoning in the context of punishment. There is only one sort of candidate for the worst-off position when we are talking about distributive justice—the poorest. In the case of penal justice, there are two sorts of candidates for the worst-off position—victims of crime, and the people punished—and these two positions compete in the penal justice system. That is, if we assume that punishment deters, then changing our principles of punishment to make things better for victims tends to make things worse for the people punished, and changing our principles of punishment to make things better for the people punished tends to make things worse for the victims. (I will later argue that this competition should not be understood as fundamental, since all rational adopters of a social contract secure better outcomes for themselves than they can expect in the state of nature.) The technological and social strategies that might eliminate this competition in a desirable way are limited in contemporary society. Today we might strive to eliminate the position of victim with blanket surveillance and militarized policing, but even this would not prevent all crimes. It would in effect punish everyone, and it would be irrational to consent to universal punishment. We might someday put vast numbers of artificially intelligent ticklebots on patrol, which had the legal expertise and speed necessary to ensure that would-be criminals collapsed in helpless giggles before completing their crimes, obviating the need for punishment, but without social practices yet unimagined even the cheeriest ticklebots could easily become repressive tools of a security state. We could of course eliminate the position of the punished today by ceasing to punish people altogether, but we would worry that violent crime would become ubiquitous and we would be cast into the state of nature. If, for the foreseeable future, criminals and victims will compete in our institutions of punishment, how should we weigh their interests in original position deliberation? We seem compelled to make assumptions behind the veil of ignorance about the probabilities of finding ourselves among the punished, and among the unpunished.

Rawls himself holds that the veil of ignorance must be understood as screening out knowledge of the probabilities of finding ourselves in various social roles. This helps him motivate maximin reasoning—if we do not know the probabilities of finding ourselves in various social roles, it is more plausible to claim that self-interest requires us to design the worst role to be the best it can be. But as I have explained, there are also broadly Kantian moral reasons for preferring maximin reasoning to consequentialist reasoning, and this screening-out helps original position deliberation to model these reasons. If we could know that the probability of ending up in the worst-off position was very low, and we could make all the other positions better by making the worst-off position very bad (say by making the worst-off position that of a slave doing the bidding of everyone else) then it might be rational in a self-interested sense to make the worst-off position very bad and roll the dice. However, when our choice of the principles of punishment confronts us with two candidates for the worst-off position, it seems impossible to conduct original position deliberation without making some assumption about the probability of finding oneself in one position or the other. Given the non-consequentialist purpose of original position deliberation, and the importance of constraining deliberators' knowledge of probabilities in achieving that purpose, care must be taken in determining the probabilities we ought to assume. Suppose that original position deliberators knew that they could have a negligible chance of ending up among the punished if they chose principles of punishment that imposed maximally painful punishments on just a few of the punished but yielded a horribly effective deterrent. This would give them reasons of self-interest to select these principles of punishment. But the moral reasoning that original position deliberation would thereby model would be little different from that of simple utilitarianism.

Since the moral purpose of Rawlsian original position deliberation is to model broadly Kantian notions of equality and fairness, I can see no better way to select a probability assumption than by falling back on these notions, and holding that a principle is fair to competing parties if I would choose it under the assumption that my probability of being benefited by it is equal to my probability of being harmed by it. There may be other plausible non-consequentialist approaches to fairness and equality in selecting a probability assumption, but I lack the space to explore that possibility in this paper. It is clear that moral constraints on the probability assumption are necessary to preserve non-consequentialism in original position deliberation, and it seems reasonable to claim that an assumption of equal probability is one plausible resolution. On this basis, I claim that the principles of punishment are fair if I would choose them under the assumption that I am just as likely to be the person punished as I am to be a potential victim.

I identify potential victims as the relevant beneficiaries, rather than actual victims, because potential victims have more to gain from punishment than actual victims. Actual victims have already suffered the harm we would hope to avoid in the original position. For example, if the positions I considered were those of actual victims of serious violence and punished people, and I assumed I was equally likely to end up in either position, then I might reason that since the harm has already been done, I would gain little from punishment if I turn out to be the victim, and I would have a lot to lose if I turn out to be the punished person. I might conclude that I am better off in not endorsing any institution of punishment at all. Further, if we identify the relevant beneficiaries as potential victims, then we do not leave out anyone who can benefit from punishment, since everyone who has been victimized can potentially be re-victimized except victims of fatal violence. I identify the actual punished as the relevant harmed parties, rather than the potential punished, because the potential punished who do not become the actual punished are not actually harmed by the institution of punishment.

What principles of punishment would I choose if I had to assume that I was just as likely to be harmed by punishment as I was to benefit from it? As already suggested, it seems safe to assume that deliberators would have a strong initial preference for a society that had *no* institution of punishment. They would choose to pour social resources into the development of technology and social practices that allowed us to build a free and just society that did not punish. Recognizing that we do not yet have these tools, they would also choose to invest heavily in non-coercive strategies for diminishing incentives to commit crimes, such as improved access to employment, education, public services, and voluntary therapy for those most at risk of committing crimes. They would also recognize that as things stand today, some system of after-the-fact punitive constraint is necessary to avoid a collapse into the state of nature, but they would deem this legitimate only insofar as it was coupled with progress toward the abolition of punishment. This point is important because it implies that the approach I recommend can only be a justification of punishment in a relative sense, in that it sets abolition as an ideal and legitimizes punishment only as a temporary measure as we work towards its abolition.[6]

What sort of after-the-fact punitive practices would original position deliberators endorse under the assumption that they are equally likely to find themselves punished and unpunished? It is crucial to recognize that the trade-off we are talking about is one that imposes significant harm on the person punished in order to confer what may be a very modest benefit on the potential victim. The badness of life in prison—the control, and the separation from friends, family, and community—would be a significant harm even in a radically reformed prison of the sort that non-retributive ethicists could contemplate. But a reduction in someone's odds of becoming a victim of crime does not confer a similarly significant benefit, so long as he is not the sort of person who worries obsessively about his odds of becoming a victim. If I could know with certainty that choosing to use some particular kind of punishment in my society would make the difference between my remaining a merely potential victim and my becoming an actual victim, then the benefit to

[6] I refer to this approach as "ideal abolitionism" in Vilhauer (2017).

me might be just as great as the harm to the punished. But I cannot know this in the original position.

If we look at society as a whole, the benefit of punishment may be much greater than the harm it imposes: even if we cannot know whether particular individuals will be spared victimization, we may be able to know that there will be a substantial overall reduction in victimization. But this is not relevant in the original position. Rawls emphasizes that a key function of original position deliberation is to make us think about social outcomes one person at a time, and he criticizes utilitarianism for disregarding the boundaries between persons. The aggregate reduction in victimization is not something that happens to a person—it is an abstraction which is a function of many people. The fact that original position deliberation disregards aggregate harm reduction is part of what makes this approach a deontological approach. That is, it helps safeguard against the instrumentalization of criminals to which utilitarianism resorts in its unconstrained pursuit of harm reduction.

4. Deterrence and the Rights of Criminals

In light of these considerations, the question to ask is the following: how much harm am I willing to impose on the person punished for the sake of bringing the much-smaller benefit of reduced odds of victimization to the potential victim, assuming that I am just as likely to be the former as I am to be the latter?

We would be unwilling to risk imprisonment to protect ourselves against non-violent crime—we would prefer, for example, principles that required thieves to compensate their victims. We would take a different attitude toward violent crime, however. We would be willing to risk imprisonment to protect ourselves against crimes of violence, so long as the conditions of imprisonment were very different from those of contemporary prisons—that is, so long as they were humane, offered opportunities for voluntary therapy and rehabilitative treatment, and left room for a worthwhile life, including things like regular visits from friends and loved ones and opportunities for meaningful work.

The mere fact of imprisonment would prevent violent offenders from repeating their crimes, and this would be attractive to original position deliberators. However, their general knowledge of human psychology, sociology, and economics would make them realize that prison conditions had to be calibrated to create some degree of deterrence to avoid funishment. It seems clear that the mere fact of imprisonment could provide a substantial deterrent even if prison conditions were comfortable, since almost everyone would prefer not to have their actions restricted in the way that prison restricts action, even under humane conditions. Further, in a society that selected all its basic principles in the original position, our principles of penal justice would be implemented along with the principles of distributive justice chosen in the original position, which require that everyone's basic needs are met, that there is equality of rights and political liberties, and that the poorest are as wealthy as possible. As a result, there would be less upward pressure on crime rates from the problems caused in our own society by poverty and oppression, so less deterrent force would needed than is needed in our society. It must also be kept in mind that we almost certainly cannot deter everyone, and that we need not deter everyone to maintain a society of law and order. So there is reason to hope that a modest deterrent would be strong enough.

Original position deliberators would, however, need to be certain to avoid a deterrent so weak that it would amount to funishment and the consequent collapse of the law and order which society needs to survive, since the alternative to society is the state of nature, with its war of all against all. Original position deliberators would select principles of punishment allowing enough sensitivity to actual social situations to calibrate a deterrent unpleasant enough to ensure this. It seems clear, however, that original position deliberators would not risk imprisonment that deterred *more strongly* than necessary to avoid funishment and the state of nature, since they must assume that they are as likely to live in the conditions that create that deterrence as they are to live under the protection of that deterrence. It should be emphasized that unpleasantness is often a relative matter—we avoid a less-pleasant alternative when we have a more-pleasant alternative available, even if the less-pleasant option is not intrinsically unpleasant. The minimum level of unpleasantness necessary to avoid funishment would be relative to the conditions of life outside prison, and the prioritization of non-coercive preventative measures demanded by original position deliberators (such as education, jobs, and social services) suggests that imprisonment would not have to be intrinsically unpleasant for it to be less pleasant than life on the outside. It is for this reason that I think that original position deliberation sets a limit on harsh treatment of prisoners at the same level that I take to be required by the quarantine justification.

How should we put these principles into practice? If original position deliberators have good but not perfect empirical knowledge of human nature, then I think they would select a program of rapid improvements in distributive justice toward the Rawlsian ideal, with matching but slightly slower improvements in prison conditions, and heavy spending on empirical research to determine the point at which deterrence begins to decay, so that they could slow the improvements in prison conditions, or stop them if needed, before arriving at funishment.

If my reasons for choosing principles of punishment in the original position include deterrence, then this approach to punishment has a consequentialist element. But this does not imply that it is a species of consequentialist justification. The premise of this approach is that what we consent to in the original position is just, not that punishment ought to be aimed at achieving any particular outcome. I take this to be a deontological premise. Whatever consequentialism derives from original position deliberation is the result of working out the implications of this deontological premise. That is, if this line of thought is sound, it would be rational to weigh consequentialist considerations in consenting to principles of punishment, but it is legitimate for these considerations to play a role in justifying punishment only because they emerge from our rational consent. The consequentialist considerations have no independent value. This justification of deterrence implies using criminals as *means*, but it does *not* imply using them as *mere means.* Criminals can be used as means to the end of deterrence without being used as mere means if they would rationally consent to being used in this way. Translated through Rawls, the claim is that criminals can be used as means to the end of deterrence without being used as mere means if, in the original position, they would choose an institution of punishment that included deterrence.

Now let us consider whether this approach sets suitable deontological limits on how people caught up in the penal justice system can be treated. Earlier in the paper I mentioned three disturbing forms of instrumentalization that utilitarianism, and arguably all pure consequentialisms, must endorse under certain circumstances: (i) punishments of unlimited severity; (ii) violating or weakening due process; and (iii) framing and punishing non-criminals. Can the personhood-based approach defended here rule out these practices?

Let us consider limitations on the severity of punishment first. Can we justify imprisonment under harsh conditions in order to strengthen deterrence? Original position deliberators would resist this, since they face an equal probability of a significant harm and a smaller benefit. The same would even more obviously be true of the death penalty and torture. As mentioned earlier, original position deliberators would not risk imprisonment that deterred *more strongly* than necessary to maintain a society of law and order. Rejecting the avoidability of crime assumption is crucial in establishing this limit. If we assume that we can avoid crime, and thereby avoid liability to punishment, it could seem rational in the original position to endorse very severe punishments in order to strengthen deterrence (at least if we have strong institutions of due process in place).

Next, can this approach explain why we should not weaken or violate practices of due process when the utilitarian calculus shows that doing so would reduce overall suffering? On this point, rejection of the avoidability of crime assumption may appear to be more of a liability than an advantage. That is, if we assume that we can avoid committing crimes, then it makes sense to insist on a very strong institution of due process, because this will allow us to be confident that by refraining from crime we can escape punishment. Free will deniers need a different solution. As argued before, original position deliberation about principles of penal justice requires us to distinguish the people harmed and the people benefited. Previously, the focus has been on choosing the *principles of punishment,* that is, on determining how we should treat people who have *already been selected* for punishment. But now, we are choosing the *principles of due process*: in other words, we are choosing the rules for *determining who we should punish.* So our competitors are not the people punished versus potential victims, but instead the accused versus potential victims. We must assume that we are just as likely to be the person accused of a crime as we are to be the potential victim. It can be shown that an individual accused of a crime (whether correctly or incorrectly) has more to lose from a weakened institution of due process than an individual potential victim has to gain from it. Suppose that the criminal conviction standard were to be lowered from "reasonable doubt" to something weaker. This would allow prosecutors to convict more of the accused, thereby worsening things for the accused. Some of the additional people convicted will have been correctly accused, and their conviction will result in an improvement for potential victims. But the lowered standard will also open the door to sloppy or politically motivated prosecutions that result in the conviction of non-criminals, thereby worsening things for the accused without an equivalent improvement for potential victims. So it would worsen things for the accused more than it would improve things for potential victims, and since I must assume that I am just as likely to be in either position, I would not choose to weaken due process.

Now let us turn to the issue of framing and punishing non-criminals. Suppose that we could dramatically strengthen general deterrence by occasionally framing and punishing celebrities, given all the publicity involved. In Kantian terms, an institution which aims to deter by way of penalizing anyone other than real criminals can only succeed through a systematic and global deception of the public which contradicts itself. That is, to choose a principle of punishment that allowed punishment of a framed celebrity instead of a real criminal for the sake of general deterrence, I would also have to choose that the overwhelming majority of the population be deceived about the fact that this principle was in effect. The deception would be necessary because if word got out that scapegoats were sometimes punished instead of real criminals, then the extra deterrent force which authorities had hoped to achieve with the framing would be destroyed. Since I could not assume that I would not be among the deceived, I would in effect be volunteering to be deceived about the principle I had chosen. In other words, I would be volunteering to be a mere means to the end of amplifying deterrence. Consenting to be deceived about the basic principles of one’s society undermines one's status as a rational agent in a way that parallels consenting to slavery, and should be seen as self-contradictory for parallel reasons. So we could not rationally consent to an institution of punishment that punished anyone but real criminals.

5. Objections and Responses

I would like to move toward a conclusion by responding to some objections from Pereboom (2014: 160-161 ftn). He accepts that it may be right to claim that because we are persons, we deserve to live in a society regulated by the principles we would choose in the original position, but he is skeptical about using this idea to justify the claim that criminals deserve to be punished. However, the purpose of my discussion of desert is to analyze a notion of desert sometimes taken to be monolithic in the free will literature. As explained above, I distinguish personhood-based desert from action-based desert, and then on the basis of this distinction, I make a more specific claim: that criminals have personhood-based desert claims which give them absolute deontological rights to certain standards of treatment, and that it is nonetheless possible to punish them without violating their personhood-based desert claims. The point of appealing to desert is to argue that even in the absence of action-based desert, there is another kind of desert which continues to protect the rights of criminals to be treated in particular ways.

Pereboom goes on to express a more specific concern about how my view justifies deterrence:

> We might well agree in the original position that if there was no other way to save 100 million or more people than by killing one innocent person, we should do so. But it seems implausible that this one person would then deserve to be killed. In addition, I don’t think that just because we’ve all hypothetically agreed to this resolution, the innocent person isn’t being used as a mere means when she is killed. On the contrary, it’s highly intuitive that she is. This worry carries over to a general deterrence scheme agreed upon in the original position. (2014: 160-161, ftn.)

Suppose that we contemplated the “kill one innocent to save 100 million” principle in ignorance of whether we would wind up as the one innocent person killed, or among the 100 million saved. In the act of social cooperation in which she might be killed to benefit the rest, the one innocent would be easy to recognize as the worst-off party, and original position deliberation would direct us to assume that we would find ourselves in her shoes when the veil of ignorance was lifted. So, *contra* Pereboom’s claim above, we would reject the “kill one to save 100 million” principle and we would not be compelled to endorse the use of her as a mere means. Of course (and as Pereboom would no doubt add), it will sound rigoristic and bizarre to many to suggest that we could have a moral reason to refrain from killing one and let so many die. Two points are important here. First, as I explained earlier, my goal in drawing on original position deliberation in the ethics of free will skepticism is to incorporate the non-consequentialist notion of rational choice which allows Rawls to model Kantian absolute duties to persons which cannot be overridden by considerations of aggregate well-being. Kantian duties are demanding. Kant famously rejects the saying that "It is better for *one* man to die than for an entire people to perish," claiming that "if justice goes, there is no longer any value in human beings' living on the earth" (Kant 1996, 161). Second, from a broadly Kantian perspective, I think the claim to make about the moral reasons relevant for the one innocent is this: she has a *right* not to be killed by us to benefit others, but she can nonetheless *fulfill an imperfect duty of virtue* if she sacrifices herself for the benefit of the rest. (Kant himself does not argue that self-sacrifice can be an imperfect duty of virtue, but ethicists working in the Kantian tradition could accommodate such a duty.) The distinction between duties of right and imperfect duties in Kantian ethics allows us to hold, in cases like this, that the moral reasons in favor of the one innocent's death are hers to voluntarily accept, not ours to coercively enforce, and if she voluntarily fulfills a duty in this way she is not used as a mere means. I appeal to Rawlsian methodology for help in explaining our rights against one another when harms are at issue, but I do not mean to claim that it provides a complete moral theory. I should emphasize that I do not take myself to have provided anything like a sufficient case for the claim that this Rawlsian/Kantian approach to the ethics of free will skepticism is preferable to a consequentialist approach. My goal has been merely to argue that such an approach is compatible with free will skepticism, and that it solves certain problems in a way that makes it worthy of consideration.

A more indirect response to Pereboom's concern derives from a more general feature of the view of social contracts explained earlier, that criminals are not contract-breakers cast back into the state of nature, but instead equal participants in the social contract. I take it to be a basic assumption of social contract theory that it can only be rational for us to join the contract if it gives us good reason to believe that we will fare better in society than we would fare in the state of nature. This on its own implies that I cannot contract for the possibility of being killed, because that outcome is no better than the state of nature. Further, I think that even the worst positions in society must be better than the state of nature for us to rationally join the contract, even those of prisoners. It seems plausible to think that we can imprison people under conditions calibrated to provide just enough deterrence to avoid funishment which are nonetheless preferable to the state of nature. This point adds another perspective to my defense against the mere means objection. That is, even imprisoned people gain something in return for what they contribute to society. A society that includes people serving as means to the end of deterrence is the only alternative to the state of nature, given the limited social and technological tools for abolishing punishment currently at our disposal. The people punished serve as means to the end of deterrence, but the other members of society serve the people punished as means to their end of living in better conditions than the state of nature affords. This may sound appalling. Don't we all deserve more from society than a life which is merely better than the state of nature? My answer is yes, of course—we deserve a life to which we would consent in the original position. The other members of society must serve prisoners not just as means to their end of a life better than the state of nature, but also as means to their end of a life they would accept in the original position. I am not suggesting that it is sufficient to justify a form of punishment that it passes this better-than-the-state-of-nature test, only that it is necessary. Forms of punishment chosen in the original position will pass the better-than-the-state-of-nature test, but forms of punishment that pass the better-than-the-state-of-nature test may not always be acceptable in the original position, because they may punish more harshly than necessary to avoid funishment, or because they may allow practices like violations of due process.

Nothing I have said is meant to deny that there remains a tragic kind of unfairness in the incarceration of criminals for free will deniers, given that they have done nothing to deserve their situation. But acknowledgment of that tragic unfairness isn't a reason to be less resistant to the kinds of unfairness involved in maximally painful punishment, violations of due process, and framing. The main purpose of the account I defend is to explain how free will deniers can justify punishment despite establishing deontological rights against these kinds of unfairness.

References

Caruso, Gregg. 2019. Public Health and Safety: The Social Determinants of Health and Criminal Behavior. *Justice Without Retribution,* eds. ???????

Dolovich, Sharon. 2004. Legitimate Punishment in Liberal Democracy. *Buffalo Law Review* 7(2): 314-29.

Harsanyi, John. 1975. Can the maximin principle serve as a basis for morality? A critique of John Rawls's Theory. *The American Political Science Review* 69(2): 594-606.

Kant, Immanuel. 1996. *The metaphysics of morals,* trans. and ed. Mary Gregor. New York: Cambridge University Press.

Levy, Neil. 2012. Skepticism and sanction: The benefits of rejecting moral responsibility.

Law and Philosophy 31(5): 477-493.

Pereboom, Derk. 2014. *Free will, agency, and meaning in life*. New York: Oxford University Press.

Pereboom, Derk. 2001. *Living without free will.* New York: Cambridge University Press.

Pereboom, Derk. 2013. Free will skepticism and criminal punishment. In *The Future of Punishment*, ed. Thomas Nadelhoffer, pp.49-78. New York: Oxford University Press.

Smilansky, Saul. 2011. Hard determinism and punishment: A practical reductio. *Law and Philosophy* 30 (3): 353-367.

Vilhauer, Benjamin. 2017. Kant's mature theory of punishment and a first *Critique* ideal abolitionist alternative. In *Palgrave Kant Handbook*, ed. Matthew C. Altman, pp. 617-642. New York: Palgrave Macmillan.

Vilhauer, Benjamin. 2009. Free Will Skepticism and Personhood as a Desert Base. *Canadian Journal of Philosophy* 39(3): 489-511.

Vilhauer, Benjamin. 2015. Free will and the asymmetrical justifiability of holding morally Responsible. *Philosophical Quarterly* 65(261): 772-789.

Vilhauer, Benjamin. 2013. Persons, punishment, and free will skepticism. *Philosophical*

Studies 162(2):143-163.