Diametros (2024), 79: 91–106
doi: 10.33392/diam.1928

Free Will Denial, Punishment, and Original Position Deliberation

– Benjamin Vilhauer –

Abstract: I defend a deontological social contract justification of punishment for philosophers who deny free will and moral responsibility (FW/MR). Even if nobody has FW/MR, a criminal justice system is fair to the people it targets if we would consent to it in a version of original position deliberation where we assumed that we would be targeted by the justice system when the veil is raised. Even if we assumed we would be convicted of a crime, we would consent to the imprisonment of violent criminals if prison conditions were better than the state of nature but deterring enough to prevent the state of nature.

Keywords: punishment, free will skepticism, original position, retribution, Kant, Rawls, deterrence
Submitted: 26 June 2023
Accepted: 4 November 2023
Published online: 19 April 2024

1. Introduction

I defend a deontological social contract justification of punishment for philosophers who deny free will and moral responsibility (FW/MR). Even if nobody has FW/MR, a criminal justice system is fair to the people it targets if we would consent to it in a version of original position deliberation where we assumed that we would be targeted by the justice system when the veil is raised. Even if we assumed we would be convicted of a crime, we would consent to the imprisonment of violent criminals if prison conditions were better than the state of nature but deterring enough to prevent the state of nature. And even if we assumed we would be accused of a crime, we would consent to an evidentiary standard low enough to allow a conviction rate sufficient to prevent the state of nature. After explaining this justification, I contrast it with the public health quarantine justification defended by Derk Pereboom and Gregg Caruso, and argue that the former provides a stronger response to the *using criminals as mere means* objection. Finally, I address the objection that denying FW/MR entails moral nihilism.

Benjamin Vilhauer
City University of New York
City College and Graduate Center
North Academic Center 5/144
160 Convent Ave
New York, NY 10031
United States
E-mail: bvilhauer@ccny.cuny.edu

2. A Deontological Punishment Contract for Deniers

The deontological social contract justification is founded in broadly Kantian and Rawlsian ideas about justice. These ideas are only broadly Kantian and Rawlsian, as they diverge from the views of Kant and Rawls in important ways.[2] But they have a deontological core which distinguishes the justification presented here from other justifications of punishment available to those who deny FW/MR ("deniers" henceforth), which typically have consequentialist foundations.[3] Social contract thinking plays a crucial role in formulating the justification, but it also includes moral principles which are *prior* to the contract and constrain the considerations to which we may appeal in contract bargaining.

The prior principles derive from a denialist view according to which there is a *personhood-based* kind of desert which does not depend on FW/MR in the way *action-based* desert does.[4] In other words, there are some ways we deserve to be treated just by virtue of being persons, and we deserve to be treated in these ways irrespective of any claims that might be made about how we deserve to be treated based on our actions. On this view, personhood-based desert is as metaphysically and metaethically fundamental as action-based desert, and its distinctness from action-based desert means that we deserve things just by virtue of being persons even if we lack FW/MR. Thus the denial discussed here does not undermine personhood-based desert.[5]

Derk Pereboom and Saul Smilansky disagree with this view of personhood-based desert.[6] Pereboom acknowledges different kinds of desert, but holds that free will is required for what he calls *basic* desert, and holds that desert for which free will is not required (which he thinks can be derived from consequentialist considerations) is *non-basic*.[7] Thus his view appears to be that if we lack free will, then nobody fundamentally deserves anything. Smilansky holds that *non-trivial* desert claims are action-based (or "responsibility-based," in his terminology). He allows that there is a minimum baseline of well-being we deserve just because we are persons, but argues that when we fall below it, our claims to deserve restoration of that baseline must be action-based.[8] He

[2] Kant (1996); Rawls (1999).

[3] Some think we ought not call a practice "punishment" if it does not include retribution in its justification. I think it is hazardous euphemism to avoid this term when we seek to justify coercive curtailment of wrongdoers' liberty, whatever our justification may be.

[4] I do not endorse denial, but instead a skepticism according to which it is possible that we have FW/MR, but we do not know if we do. However, I think skeptics are mostly in the same boat as deniers when it comes to criminal justice, since retribution is unjust if we don't know we have FW/MR, so I leave this distinction out of the main text. Let me note in passing that I think even deniers should acknowledge moral reasons for wrongdoers to feel remorse, since remorse is part of understanding one's wrongs, and can involve empathic care instead of self-retribution (Vilhauer 2023b).

[5] See Vilhauer (2009 and 2013) for more on personhood-based desert.

[6] Kant disagrees too. He thinks FW/MR is necessary for personhood (Kant 1996: 378). But his texts also offer ideas for the revisionist Kantian view advocated here. He says, for example, that "every rational being, as an end in itself, must be able to regard himself as also giving universal laws with respect to any law whatsoever to which he may be subject; for, it is just this fitness of his maxims for giving universal law that marks him out as an end in itself" (Kant 1996: 87). If it is just this that marks him out as an end in himself, and being an end in himself suffices for being a person, then we have a definition of personhood which does not refer to FW/MR.

[7] Pereboom (2014): 2, 126–152.

[8] Smilansky (1996): 159.

illustrates this view with an example about air pollution. Smilansky argues that one deserves clean air just because one is a person, but if one's air is polluted, one deserves compensation *only if there is nothing one has done to deserve* the pollution — that is, only if the pollution does not derive from one's own culpable actions, e.g., negligence in burning too much garbage. Smilansky thinks the relationship he claims to find in this scenario between personhood-based desert and action-based desert generalizes, such that personhood-based desert can do no non-trivial moral work unless accompanied by action-based desert claims.

Smilansky's generalization does not hold, however, because there are things we deserve as persons which we cannot cease to deserve no matter what we do. Any human rights which are universal and cannot be alienated or forfeited are rights we deserve to be accorded no matter how we act. Many laypeople as well as philosophers hold that there are such rights. It is obviously well beyond the scope of this paper to try to *prove* that there are such rights. Philosophers have debated about them for centuries, and many of the questions about them have little to do with FW/MR. The way to understand the argument of this paper is as an exploration of the implications of such widely-accepted rights for denial and punishment. Personhood understood as a desert base which is metaphysically and metaethically independent from action is the only plausible grounding for such rights. Since deniers must reject action-based desert, they must not treat targets of the criminal justice system as if they deserve to suffer for their crimes. But the independence of personhood-based desert means deniers must still treat targets in accordance with rights they deserve based on personhood.

Some may object that it is a distortion of the concept of *desert* to talk of deserving rights — that the correct concept here is *entitlement.* There are three problems with this objection. First, laypeople as well as philosophers *do* talk of deserving rights. Second, "entitlement" is often used for things people claim not for fundamental moral reasons, but instead for merely legal reasons which are morally arbitrary or even immoral. If I have a legal contract that says I own the vegetables you have grown, I am in this sense entitled to the vegetables even if your children are hungry and mine dine at gourmet restaurants. Third, the objection misses the point. What is important for the present argument is the idea that there are personhood-based claims that derive from fundamental deontological principles which are independent of FW/MR and constrain punishment.

The deontological social contract justification (DSC) does not claim that people deserve to be punished just because they are persons. Instead, it claims that punishment is fair if it does not violate the deontological constraints imposed by personhood.[9] Personhood-based desert constrains punishment, but does not set an aim for punishment. As explained below, the aim is incapacitation, along with whatever general deterrence may be needed to avoid the state of nature, but this aim only has normative significance because it arises from the rational consent of the punished. Thus the justification is fundamentally deontological despite having a consequentialist element.

[9] The DSC can acknowledge that even when its criteria for just punishment are met, a kind of unfairness remains: an unfairness rooted not in human injustice, but in the tragedy of human nature and the limits of the world. It shares features with the unfairness according to which some are afflicted with illness when others are not.

How exactly do targets of the criminal justice system deserve to be treated just by virtue of being persons? It is at this point in the story that the justification draws on contractualism. I take it to be straightforward to prompt the intuition that there are personhood-based desert claims, but I do not claim that this intuition brings with it an immediately clear inventory of our personhood-based desert claims. Historically, laypeople as well as philosophers have typically believed in action-based desert, so there has been no pressure to disentangle action- and personhood-based desert, and there is no tradition to draw on. I rely on a modified version of Rawlsian original position deliberation (OPD) to illuminate what we deserve from the criminal justice system based on personhood rather than action. OPD is sometimes criticized on the grounds that it is hard to be very confident about what deliberators would choose in OPD, so theorists can smuggle in assumptions that reflect their biases to get the outcomes they want. Here are three reasons it is worth exploring despite this hazard. First, social contract thinking offers the prospect of answering the mere means objection directly and in its own terms, because of the way rational consent demonstrates that we do not use people as mere means. Second, *actual* consent theories of rational consent (typically in the form of *implicit* consent theories in the context of punishment) are vulnerable to the objection that actual consent is morally significant only when agents with FW/MR *act* to give consent, in a way that *hypothetical* consent theories like Rawls' are not (see section 4 for more on this). Third, OPD should be of interest to deniers because Rawls himself recommends it partly because it screens out undeserved inequalities.[10]

Rawls applies OPD to distributive justice, not criminal justice. The DSC extends OPD to criminal justice.[11] The DSC is conditioned on Rawlsian distributive justice — a criminal justice system chosen in OPD could only be implemented in society in a fully justified way if the distributive justice system chosen in OPD is fully implemented in society too.

In OPD, deliberators use maximin reasoning to select principles that make the minimum (worst-off) social position as good as possible.[12] A disanalogy between distributive and criminal justice complicates DSC maximin reasoning. There is only one minimum position in distributive justice – the poorest. Criminal justice has two candidates for minima, crime victims and the punished, and they compete for the worst-off position. If punishment deters, then calibrating punishment to improve one position worsens the other. Social and technological innovations might someday eliminate this competition. Perhaps society will someday be ruled by laws so just that legitimate civil disobedience will be impossible. Perhaps swarms of soft, sagacious nanodrones will apply persistent gentle pressure to the limbs of everyone who begins to commit a crime, overwhelming

[10] Rawls (1999): 86–89.

[11] OPD justifications of punishment are found in Murphy (1973), Sterba (1977), Clark (2004), and Dolovich (2004), but they rely on retributivist ideas which deniers cannot allow. Clark claims to offer a non-retributive Kantian approach to punishment, but allows what he calls "negative retributivism" in his account of why we should punish the guilty rather than the innocent, and denial rules this out. My approach is based on a more general Rawlsian approach to social justice for deniers in Vilhauer (2009).

[12] Rawls defends the risk-aversiveness of OPD as rational under uncertainty. This has been disputed (e.g. Harsanyi 1975). However, Rawls also offers a separable defense of OPD as conforming to moral facts that are prior to contract bargaining. I endorse the latter defense and am agnostic about the former.

them with fatigue before they succeed. But if criminals and victims will compete for the foreseeable future, then deniers must apply OPD in a way that is *fair* in view of the fact that neither party deserves to suffer the consequences of crime based on their actions. Rawls does not address such competition, so his account offers no guidance about how to understand fairness here. One intuitive way of thinking about fairness in competition is in terms of having an equal chance of being harmed and benefited.[13] Thus, in previous work, I explored a model of fairness in OPD bargaining about criminal justice which assumes a 50/50 chance of finding ourselves among those targeted by the criminal justice system, and among potential victims, when the veil is raised.[14] I continue to think this approach has merits. Here, however, I will consider a model of fairness in which we are *certain to be targets* of the criminal justice system (to be convicted or accused of crime) when the veil is raised.[15] This allows OPD deliberators to work with only one minimum position, the position of the punished.

How could it be fair to assume that we will certainly be targets when the veil is raised, given that victims suffer terribly too, and compete with targets? After all, a guiding principle in OPD is to bargain solely on the basis of self-interest. Assuming we will be targets is in effect assuming that targets' self-interests matter but victims' self-interests don't matter.

Two points about Rawls' conception of self-interest help us respond to this concern about fairness. First, OPD uses self-interest as a heuristic device — it does not claim that the only moral interests are self-interests.[16] OPD focuses on the self-interests of people in the minimum position to ensure that the principles selected are fair to people in the minimum position, because the social contract as a whole is fair only if it is fair to expect everyone to consent. When the veil falls, it is often *not* in our self-interest to follow the principles we choose behind the veil — when the veil falls, those principles are revealed as obligations which we ought to accept for moral reasons, even if reluctantly. Second, OPD involves a moralized and rationalized conception of self-interests, which Rawls explains as interests in "primary goods…answering to [our] needs as citizens as opposed to [our] preferences and desires."[17] Primary goods include rights, liberties, opportunities, and income and wealth.[18] Perhaps even malicious desires give people interests, such that serial killers' desires to kill give them malicious interests in killing, but malicious interests are not weighed in OPD. Since the DSC appeals to contractarian thinking to unpack the implications of moral commitments which are prior to the contract, the moralization of interests in OPD does not burden the DSC with additional commitments.

However, the point that self-interest functions as a moralized heuristic device does not suffice to explain how OPD that weighs self-interests of targets but not victims

[13] Rawls himself rules out probabilistic reasoning about social outcomes behind the veil, but his target is probabilistic thinking based on assumptions about empirical distributions of outcomes in society. If probabilistic reasoning is needed in the DSC, it must be derived from *a priori* moral ideas about fairness under competition.

[14] Vilhauer (2013; 2023b).

[15] I was prompted to explore this new approach by John Lemos' critique of the previous approach (Lemos 2023: 177–192).

[16] Rawls (1999): 128.

[17] Ibidem: xiii.

[18] Ibidem: 79.

could be fair. We need another key point to explain why it could be fair. When we decide to punish people, we intentionally impose serious harm, and intentionally imposed serious harm must meet a very high standard of justification. It might seem that there is a parallel imposition of harm on victims when the criminal justice system fails and crimes are committed. We probably have a duty to design a criminal justice system that prevents victimization, and the occurrence of crime probably implies that we are failing to fulfil this duty and in this sense *allowing* people to become victims. But the distinction between *intentionally doing* and *allowing* is important in deontological ethics. Criminals are the ones who intentionally harm victims. Law-abiding members of society do not intentionally harm victims through the justice system.

The clearest way to be confident that we are justified in intentionally harming people is to have an account that the people harmed would accept themselves if they thought rationally about it, on the basis of their own reasons and nobody else's. Many doubt that consequentialist reasons of general deterrence suffice for such an account. Why should I accept that I should be harmed merely as a means to benefiting others? Retribution plays a key role in providing such an account in most theories of punishment. If there are reasons of retributive desert, rational wrongdoers must acknowledge that they deserve harm. But deniers reject retributive desert. Reasons of self-defense are another candidate, and they are important in the Pereboom/Caruso public health quarantine justification discussed later. But reasons of self-defense become weaker and vaguer when offenders' immediate threats have been neutralized and they are under our power, and it is at this point that the hard questions about punishment arise. The DSC's reason is OPD consent. We can be more confident that OPD fairly models the consent of the punished if OPD weighs only the interests of the punished. If OPD weighs the interests of the punished *and* the interests of those benefited by punishment, then the DSC is confronted with an objection parallel to the mere means objection: why should the interests of people who would *benefit* from my punishment give *me* a reason to consent to punishment? I do not claim to have established that the interests of those protected by punishment do not matter at all for fairness. The point is rather that we can be especially confident that we are not imposing unfair harms on the punished if we weigh only their interests in OPD. Since intentionally imposed harm must meet an especially high justificatory standard, it is worth exploring an account that makes us especially confident that the harms we are imposing are fair. Below, I apply OPD in which deliberators assume they will be targets of the criminal justice system to two kinds of targets: first, people convicted of crimes, and second, people accused of crimes.

First, how would deliberators think about punishment if they assume they will find themselves convicted when the veil is raised, and they will be among the punished if any are punished? Will they consent to punishment? They will begin by interrogating the assumption. As mentioned, someday technological and social innovations may prevent all crime and thereby abolish punishment. From our standpoint outside OPD, we know our society is not post-crime. But in OPD, deliberators are ignorant about the social and technological development they will find when the veil falls. They can hope they will arrive post-crime, and will choose a principle that facilitates abolition. We might worry that deliberators in fear of punishment would seize upon authoritarian

measures that suppress crime but also liberty. But OPD prioritizes liberty in the primary goods pursued by deliberators, so it rules out authoritarian measures. As explained, OPD is moralized in a way that deniers should endorse. Instead, deliberators would choose a principle of investing as much social energy as possible in non-authoritarian crime reduction measures as is compatible with meeting their other basic needs. Call this the *abolition principle*, since it aims at the abolition of punishment as an ideal that may someday become actual. Outside OPD, in today's society, this principle implies an obligation for social spending on non-authoritarian crime control. It may be beyond the resources of the DSC to specify a spending plan with precision, but it would support the development of innovative non-authoritarian social and technological crime-control strategies, and pour money into already-available non-authoritarian crime reduction measures: more jobs, education, public services, and voluntary therapy for people at risk of committing crimes. This adds a supplement of social resources for people at risk of committing crimes onto the share allocated to them by Rawlsian distributive justice.

However, deliberators would acknowledge that they may not arrive post-crime. They know enough about human psychology[19] to know that a society with uncontrolled crime is likely to have rising crime rates, and that rising crime rates can eventually lead to a collapse into the state of nature and its war of all against all. Would they consent to some form of punishment to control crime? Consenting would be in their interest if punishment were preferable to the state of nature and necessary to prevent the state of nature. Let us assume for the sake of argument that the penalties they consider for societies which are not post-crime are the ones available today, such as execution, imprisonment, and assorted lesser penalties like fines.

Deliberators would not prefer death to the state of nature, so they would not consent to a punishment system that includes execution as a penalty.[20] Deliberators' knowledge of psychology would tell them that non-violent crime is much less threatening to social order than violent crime, and they would be reluctant to consent to imprisonment as a penalty for non-violent crime. But non-violent crimes erode social order if unconstrained. Lesser penalties like fines proportioned to wealth would provide significant constraints. Since fines are clearly preferable to the state of nature, it is reasonable to assume that deliberators would consent to them to control non-violent crime.

Deliberators' knowledge of psychology would also tell them that fines would not suffice to control violent crime enough to prevent the state of nature. Deliberators would therefore endorse imprisonment as a penalty for violent crime if imprisonment could offer prison conditions preferable to the state of nature while still preventing the state of nature. Deliberators would demand prison conditions better than those in contemporary U.S. prisons, because they are probably not better than the state of nature. Deliberators

[19] Ibidem: 119.

[20] Kant endorses the death penalty for murder (Kant 1996: 474), because he is a retributivist and endorses a principle of retributive equality (ibidem: 473). He appeals to retributivism to deny criminals a "voice" in the social contract (ibidem: 476). He is wrong to be a retributivist, because he acknowledges that we lack theoretical knowledge of the transcendental freedom that would make us morally responsible (ibidem: 95), and his practical argument that we can infer from knowledge that we ought to act in certain ways to knowledge that we are transcendentally free to do so (ibidem: 163–165) is problematic. See below and Vilhauer (2024) for discussion.

would want humane and secure conditions that included education, meaningful work, voluntary therapy, excellent healthcare, regular visits from friends and loved ones, continual parole review to determine whether prisoners could be released without undue risk of repeated violence, and ongoing post-release support to help people avoid new violence.

In the interest of fairness to the targets of the criminal justice system, let us assume that deliberators choose principles governing imprisonment based upon the assumption that they will find themselves imprisoned when the veil is raised. Deliberators would want prisons' main goal to be the comfortable incapacitation of violent offenders, not general deterrence. But they could not ignore general deterrence altogether. Saul Smilansky has argued that deniers must acknowledge that it is unjust to imprison people who do not deserve imprisonment based on their actions, and must therefore acknowledge a duty to compensate them heavily during their imprisonment.[21] He argues that this compensation would make prison conditions so pleasant that they would become an incentive to commit crime. While I do not agree with the entirety of this argument, I think it shows that deniers cannot assume that prison conditions inevitably deter, and that deniers need a moral reason to calibrate prison conditions to provide a deterrent sufficient to prevent imprisonment from becoming an incentive to commit crime. Some deniers seem to see the hospitable prisons of Norway as counterexamples to Smilansky's argument.[22] But this misses his point. Even Norwegian imprisonment is a profound deprivation of liberty. If we reflect seriously on what we are inflicting on prisoners, we ought to ask not only whether we should compensate them heavily during their imprisonment, but also afterwards if they are released, when they must reintegrate into society and need support. After intentionally depriving them of so much liberty, we should give them a trust fund to help them if they can never hold down a job again. How much should we put in the fund? Just enough to get by, or to be comfortable, or to live in luxury? If deniers appreciate the deprivation we inflict with imprisonment, they should see that it is not confused to wonder if we might owe them luxury. But it should not be hard to see that luxury trust funds might create incentives to commit crime. If we take ethics seriously, we cannot arbitrarily decree a limit to the desirability of imprisonment. We need a moral reason to draw a line.

The moral reason offered by the DSC is OPD consent. OPD deliberators' knowledge of human psychology would make them worry that excessively pleasant prison conditions would accelerate a collapse into the state of nature rather than preventing it. If their psychological knowledge is no more complete than ours, they would of course see this as a matter requiring further empirical study. Potential criminals' motivations might turn out to be too causally isolated from prison conditions for limits on the desirability of prison conditions to make a difference. So they would not choose a principle that *required* prison conditions to be unpleasant if further study showed this had no effect. But they would choose a principle that *permits* calibration of prison conditions for deterrence if it turns out that limits on prison desirability *do* deter — a principle that permits calibration to a level of deterrence sufficient to avoid creating an incentive for crime, but

[21] Smilansky (2011).

[22] This point is in response to an objection from a reviewer.

no higher. It is crucial to emphasize that conditions would not have to be unpleasant in any absolute sense to provide substantial deterrence, only unpleasant relative to life outside prison.[23] Under the social conditions required by the DSC, life outside prison would be much better for people at risk of offending than it is today, for two reasons mentioned earlier. First, OPD criminal justice is conditioned on OPD distributive justice, so it entails that imprisonment could only be fully justified when OPD distributive justice is achieved, and the incentives for crime which derive from distributive injustice are absent. Second, the abolition principle supplements the social resources allocated to people likely to offend, funding jobs, education, public services, and voluntary therapy. These social conditions make life better for likely offenders outside prison, so it can get better inside while still deterring. It seems quite plausible that, under these conditions, general deterrence sufficient to avoid the state of nature could be maintained despite making prison conditions dramatically better than the state of nature.

General deterrence inevitably uses the punished as *means,* but their consent demonstrates that they are not used as *mere* means. Further, the punished and the people protected by punishment use each other as means *reciprocally*. Both parties seek a life better than the state of nature. The protected pursue this end by using the punished as means to general deterrence. The punished pursue this end by using the protected as means for generating the social resources necessary to provide the best prison conditions compatible with general deterrence and sufficient to prevent the state of nature. Thus there is a sense in which the competition between the punished and protected is not fundamental after all.

Next, consider the standard of evidence for conviction. Here, to be fair to targets, we must ask deliberators to make a different assumption: that they have been *accused* of a crime but not yet *convicted*. Will they consent to a standard of evidence that makes conviction possible? They would prefer not to be convicted, so it might seem that they would not. But they will understand the need for crime control to prevent the state of nature in the same way the deliberators considered earlier do. So it would be irrational for them to choose a standard of proof that is impossible to satisfy. They would choose a principle that sets the standard low enough to allow a conviction rate sufficient to prevent the state of nature, but no lower. The *proof beyond reasonable doubt* standard, as it is applied in the nations that endorse it, allows enough convictions to prevent the state of nature. So deliberators would not choose a standard lower than reasonable doubt. Would they choose a standard higher than reasonable doubt? Some interpretations of the reasonable doubt standard make it extremely high. Justice Brennan, in *In re Winship*, says it demands "utmost certainty."[24] Laurence Tribe says it requires "as close an approximation to certainty as seems humanly attainable in the circumstances."[25] On Alex Stein's interpretation, "[a] legal system may justifiably convict a person *only if it did its*

[23] It of course does not make sense to aim at *perfect* deterrence, since that will not be possible without social and technological innovations like those that would make for a post-crime society. In circumstances like those of contemporary society, we must assume that the motivations of at least *some* potential criminals will be too causally isolated from prison conditions for such conditions to make a difference, due e.g., to ignorance, mental illness, or the strength of their desires to commit crime.

[24] US Supreme Court (1970): 364.

[25] Tribe (1971): 1374.

best in protecting that person from the risk of erroneous conviction and if it does not provide better protection to other individuals."[26] These remarks suggest that the reasonable doubt standard demands a probability close to 100%. It is hard to see how a higher standard could allow enough convictions to prevent the state of nature. So deliberators would not choose a higher standard.

3. The Quarantine Justification and the Mere Means Objection

The mere means objection is an important objection to consequentialist justifications of general deterrence.[27] If our rationale for punishing criminals is just that it produces the good consequence of general deterrence, we are using criminals as mere means to social ends. The DSC has a strong response to this objection. The Kantian test for whether we treat someone as a mere means is whether they would rationally consent to the treatment. The argument that we would rationally consent to punishment even if we assumed we would be convicted forms the core of the DSC. This response is stronger than the one offered by the public-health quarantine justification, which is defended by Derk Pereboom and Gregg Caruso in joint papers as well as independent papers and books.[28] The prison conditions advocated by Pereboom and Caruso have a great deal in common with those advocated here, indeed, so much that one might wonder whether their principles of imprisonment might be those selected in the original position. If this is right, then the DSC might serve as a deontological scaffolding for their view, and it might be a useful and interesting scaffolding to have even if that were its only contribution. But I do not think this is the case, because the views handle deterrence and the mere means objection in different ways.

Pereboom and Caruso hold that the right to self-defense justifies quarantining carriers of dangerous diseases despite the fact that they do not deserve to be sick. They draw an analogy between quarantining the sick and imprisoning the violent, and argue that we have as much right to imprison the violent as we do to quarantine the sick, even if the violent do not deserve imprisonment. As I understand their view of ethics as a whole, it is predominantly consequentialist. Smilansky agrees.[29] However, they hold that the quarantine justification is *not* consequentialist because they hold (A) that the right to self-defense need not be understood consequentially, and (B) that their justification offers sufficient protection of criminals' rights not to be treated as mere means. (A) is correct, though positing non-consequentialist rights within a predominantly consequentialist theory can be questioned on grounds of parsimony. (B) is problematic because their account of the right not to be used as a mere means is puzzling.

Pereboom's position on the right not to be used as a mere means has evolved. In *Free Will, Agency, and Meaning in Life*, he holds that people who I "harm in self-defense" are "being used merely as a means," and though this is a moral concern, it is "outweighed by the right to harm in self-defense," so long as "the harm inflicted is the minimal amount

[26] Stein (2005): 175.
[27] This section adapts some remarks from Vilhauer (2023b).
[28] See e.g. Pereboom (2001; 2014; 2021); Caruso and Pereboom (2020); Caruso (2021).
[29] Smilansky (2019).

reasonably required."[30] But in non-consequentialist moral theories, the right not to be used as a mere means is typically understood to be *absolute*, not as a right that can be overridden. From the Kantian perspective, using people as means is *never* permissible, unless they would rationally consent, in which case they are *not* treated as mere means. The DSC demonstrates this with respect to punishment.

More recently, Pereboom and Caruso hold that using people as means without their consent only requires a special justification if we use them *manipulatively* toward ends other than self-defense, such as general deterrence.[31] They claim that the quarantine analogy yields what they call "free general deterrence," that is, general deterrence from which we can benefit *without* taking on any obligation to provide a moral reason for using people as means. They think almost nobody wants to be quarantined, so quarantine inevitably produces deterrence as a side effect, and think the same is true for imprisonment. However, as John Lemos and I have argued in different ways, quarantine does *not* inevitably deter.[32] The COVID-19 era showed that many are not very distressed by quarantine, especially when the authorities send checks that allay income anxiety. The argument from Smilansky discussed earlier shows that when the authorities constrain people's freedoms, they have an obligation to compensate them.[33] So the authorities have an obligation to send people checks when quarantine is imposed. But problems would arise if checks got too big. Excessive checks would give people at low risk of serious illness (because they are youthful or hardy) an incentive to intentionally expose themselves to pathogens in order to get quarantined and paid. This would make quarantine ineffective. The authorities would need to ensure that checks were big enough, but not too big. This might seem to raise no moral concerns. Some might see any checks at all as mere largesse, a matter of supererogatory generosity, and might therefore suppose the authorities could limit checks as they wished without providing a moral reason for the limit. But this is a mistake, because the checks are an obligatory response to the authorities' constraints on liberties, and the authorities must therefore have a moral reason for setting limits. Said differently, quarantine practices must be calibrated to make life in quarantine as pleasant as possible while still making it *unpleasant* enough to *deter* people from intentional exposure. We need a moral reason to do this, and the kind of moral reason we need is quite precisely a reason for general deterrence. This shows that general deterrence does not come as a free byproduct – we cannot justify effective quarantine without justifying general deterrence.

If the analogy Pereboom and Caruso draw between quarantine and punishment holds up, then justifying effective punishment also requires justifying general deterrence. Arguments earlier in this paper also show this to be true. But here is an example to illustrate the point. Suppose I am a prison warden, and my goal is merely to detain violent offenders in comfortable conditions. Suppose I believe I am not entitled to aim at general deterrence, since I think calibrating prison conditions for general deterrence

[30] Pereboom (2014): 167.

[31] Caruso (2021); Pereboom (2021); also see Shaw (2019).

[32] Lemos (2016); Vilhauer (2019). Both draw on arguments from Smilansky (2011).

[33] Pereboom responds to Smilansky by noting that theorists who reject basic desert (which is equivalent to what is called "action-based desert" here) see "no basic desert requirement to compensate" the detained, and that consequentialists see no reason to compensate them heavily (Pereboom 2014: 172–173). But this response is inadequate if deniers should question consequentialism and not everything we fundamentally deserve is action-based.

would nonconsensually and impermissibly use the imprisoned. Now suppose I discover that conditions are producing general deterrence as a side effect. Perhaps general deterrence comes for free until I discover this, since I did not intend to create general deterrence. But upon discovery, it is no longer free: my belief about the impermissibility of calibrating conditions for general deterrence gives me a reason to *improve* conditions. If I wish to preserve general deterrence, I need a justification of general deterrence. I could appeal to the DSC to show why people would rationally consent to being used for general deterrence, or I could appeal to consequentialism and argue that it is permissible to nonconsensually use people for general deterrence after all.

In response to considerations like this, Pereboom now endorses a view that supplements free general deterrence with a straightforwardly consequentialist argument for general deterrence,[34] and holds that even nonconsensual, manipulative use is consistent with people's right not to be treated as mere means[35] so long as they are not treated too severely.[36] This distorts the right not to be used as a mere means. Caruso still holds that the only general deterrence we should endorse is free general deterrence.[37] But there is no free general deterrence, so appeals to it obscure a justificatory burden that punishment theories must bear. The DSC can bear this burden.

4. Avoiding Moral Nihilism

Some may object that deniers must deny morality altogether, and accept moral nihilism. If this is right, then deniers' efforts to justify punishment (or anything else) founder in absurdity. This objection may be at its sharpest when it addresses moral theories (like the one advocated here) which understand moral reasons as propositions about what agents *ought* to do, and it focuses on the widely-accepted "*ought implies can*" principle (OIC), according to which it can only be true that I ought to do *x* if I can do *x*. It is natural to understand deniers' position as radically diminishing the range of things we can do. Determinism entails that there is at most one action which it is physically possible for an agent to take at any given time, and deniers typically think this entails that there is at most one way a deterministic agent *can* act at a given time. Deniers also typically think indeterministically caused actions are not under our control in the deep sense necessary for it to be true that we have alternative possibilities of action in the way that matters for free will and moral responsibility. Deniers who hold these views but wish to preserve OIC as well as the basic principle of deontic logic that says agents sometimes do things they ought not do are threatened with the objection that, on their view, alternative possibilities are not *accessible* in the way that is necessary for it to be true that we *can* do anything but what we *actually* do, and that it is therefore false that we ought to do anything we do not actually do.[38]

[34] Pereboom (2021): 91, 102.
[35] Ibidem: 95.
[36] Ibidem: 85.
[37] Caruso (2021): 312.
[38] Some hold that failing to do as one ought entails blameworthiness (FOEB). See e.g. Dahl (1967). Deniers who accept "oughts" must reject FOEB. But FOEB proponents must acknowledge FOEB to be a much less fundamental feature of deontic logic than the feature according to which agents sometimes do things they ought not do. Deniers can preserve this more fundamental feature. There are reasons to doubt FOEB that have nothing to do with denial. FOEB implies that we cannot act when moral dilemmas obtain without blameworthiness.

Deniers can respond by rejecting OIC. One prominent OIC rejecter is John Martin Fischer, a semicompatibilist who holds that moral responsibility is compatible with determinism even if alternative possibilities that support OIC are not.[39] If semicompatibilists can reject OIC, then why not deniers? It may, however, be more attractive for deniers to find a way to accept OIC. We arguably cannot deliberate about how we ought to act at some point in time unless we believe we can act in alternative ways at that time. Deniers can accommodate this deliberation requirement as well as OIC by adopting an epistemic interpretation of the "cans" needed to ground "oughts." According to this view, there are (at least) two different senses of "can" necessary in the free will debate, which correspond to different kinds of accessibility of alternatives. The "can" that matters for FW/MR is *accessibility in action*. The "can" that matters for OIC is *accessibility in deliberation*. On the epistemic interpretation, alternatives need only be epistemically available to be accessible in deliberation: to say that I *can* do x is to say that, to the best of my knowledge, it is possible that I *will* do x.[40] On this interpretation of "can," it remains true that I can act in alternative ways at any future time of action no matter what metaphysics of accessibility I accept, so long as I make reasonable assumptions about the limits of my knowledge about my future actions.[41] Even if determinism is true, it is quite unlikely that we will ever be able to predict how we will act in any detail, given the vast number of variables that would have to be measured. Theories are simpler when they do not multiply senses, but different senses of "can" are necessary in different contexts, in ordinary language as well as in philosophy.[42]

Objectors may now argue that even if the epistemic approach to "can" has some merits, it cannot plausibly support the "ought" claims about rational consent which are fundamental for the DSC. Objectors may point out that giving consent is an action, and argue that it could only be legitimate to treat an agent in some way *x* based on the claim that he *would* have rationally consented to *x* if it is the case that he *actually could* have rationally consented to *x*, in the sense that I called accessibility in action just above. This would imply that rational consent can only have the significance the DSC requires if possible worlds in which agents give rational consent are available to them with the same kind of accessibility which suffices for FW/MR. If this reasoning is sound, then deniers have no principled way to bring in rational consent without bringing in FW/MR too, and thereby giving up on denial. But it is not sound, because rational consent matters even if it is entirely hypothetical, entirely a matter of counterfactual consent that

[39] Fischer (2003).

[40] For discussion, see Pereboom (2001): 137.

[41] If "cans" are based only on agents' ignorance about what they will do, then if an agent acts in a way she prospectively believed she ought not act, her belief is threatened with retrospective falsification. Retrospective falsification does less damage to deontic logic than the elimination of "oughts" from practical reasoning altogether, so it is worthwhile for deniers to preserve "oughts" even at the price of retrospective falsification. But retrospective falsification can arguably be avoided by time-indexing ought claims. Skeptics like myself, who hold that alternatives *may* be accessible in action but we just don't know, have a stronger defense against retrospective falsification than deniers (Vilhauer: 2023a). But since skeptics and deniers are mostly in the same boat when it comes to criminal justice, I leave this issue out of the main text.

[42] Ben Schwan argues that "the truth of an ability ascription depends on an (almost always implicit) characterization of the relevant possibility space" (Schwan 2018: 703).

would have been given under different circumstances. Suppose a cult kidnapped and enslaved our friend, and brainwashed him to be content with his enslavement. Suppose the brainwashing is so complete that he not only *does* not but also *cannot* consent to deprogramming therapy in the actual world. Suppose his thinking in the actual world is so deeply irrational that a world in which he rationally consents is not accessible to him in action *or* in deliberation, no matter what metaphysics of alternatives we adopt. It is nonetheless clear that he *would* consent to deprogramming therapy if he thought rationally about it, and this contributes to justifying the belief that we ought to coerce him into deprogramming therapy. Rational consent fits into the DSC in a parallel way. Since rational consent matters even when it is entirely counterfactual, deniers can incorporate it into their ethics without endorsing FW/MR. Appeals to entirely counterfactual rational consent should be scrutinized more closely when made in a justification of punishment than when made in a justification of liberating someone from slavery, since punishment is a serious harm and liberation is a benefit, and justifications for serious harm bear a heavy justificatory burden. But the modal structure of the appeal is the same, and it stands up to scrutiny. If the appeal fails, then I think it must be for reasons other than its modal structure.

5. Conclusion

OPD deliberation which assumes we will be targets of the criminal justice system gives us high confidence that we can justify a criminal justice system with a reasonable doubt conviction standard and prison conditions deterring enough to prevent the state of nature. But as discussed, there is room for debate about whether the criminal justice system we endorse under the assumption that we will be targets is fair to the people the criminal justice system should protect. I have not argued that it is fair to them. Instead, I have argued that we can be especially confident that we will not treat targets unfairly if we follow the principles to which we would consent under the assumption that we will be targets. A model of fairness in which OPD deliberators weigh interests of both the targeted and protected may justify a more stringent criminal justice system, one which deters more crime through more unpleasant prisons than those described here, and uses a conviction standard lower than the "reasonable doubt" standard as interpreted here (which demands a probability close to 100%). The arguments in this paper imply that we can have less confidence that such a system is fair to its targets. But as far as the argument here is concerned, it *may* be fair to its targets despite our lower confidence. However, critics of free will denial sometimes argue that deniers cannot justify punishment at all,[43] and sometimes argue that deniers must endorse practices that strike our moral intuitions as excessively punitive.[44] So it is valuable for deniers to have a justification which offers high confidence that we can treat targets fairly despite treating them stringently enough to avoid the state of nature.

[43] E.g. Smilansky (2011).
[44] E.g. Lemos (2023), especially chapters 7 and 8.

Acknowledgments: Thanks to Matthew Altman, Gregg Caruso, Michael Corrado, Sofia Jeppsson, John Lemos, Derk Pereboom, Elizabeth Shaw, Saul Smilanksy, Bruce Waller, Przemysław Zawadzki, and *Diametros* reviewers for helpful communications at various points in the development of this view.

Funding: This article received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest: The author has no conflict of interest to declare.

References

Caruso G.D. (2021), *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice*, Cambridge University Press, Cambridge.

Caruso G.D., Pereboom D. (2020), "A Non-Punitive Alternative to Punishment," [in:] *The Routledge Handbook on the Philosophy and Science of Punishment*, F. Focquaert, E. Shaw, B.N. Waller (eds.), Routledge, New York: 355–365.

Clark M. (2004), "A Non-Retributive Kantian Approach to Punishment," *Ratio* 17 (1): 12–27.

Dahl N.O. (1967), "'Ought' and Blameworthiness," *The Journal of Philosophy* 64 (13): 418–428.

Dolovich S. (2004), "Legitimate Punishment in Liberal Democracy," *Buffalo Criminal Law Review* 7 (2): 307–442.

Fischer J.M. (2003), "'Ought-Implies-Can', Causal Determinism and Moral Responsibility," *Analysis* 63 (3): 244–250.

Harsanyi J. (1975), "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory," *American Political Science Review* 69 (2): 594–606.

Kant I. (1996), *Practical Philosophy*, trans. and ed. M.J. Gregor, Cambridge University Press, Cambridge.

Lemos J. (2016), "Moral Concerns about Responsibility Denial and the Quarantine of Violent Criminals," *Law and Philosophy* 35 (5): 461–483.

Lemos J. (2023), *Free Will's Value. Criminal Justice, Pride, and Love*, Routledge, New York.

Murphy J.G. (1973), "Marxism and Retribution," *Philosophy and Public Affairs* 2 (3): 217–243.

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

Pereboom D. (2014), *Free Will, Agency, and Meaning in Life*, Oxford University Press, Oxford.

Pereboom D. (2021), *Wrongdoing and the Moral Emotions*, Oxford University Press, Oxford.

Rawls J. (1999), *A Theory of Justice: Revised Edition*, Harvard University Press, Cambridge (MA).

Schwan B. (2018), "What Ability Can Do," *Philosophical Studies* 175 (3): 703–723.

Shaw E. (2019), "Justice without Moral Responsibility?," *Journal of Information Ethics* 28 (1): 95–130.

Smilansky S. (1996), "Responsibility and Desert: Defending the Connection," *Mind* 105 (417): 157–163.

Smilansky S. (2000), *Free Will and Illusion*, Oxford Clarendon, Oxford.
Smilansky S. (2011), "Hard Determinism and Punishment: A Practical Reductio," *Law and Philosophy* 30 (3): 353–367.
Smilansky S. (2019), "Free Will Skepticism and Deontological Constraints," [in:] *Free Will Skepticism in Law and Society: Challenging Retributive Justice*, E. Shaw, D. Pereboom, G.D. Caruso (eds.), Cambridge University Press, Cambridge: 29–42.
Stein A. (2005), *Foundations of Evidence Law*, Oxford University Press, Oxford.
Sterba J.P. (1977), "Retributive Justice," *Political Theory* 5 (3): 349–362.
US Supreme Court (1970), *In re Winship*, 397 U.S. 358.
Tribe L.H. (1971), "Trial by Mathematics: Precision and Ritual in the Legal Process," *Harvard Law Review* 84 (6): 1329–1393.
Vilhauer B. (2009), "Free Will Skepticism and Personhood as a Desert Base," *Canadian Journal of Philosophy* 39 (3): 489–511.
Vilhauer B. (2013), "Persons, Punishment, and Free Will Skepticism," *Philosophical Studies* 162 (2): 143–163.
Vilhauer B. (2019), "Deontology and Deterrence for Free Will Deniers," [in:] *Free Will Skepticism in Law and Society: Challenging Retributive Justice*, E. Shaw, D. Pereboom, G.D. Caruso (eds.), Cambridge University Press, Cambridge: 116-138.
Vilhauer B. (2023a), "An Asymmetrical Approach to Kant's Theory of Freedom," [in:] *The Idea of Freedom: New Essays on the Kantian Theory of Freedom*, D. Heide, E. Tiffany (eds.), Oxford University Press, Oxford: 130–149.
Vilhauer B. (2023b), "Free Will Skepticism and Criminals as Ends in Themselves," [in:] *The Palgrave Handbook on the Philosophy of Punishment*, M.C. Altman (ed.), Palgrave-Macmillan, New York: 535–556.
Vilhauer B. (2024), "Five Perspectives on Holding Wrongdoers Responsible in Kant," *British Journal for the History of Philosophy* 32 (1): 100–125.