Free Will Skepticism and Criminals as Ends in Themselves

Benjamin Vilhauer

Introduction

Free will skepticism undermines retributivism, which plays a crucial role in many justifications of punishment and remorse. Free will skeptics often think that retributivism is cruel and undermining it is a victory. But they must confront difficult questions about developing a non-retributive ethics for criminal wrongdoing. It is perfectly coherent for free will skeptics to hold that no form of punishment is justified and that we should never feel remorse. However, it is natural to worry that a society which ceased to punish altogether would revert to the state of nature, and that advocating remorselessness is too much like advocating psychopathy. If free will skeptics seek to justify punishment and remorse, their most obvious option is consequentialism, which has its own problems. It is plausible that punishment and remorse have good consequences – that punishment deters, and remorse improves behavior. But if these are our only justifications, we use criminals as mere means to ends, and remorse is a mere means rather than an experience that "fits" our wrongs.

This chapter offers non-retributive, broadly Kantian justifications that resolve these problems and can be endorsed by free will skeptics. The justification of punishment draws on non-retributive original position deliberation: we consent to humane punishment if we assume we are equally likely to be among the punished and among those protected by punishment when the veil is raised. It has consequentialist features but is deontologically founded in the duty to avoid using others as mere means. The justification of remorse is care-based. It draws on the value of sympathizing with people we have wronged, which has a Kantian ground in the duty to take others' ends as our own.

These accounts of punishment and remorse are Kantian in that they emphasize the idea that people are ends in themselves. As understood in this chapter, to be an end in oneself (which I will typically abbreviate as "to be an end") is to have a moral status with four key features:

(1) One ought never to be coerced or deceived into serving as a *mere means* to ends to which one would not rationally take as one's own. This element grounds a *perfect* duty: we must *never* treat others as mere means. We need not avoid treating each other as means in all cases, since it is sometimes rational to consent to serving as a means to another's end, for example, when they are reciprocally serving as a means to one's own end. If I have the end of teaching, then my students are among my means, and if they have the end of learning, then I am among their means. Such reciprocity is central for the account of punishment below.

(2) One's own permissible ends are valuable in a way that gives others reasons to take those ends as their own. This feature grounds an *imperfect* duty that grants us latitude – we must *sometimes* take others' ends as our own. It is because of this latitude that Kantianism is often thought to be less demanding than utilitarianism. But even for Kant, this duty becomes more demanding, in a way that brings it closer to perfect duty, when we have relationships with morally salient others such as friends and family. This point figures in the account of remorse to be presented here. Adopting others' ends requires sympathy, and sympathy sometimes involves pain. It is intuitive to think that relationships with one's victims are as morally salient as relationships with family and friends.

(3) One ought to treat others in conformity with (1) and (2).

(4) The fact that something is an end does not imply that we have reasons to make *more* things like it. To be an end is to be a normative node in the practical reasoning of rational agents such that (1), (2), and (3) constrain their reasoning. This is a key feature of Kant's non-consequentialist moral teleology. So, recognizing criminals as ends does not give us a reason to make more criminals.

The accounts of punishment and remorse are only *broadly* Kantian, however, since they modify his views in (at least) three important ways:

i) Kant's mature ethics adds a fifth key feature to the list of four above: we can only regard ourselves as ends if we know we have free will of a libertarian kind, which he calls "transcendental freedom." But the kind of autonomy we have when we are treated as ends is not conceptually dependent upon free will. Agents who are treated as ends are independent of others' undue control. Agents who have transcendental freedom are independent from their causal histories. We can have the former without the latter.

ii) Retributivism plays an essential role in Kant's account of punishment, though he offers little in the way of argument for retributivism. He appears to see it as the only way to avoid endorsing punishing people as mere means (Kant 1996b, 6:331 [p. 473]).

iii) Retributivism plays an essential role in Kant's account of remorse (Kant 1996b, 6:394 [p. 524]) and he does not consider a sympathy-based alternative, but he has an account of rational sympathy that provides a basis for this alternative.

Readers may wonder: if Kant has to be revised so much to be useful to free will skeptics, then why bother with Kant in the first place? The reason is that it has long been thought, not just by Kant but by many philosophers working on free will and moral responsibility, that free will skepticism inevitably pushes us away from Kantian ethics and toward consequentialism. For example, one of P. F. Strawson's main arguments in his influential paper "Freedom and Resentment" (1962) is that free will skepticism requires us to adopt a depersonalizing "objective attitude" in which our only moral reasons are utilitarian. Strawson and his many followers see this as a reason to endorse free will so that we can retain broadly Kantian ethical views. Saul Smilansky (2000) thinks that we lack free will, but he also thinks that free will denial pushes us away from broadly Kantian commitments which are so important that we should cultivate the illusion that we have free will. On the other hand, some theorists think that Kant's influence has been largely inimical, so moving away from him is valuable progress (e.g., Waller 2015a, 2015b, and his chapter in this volume). If it can be shown that key Kantian ideas can be usefully retained by free will skeptics, this gives skeptics a more resourceful moral theory, especially when they respond to criminal wrongdoing.[1]

Defining and motivating free will skepticism

As understood here, *free will* is the control condition for moral responsibility, and *moral responsibility* is the relationship to our actions that would legitimize *action-based desert claims*, such as claims that people deserve praise, blame, reward, or punishment for their actions. *Retributivism* is the view that action-based desert claims play a necessary role in justifying punishment.

Free will skepticism is the view that we do not know whether we have free will.[2] Kant is a skeptic about free will from the perspective of theoretical rationality, but he claims practical

[1] To my knowledge, Pereboom (2001, 150-52) is the first contemporary free will denier to argue that Kantian ideas can be preserved by skeptics, but Kantianism plays only a limited role in his response to criminal wrongdoing (see below).

[2] Some writers in this literature use "free will skepticism" to refer to the view I call "free will denial." But these are importantly different positions that should be distinguished appropriately. "Skepticism" traditionally refers to claims

knowledge that we have free will (1996a, 5:30-31 [p. 163]). Free will skepticism as understood in this chapter accepts Kant's theoretical skepticism but rejects his practical knowledge claim.

Free will skepticism is different from *free will denial,* which is the view that we do not have free will. Free will skepticism is more epistemically modest and theoretically conservative than free will denial. This makes it more plausible than free will denial in important ways. Free will skeptics need not prove that compatibilist accounts of free will are inadequate or that there are no libertarian metaphysical sources for free will. They need only introduce reasonable doubts that suffice to undermine knowledge claims. The raging controversies about the viability of compatibilism and the plausibility of libertarian metaphysics suggest that such reasonable doubts are readily available, so this chapter will not provide an inventory. Free will skepticism is also more theoretically conservative with respect to widely accepted features of moral theory. I argue that the possibility that we have free will is sufficient to ground "cans" corresponding to "oughts" in the "ought implies can" principle (Vilhauer 2012, 2015). I also argue for an asymmetry in the standards of justification we must meet to hold people morally responsible in the contexts of "positive" and "negative" responsibility-attribution practices, such that the possibility of free will suffices for some forms of praise – for example, praise that helps people understand themselves as valuable moral agents (Vilhauer 2012, 2015).

However, the possibility that we have free will is not sufficient to support retributivism about criminal wrongdoing. As explained above, action-based desert entails moral responsibility, and moral responsibility entails free will. This means that free will denial entails the denial of

about what we can know about things, not claims about how things are. Skeptics about other minds rarely hold that there are no other minds. So we ought not refer to the view that there is no free will as "free will skepticism." (I must confess, however, that I too used the term in this problematic way in some earlier papers.)

retributivism. But it also means that free will skeptics' reasonable doubts about free will entail reasonable doubts about retributivism. Reasonable doubts about retributivism are enough to undermine retributivism, at least in the context of the ethics of criminal wrongdoing, where retributivists rely on the claim that we have free will to justify profound suffering. It is obvious that criminal punishment often involves profound suffering. The suffering of remorse can also be profound. People sometimes choose suicide over continuing to live with remorse. The remorseful suffering typically expected in response to serious crimes is especially profound. It is widely acknowledged that justifications for inflicting the suffering involved in criminal punishment must meet an extremely high justificatory standard. This is why so many people find it intuitive that arguments in the criminal court must be proven beyond reasonable doubt. A similarly high standard should be met for remorse. Retributive justifications of such suffering cannot meet these standards in light of free will skeptics' reasonable doubts (Pereboom 2001; Vilhauer 2009b; 2012; Caruso 2021; Jeppson 2021). So, it is not only the radical position of free will denial that undermines retributivism – a modest and comparatively conservative free will skepticism undermines it too.

The variety of desert bases

As mentioned earlier, consequentialism is the most obvious option for skeptics seeking non-retributive justifications of punishment and remorse. The discussion in this chapter primarily addresses utilitarianism, because it is a simple and clear form of consequentialism. Skeptics may be troubled by utilitarianism's implications for criminal wrongdoing. Objectors may protest that it is disingenuous for skeptics to claim to be troubled by utilitarianism, because if nobody really deserves anything, intellectual honesty demands that we bite the bullet and accept that the only moral reasons still standing are utilitarian. But neither skeptics nor deniers should hold that nobody deserves anything.[3] Skepticism undermines action-based desert. But it is plausible that action is not the only desert base (that is, the thing that grounds legitimate claims about desert).

Personhood is a desert base that is plausibly distinct from action (Vilhauer 2009a, 2013). For example, we deserve to be treated as innocent until proven guilty just because we are people, and there is a special kind of kind of respect we deserve just because we are people.[4] We do not deserve these things because we have *acted* in a way that makes us deserve them, and there is no way we could act that would make us *cease* to deserve them. Even if we felt confident that somebody was a murderer, they would not cease to deserve the presumption of innocence, and even if they were convicted, they would deserve protection from punishment that disrespects their humanity. Universal human rights that cannot be alienated or forfeited are plausibly grounded in personhood as a desert base.

Need is another desert base that is plausibly distinct from both action and personhood. We respond to need with forms of care. Children deserve love from their parents even when they are too young to have done anything to *earn* their love.[5] People lying bleeding on the sidewalk deserve a call for an ambulance from passersby even if they *did* play a role in the circumstances that led to their injuries. Need is a more complex desert base than personhood, as the obligations it confers do not always universalize. Children deserve love from their parents in a way they do

[3] Pereboom holds that free will is required for what he calls "basic desert," but he endorses "forward-looking" moral responsibility, which makes room for non-basic desert (2014, 2, 126-52). However, his view seems to be that without free will nobody *fundamentally* deserves anything.

[4] By contrast, Smilansky holds that "the idea of respect for persons … is control related, and not personhood related as for Vilhauer. And the concern with 'being used merely as a means' … is also firmly embedded in control-related ideas" (2019, 31). However, as mentioned in the introduction, while there is a kind of voluntariness at issue in treating people as ends, it is conceptually independent of free will.

[5] McLeod (2013) makes a similar argument.

not deserve it from others, while people who lie bleeding on the sidewalk deserve an ambulance call from anyone passing by.

The independence of action, personhood, and need as desert bases makes it sensible for skeptics and deniers to acknowledge desert claims based on personhood and need despite doubting or rejecting desert claims based on action. Multiple approaches to moral theory may allow this, but the focus here is on non-retributive Kantianism. From this perspective, the most fundamental way to understand what we deserve based on our personhood is to be treated as we would rationally consent to be treated if we had only our personhood in view, and the most fundamental way to understand what others deserve from us based on their need is in terms of our duty to adopt others' ends as our own. The accounts of punishment and remorse below aim to unpack these ideas in the context of criminal wrongdoing.

Non-retributive Kantian punishment for skeptics

As already noted, utilitarianism is a ready option for skeptics seeking to justify punishment. Utilitarians think criminals' pain is as bad as anybody else's, and they see the pain involved in punishment as justified only insofar as it diminishes overall suffering in society, for example, by incapacitating criminals and deterring potential criminals. So utilitarianism can seem less cruel than retributivism. However, if our only reason to punish criminals is to reduce suffering elsewhere in society, then we are using criminals as mere means. Further, suppose we get the best ratio of pain inflicted to pain prevented with practices that violate moral intuitions most ethicists want to preserve, such as framing people and weakening due process. Utilitarians have to endorse these practices if they cannot find a way to rule them out.[6] Some philosophers adopt a hybrid view according to which punishment must increase utility *and* be deserved based on actions. Skeptics do not have this option.

Kantian skeptics can point out that, while the action-based desert claims favored by retributivists are undermined by skepticism, personhood-based desert claims are not, and we have a personhood-based desert claim not to be used as mere means. So, they can propose an institution of punishment to which we could rationally consent. The notion of consent to punishment can seem bizarre. Few criminals *actually* consent to punishment – coercion is in the nature of punishment, and it is in the nature of coercion that we do not consent to it. Thus any plausible account of consent to punishment must rest on hypothetical consent, consent we *would* give under appropriate hypothetical conditions. We can model this with social contract theories based on hypothetical consent. John Rawls's approach is a natural fit for free will skeptics who want to model personhood- but not action-based desert, since Rawls himself recommends original position deliberation (OPD) in part because it screens out undeserved inequalities (1999, 86-89).

In OPD, deliberators use maximin reasoning to select basic social principles that make the circumstances of the worst-off as good as possible. Rawls argues for OPD as risk-averse in a way that is rational under uncertainty, and this has been disputed (e.g., Harsanyi 1975). However, he also defends OPD as conforming to moral facts that are normatively prior to the social contract: OPD procedurally specifies an underlying Kantian conception of fairness and equality among rational beings. This chapter draws on the latter approach.

[6] Perhaps there is a rule-utilitarianism that can adequately explain why we should follow rules like "do not frame" when breaking them would maximize utility. It seems unlikely to me. But even if there is, we should also know whether a non-retributive Kantian option is available.

Rawls applies OPD to distributive justice, not punishment. Although OPD can be extended to punishment,[7] a disanalogy between distributive justice and punishment complicates maximin reasoning. There is only one worst-off social position in distributive justice – the poorest. In punishment, crime victims and the people punished compete for the worst-off position (though later it will be argued that this competition is not fundamental). If punishment deters, then adjusting punishment to improve things for one party worsens things for the other. Perhaps technological and social innovations can someday eliminate this competition in a morally attractive way. Someday speedy AI ticklebot police may overwhelm all would-be criminals with incapacitating but harmless giggles before they complete crimes, eliminating the need for punishment and the position of the punished in the competition equation. But yet-unimagined social innovations would be required to prevent ticklebots from becoming tools of authoritarian repression. We could of course eliminate the position of the punished by ceasing to punish criminals, but we would worry that crime might explode and cast us into the state of nature.

If criminals and victims will compete for the foreseeable future, how should OPD weigh their interests? To be fair to both, we must assume that we have an equal chance of ending up in each position – that we are just as likely to be harmed by punishment as we are to benefit. The relevant harmed parties are obviously the people punished. The beneficiaries upon whom we

[7] For previous justifications of punishment drawing on OPD, see Murphy (1973), Sterba (1977), Clark (2004), and Dolovich (2004). The justification provided in this chapter is novel in its claim to use OPD to unpack what we deserve based on personhood but not action. The other justifications import retributivist premises and cannot be endorsed by skeptics. Clark aims to provide a non-retributive Kantian approach to punishment, but he allows a “negative retributivism” that explains why we should punish the guilty rather than the innocent. Skeptics cannot allow this. The justification provided in this chapter shares more features with Dolovich’s account than the others, despite the fact that I developed this justification before I became aware of her paper.

should focus are potential victims rather than actual victims, because victims have already suffered the harm OPD deliberators would hope to avoid by instituting punishment.

What principles of punishment would we choose under this assumption? Fear of punishment would make OPD deliberators initially prefer a society that did not punish. They would invest in research on technologies and practices (like authoritarian-proof ticklebots) that would yield a just but crime-free society with no need for punishment. Since such innovations are not yet on hand, they would also invest in attractive crime-prevention measures already available: more jobs, education, public services, and voluntary therapy for those most at risk of committing crimes. But they would endorse some form of punishment in order to avoid the state of nature. This motivation for endorsing punishment is consequentialist, but it only has normative significance because it unfolds from rational consent. Thus, this justification of punishment is fundamentally deontological despite the consequentialist motivation.[8]

What particular form of punishment would OPD deliberators choose based on these general principles? Punishment imposes significant harm on the punished to confer what may be a very modest benefit on the potential victim. Even radically reformed prisons would cause significant harm by blocking prisoners' freedom of movement and damaging their social relations. A reduction in someone's odds of becoming a victim does not confer a similarly tangible benefit. If I was confident that OPD adoption of some particular form of punishment would ensure that I did not find myself an actual victim when the veil is lifted, I might think the benefit of punishment to potential victims was equal to the harm to the punished. But I cannot be confident about this in OPD. The aggregate benefit of punishment to society as a whole may be

[8] Kant's own justification of punishment has consequentialist elements (1997, 27:286 [p. 79]), so this feature of the OPD approach is compatible with its Kantian foundations. See Vilhauer (2017) for discussion.

much greater than the harm it imposes: even if we cannot know which individuals will be saved from victimization, we may confident that victimization will dramatically decline overall. But this is not relevant in OPD, since it requires us to consider social outcomes one person at a time, thereby avoiding the utilitarian deletion of the boundaries between persons. This is part of why it is a deontological justification.

In OPD, we would be unwilling to risk imprisonment to protect ourselves against non-violent crime. We would choose less-intrusive alternatives such as fines and ankle monitors. On the other hand, we would risk imprisonment to protect ourselves against crimes of violence. But we would insist on humane prisons that offered education, meaningful work, voluntary therapy, regular visits from friends and loved ones, very frequent parole review to determine whether prisoners could be released without undue risk of repeated violence, and radically enhanced post-release support to help people avoid new violence. The main function of such prisons would be incapacitation rather than deterrence. But we could not ignore deterrence, since a primary reason to want punishment in OPD would be maintenance of enough order to avoid the state of nature. Prison would provide a substantial deterrent even if prison conditions were comfortable, since we prefer not to be controlled and separated from loved ones. But if conditions were *too* comfortable, they would become an incentive to commit violent crime, and prison would work against its intended purposes. We would therefore choose a policy of calibrating prison conditions at a level of unpleasantness high enough to maintain deterrence, but no higher. Conditions would not have to be intrinsically unpleasant to deter, just unpleasant relative to life outside prison. As discussed earlier, OPD deliberators' fear of punishment would prompt them to diminish incentives for crime by funding jobs, education, and social services. When life gets better outside prison, it can get better inside while still deterring.

According to the OPD approach, both special and general deterrence are justified because we would rationally consent to them. Rational consent means that general deterrence uses criminals as *means,* but not *mere* means. But it is easiest to be satisfied with contract-based claims about rational consent when we identify ends of people burdened by the contract that are achieved despite their burdens. Any plausible social contract theory acknowledges that it is not rational to consent unless the contract gives us lives better than the state of nature. If life inside prison is no better than the state of nature, then we can only justify it through contract if we represent criminals as contract-breakers who have forfeited their contractual claims. But this move turns on an implicit retributivism, and skeptics must reject it. They must instead make life in prison better than the state of nature. This is achievable in the humane prisons endorsed in OPD, since we make prison conditions as good as possible without undermining deterrence. It seems reasonable to think that life in prison could be quite a lot better than the state of nature when it is designed this way. Further, the punished and the protected *use each other reciprocally*. Both parties pursue the end of a life better than the state of nature. The protected pursue this end by using the punished to generate deterrence. The punished pursue this end by using the protected to generate the social resources necessary to provide the best prison conditions possible. If deterrence is necessary to avoid the state of nature, then it is only possible for the punished to have lives better than the state of nature if they consent to serve as means to deterrence. It is in this sense that the competition between the punished and the protected is not fundamental.

What if future research shows that we do not need prison conditions to deter after all, perhaps because the prospect of unpleasant imprisonment does not play a significant role in potential criminals’ decision-making, or because the mere fact of imprisoning violent offenders for a time prompts enough behavior improvement? The OPD justification is sensitive to this possibility, since such research would prompt deliberators to imprison only to incapacitate. But the OPD justification is equally sensitive to the possibility that future research will reinforce the need for deterrence, and it has the resources to justify it.

On this point, it may be useful to compare the OPD justification and another justification based on the concept of *quarantine*, which Smilansky has called the "two most developed denialist attempts to defend deontological constraints concerning punishment" (2019, 30). The quarantine justification is defended by Derk Pereboom and Gregg Caruso in a number of joint papers, and independent papers and books (e.g., Pereboom 2001, 2014, 2021; Pereboom and Caruso 2018; Caruso and Pereboom 2020; Caruso 2021; and his chapter in this volume). They hold that our right to *self-defense* makes it permissible to quarantine carriers of dangerous diseases even though they do not deserve to be sick, and they draw an analogy between quarantine and imprisoning violent criminals, arguing that we have as much right to imprison violent people as we do to quarantine carriers of dangerous diseases, even if violent people do not deserve to be incarcerated. As I understand their overall ethical theory, it is fundamentally consequentialist (Smilansky [2019] interprets it the same way). However, they think the quarantine justification is not consequentialist because they think (1) the right to self-defense need not be construed consequentially, and (2) their theory adequately protects criminals' rights not to be treated as mere means. (1) seems plausible, though simply positing non-consequentialist rights within a basically consequentialist view raises questions. (2) raises similar questions, but there are further puzzles about how their theory interprets the right not to be used as a mere means.

Pereboom's position on this point seems to have 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 while this is a concern, it is "outweighed by the right to harm in self-defense," so long as "the harm inflicted is the minimal amount reasonably required" (2014, 167). But Kantian perfect duties are absolute and cannot be outweighed – instead we need to show why people would rationally consent to be used, as the OPD approach endeavors to do. In more recent work, Pereboom and Caruso hold that using people without their consent is only problematic if we use them *manipulatively* toward ends other than self-defense, such as general deterrence (Caruso 2021; Pereboom 2021;see also Shaw 2019). Similar Kantian objections should be made to this move.

Having made this move, Pereboom and Caruso claim to justify quarantine based on self-defense without licensing illegitimate use, and to justify imprisonment via analogy with quarantine. They then claim that the quarantine analogy yields "free general deterrence," that is, general deterrence we can rely on without having to justify using people as means to general deterrence. The idea is that hardly anyone wants to be quarantined, so quarantine inevitably produces deterrence as a *side effect*, and the same is true for imprisonment. However, as John Lemos (2016) and I (2019) have argued in different ways, quarantine does *not* inevitably deter.[9] The COVID-19 era has shown that many of us are not unduly distressed by being required to stay at home. Many would not be distressed at all if the state sent checks to everybody required to quarantine. Presumably the state would have a *moral reason* to send such checks, to compensate the quarantined for the undeserved restriction of their freedom. But we would want to ensure the checks were not too big, because then people at low risk of dying might intentionally expose themselves to the coronavirus to get the checks, and our quarantine practices would work against

[9] Both Lemos and I draw on arguments from Smilansky (2011).

their intended purposes. So, we need quarantine practices calibrated to make life as good as possible in quarantine while still *deterring* people from intentional exposure. We need a justification of general deterrence to justify such calibration. This means that general deterrence does not come as a “free” byproduct – we cannot justify *effective* quarantine without justifying general deterrence, and the discussion earlier in the paper has shown how the same considerations apply to imprisonment.

Here is another way of seeing the same point. Suppose that my aim in setting up a prison is merely to incapacitate violent offenders in comfortable conditions, because I think I am not entitled to aim at general deterrence, since I think calibrating conditions for general deterrence would nonconsensually and impermissibly use the imprisoned. And suppose I discover that the conditions are producing general deterrence as a side effect. Perhaps general deterrence comes for free until I discover this, since I meant well in setting up the prison. 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 consequentialism and argue that it is not impermissible to nonconsensually use people for general deterrence after all, or I could appeal to a Kantian contractualism to show why people would rationally consent to such use.

Perhaps in response to arguments like this, Pereboom now offers a theory that supplements free deterrence with a straightforwardly consequentialist argument for general deterrence (2021, 101), and he holds that even nonconsensual, manipulative use is consistent with treating someone as an end (95) so long as they are not treated too severely (85). This takes us far from Kantian foundations. Caruso remains committed to the view that the only general deterrence we should seek is free general deterrence (2020, 312). But if the argument above is correct, there is no free deterrence, so relying on it obscures a demand for justification that punishment theorists should confront. The OPD approach confronts this demand.

Let me now turn to two other problems for utilitarian punishment mentioned earlier: framing and weakening due process. Imagine that we could strengthen general deterrence by occasionally framing and punishing celebrities, because of all the media attention it provides. To rule this out, we have to be a bit creative with OPD as Rawls understands it, so that we can use it to capture Kant's notion that *deception* is a way of using people as mere means. A practice that aims to deter by penalizing anybody other than actual criminals can only succeed by deception: punishing a framed celebrity is only effective if almost everyone is deceived about the framing. If the framing becomes widely known, we get less deterrence, not more. In OPD, I must assume that I may be among the deceived, so I would be volunteering to be deceived and thereby using myself as a mere means to increase deterrence.[10]

How do we rule out weakening due process? The guiding principle in applying OPD to criminal wrongdoing is to be fair, by assuming equal odds of being harmed and benefited when the veil is lifted. When we choose principles for due process, the competing parties are no longer potential victims and the punished – now they are potential victims and the *accused.* The accused have more to lose by weakening due process than potential victims have to gain. If we lower the conviction standard from "reasonable doubt" to (say) "preponderance of the evidence," we make things worse for the accused by increasing their odds of conviction. Some additional convicts will have been correctly accused, and getting the violent ones off the street will improve things for potential victims. However, the lowered standard will also facilitate sloppy or politicized

[10] See Kant (1996c, 8:381 [p. 347]) for a related argument and Vilhauer (2017) for discussion.

prosecutions that convict non-criminals, worsening things for the accused *without* improving things for potential victims. Since this would harm the accused more than it would benefit potential victims, OPD deliberators would not choose to weaken due process. As explained earlier, this holds even if weakening due process yields an aggregate reduction in victimization, because aggregate effects are irrelevant in OPD, since it makes us focus on one person at a time, and this is part of what makes it a deontological alternative to utilitarianism.

Non-retributive Kantian remorse for skeptics

It is intuitive to think that remorse plays an important role in the moral experience of anyone who is not perfect, and it has a special importance for philosophy of punishment, as it is often a mitigating factor in sentencing (see, e.g., Maslen 2015). This is a problem for skeptics, because it is not obvious how we could have a reason for remorse if we do not deserve to suffer, and without a reason for remorse we have no reason to treat it as a mitigating factor in sentencing. The very frequent parole review chosen in OPD would transform sentencing practices, but intuitively remorse would still be helpful in gauging the dangerousness of people we imprison, which is important for imprisoning them safely. So, it is worth exploring whether skeptics can justify remorse.

It would be perfectly consistent with the basic principles of skepticism to advocate remorse-elimination therapy, but this sounds uncomfortably like therapy for inducing psychopathy. The most obvious strategy for skeptics who want to justify remorse is (once again) utilitarianism: it seems reasonable to suppose that the pain of remorse may improve behavior, if only as a sort of self-administered aversion therapy, and that the pain of remorse is outweighed by the pain it prevents. But if this is my *only* reason for feeling remorse, then I am using remorse as a mere means: I experience remorse not because there is anything *fitting* (morally appropriate) about this experience, but merely because it improves behavior. Clearly remorse is not the kind of thing I can *wrong* by using it as mere means. So, it might seem that feeling remorse as a mere means to behavior improvement is no more problematic than enduring painful physical therapy as a mere means to mobility improvement in my knee after an injury. However, it would be perverse to choose painful knee therapy if technicians offer me a device I can strap to my knee that makes me feel a soothing warmth but improves mobility just as well. It does not seem perverse to experience remorse after I commit a murder instead of strapping a device to my head that causes soothing warmth and improves my behavior just as well. This is because of an intuition that remorse is morally important not only as a means to improved behavior, but also because it is fitting.

Retributivists can explain the fittingness of remorse in terms of action-based desert, but skeptics must reject this explanation. They can instead explain the fittingness of remorse in terms of the value of care. As understood here, sympathy is part of care.[11] When I care about someone, I sympathetically share their joy but also their pain. The value of care is not grounded in action-based desert. Even if we are not skeptics but instead believe firmly in free will, it would be absurd to think that, in befriending someone, I have acted in some way that entails that I deserve to suffer when my friend suffers. Care *is* quite plausibly grounded in need-based desert, but as explained earlier, need is a distinct desert base from action and is not undermined by skepticism.

[11] As I will discuss below, the emotional orientation called "sympathy" here could just as well be called "empathy," and "empathy" might in some ways be a better fit for the contemporary literature. However, Kant calls it "sympathy" (*Sympathie* in Kant's German), and given the present chapter's goal of providing a Kantian ethics for skeptics, Kant's term will be used here. Since there is no generally accepted account of how we might usefully distinguish sympathy and empathy (Stueber 2006, 27), this poses no conceptual problem.

According to the care justification of remorse, we should have this sort of sympathetic connection not just to friends but to a broader range of morally salient others, including people we have wronged. Wronging someone thus gives us a reason to care about them that parallels our reason to care about a friend. All forms of sympathy give us a reason to alleviate the pain of those with whom we sympathize by removing the causes of their pain. When we sympathize with people we have wronged, the cause is *our own actions*, and this gives us reasons to be pained *by* our actions, to alleviate their pain by making amends, and to improve our behavior toward others in general.

Human nature seems to contain deeply embedded desires about wrongdoer's responses to their wrongs. When people violate moral norms and hurt us, we desire not only that they make amends and improve their behavior, but also that they *understand* their wrong, in a way that involves not only cognition but also painful feeling. The idea that painful feeling is part of understanding one's wrongs helps explain the idea that painful feeling can be fitting. Some philosophers may wish to model this desire for wrongdoers to suffer in terms of Strawson's (1962) "reactive attitudes," which can be understood as essentially involving desires for the wrongdoer to experience deserved suffering. But to assume this model is to over-theorize our experience. Wrongdoers' sympathetic pain sometimes satisfies victims' desires. Since the value of sympathy is not grounded in action-based desert, victims' desires need not always be understood as retributive.

Since the care-based justification includes reasons to make amends and improve our behavior, it has a consequentialist dimension. But it is not fundamentally consequentialist, because care is valuable even when it does not have good consequences. Care requires sympathizing even when there is nothing we can do to help.[12] If I am marooned on a desert island and receive a message in a bottle informing me that my friend is in pain, and I feel no sympathetic pain just because I cannot help, my claim to care is undermined. This is also true for people we have wronged: care motivates us to make amends if we can, but if we cannot, we still sympathize, because we care. In this way, the care-based justification of remorse can explain why remorse is fitting, rather than a mere means to the end of good consequences. The care-based justification of remorse can be included within any moral theory that recognizes sympathetic pain as fitting. This includes varieties of care ethics and virtue ethics, and Kantian ethics as well.[13]

The claim that Kantian ethics can value sympathetic pain probably sounds strange to philosophers with limited familiarity with Kant, and even to some with considerable familiarity. So let me discuss Kant in a bit more detail than I did in the justification of punishment, which drew on more familiar Kantian ideas. Kant makes dismissive-sounding remarks about sympathy in his most famous moral works, the *Groundwork for the Metaphysics of Morals* and the *Critique of Practical Reason*, which can easily seem to imply that the sort of sympathy just discussed does not count as a moral emotion. But those remarks elide a distinction that is important in his ethics between two ways of sympathizing.[14] One is what we might call *natural sympathy* – an instinctive, pre-reflective reactivity to others' feelings that we share with many other animal species, which can overwhelm us and make it difficult to act prudently or morally. The other is

[12] Kantian "sages" can appear to reject sympathy when they cannot help (Kant 1996b, 6:457 [p. 575]), but what they reject is natural sympathy, not rational sympathy. This distinction is explained below. See Vilhauer (forthcoming a) for discussion.

[13] See Vilhauer (2004) for a virtue ethics approach.

[14] For passages that illustrate this distinction, see Kant (1996b, 6:457 [p. 575]; 1997, 27:677-78 [pp. 408-9]; 2007, 7:235-38 [pp. 338-40]; 2012, 25:606-7, 1320-21 [pp. 156-57, 429]). See Vilhauer (2021a) for commentary.

what we might call *rational sympathy* – it is what we experience when we reflectively regulate that animal capacity according to moral reasons. In the context of Kant's corpus as a whole, it is clear that Kant denies natural sympathy a role in moral feeling, but not rational sympathy. Rational sympathy is an intentional activity of the imagination that puts us "in the other's place" (Kant 2012, 25:476): we imagine what it is like to be in the other's situation, and this prompts sympathetic feelings.[15] Kant's distinction between natural sympathy and rational sympathy corresponds (and in fact appears identical) to a distinction in contemporary psychology between *empathic distress* and *empathic concern*.[16]

Kant clearly thinks sympathy is related to the duty to take others' ends as one's own, but the nature of the relationship is a matter of controversy.[17] I have argued that sympathy is necessary for taking others' ends as one's own, based on a distinction between *adopting* and *promoting* others' ends (Vilhauer 2021b). Many of others' permissible ends are ends they have because of the particular things that make them happy, due to features of their personality that are contingent relative to the more abstract perspective of Kantian rational agency. Rational sympathy lets me imagine my way into others' perspectives and call up feelings like theirs. This gives me sympathetic joy when they achieve their ends and sympathetic pain when they do not. In this way, I not only promote but also adopt their ends. I can *promote* others' ends without sympathy as means to *different* ends that others do *not* have. If somebody wants their bleeding stopped, I may bandage their wound because the sight of blood disgusts me, because I think it

[15] For passages that illustrate the connection between sympathy and imagination, see Kant (1996b, 6:321, 456-57 [pp. 464. 575]; 1997, 27:58, 65 [pp. 25, 30]; 2007, 7:179, 238 [pp. 288, 341]; 2012, 25:476, 574-76, 606-7 [pp. 52, 130, 156]).

[16] Empathic concern is an "intentional capacity" that involves "emotion-regulation" – it "involves an explicit representation of the subjectivity of the other" rather than "a simple resonance of affect between the self and other," while empathic distress involves "emotional contagion" (Decety, Jackson, and Brunet 2007, 254).

[17] See Fahmy (2009) for an influential alternative.

will help my reputation, or because I think it will help fulfill my duty. I promote their end in all these ways, and there may be no difference at all in the consequences I produce, but I do it as means to ends they do not have.

The Kantian idea that we should universalize our maxims may seem to require us to sympathize equally with everyone, and this may prohibit us from cultivating especially strong sympathy for particular others. But Kant does not advocate this. We have a duty of friendship (Kant 1996b, 6:469 [p. 585]), and while we ought to have "general good-will toward everyone," "to be everybody's friend will not do, for he who is a friend to all has no particular friend; but friendship is a particular bond" (Kant 1997, 27:430 [p. 190]). Friendship is an "ideal of each sympathizing and communicating about the other's wellbeing," which guides us toward a "maximum" (Kant 1996b, 6:469 [p. 585]; translation modified) in which "each mutually sympathizes [*teilnehmen*] with every situation of the other, as if it were encountered by himself" (Kant 1997, 27:677 [p. 408]; translation modified).

Kant himself does not propose a sympathy-based account of remorse. His own account relies on retributivism (Kant 1996b, 6:394 [p. 524]; see Vilhauer [forthcoming b] for commentary). But it is just as intuitively plausible to think that we ought to sympathize in a special way with people we have wronged as it is to think this about friends. So, it is a natural extension of Kant's view to hold that we ought to extend such sympathy to our victims. In fact, Kant nearly suggests this account himself in a discussion of why oppressors should sympathize with the oppressed (Kant 2012, 25:606 [p. 156]).

Objectors may argue that self-retribution is part of the nature of remorse, so it misdescribes the sympathetic pain discussed here to call it a kind of remorse. I think this is wrong, but even if it is right, I am not sure it matters if sympathetic pain can play the roles described above in a skeptical moral psychology. Objectors may claim that sympathetic pain cannot play these roles because some wrongs that ought to prompt remorse do not cause pain. Suppose I murder someone instantly and painlessly. Where is the pain with which I ought to sympathize? On Kant's account of sympathy, we can imagine our way into the perspectives of not just actual but possible others. We do this when we read fiction (Kant 2012, 25:476 [p. 53]). So, I can imagine my way into the perspective of a fictional version of the person I murdered and their profound sorrow over the life I have stolen from them. This should not seem *ad hoc*, as the foundation of sympathy in imagination is fundamental to Kant's account.

In conversation, Smilansky has claimed that sympathetic pain is too *weak* relative to the pain of self-retribution to be a powerful enough motivator to improve behavior. If we take this claim as empirical, then it requires empirical evidence, and while the respective roles of sympathy and self-retributive pain in moral psychology are a contested matter in empirical psychology, it is clear that sympathy plays a crucial role.[18] Even if, as an empirical matter, sympathetic pain is on average a weaker motivator than self-retributive pain, it seems plausible to think it can be a *strong enough* motivator to be valuable. If we take the claim that sympathetic pain is too weak as a claim about the *concept* of sympathy, then it is clearly false, at least as Kant construes sympathy. As we saw above, Kant's ideal of friendship is an ideal of *maximal* sympathy: I should imagine myself in my friend's place and try to feel their feelings as vividly as if they were my own. If we think victims are as morally salient as friends, and merit the same kind of sympathetic attachment, then Kant's ideal suggests that wrongdoers should imaginatively undergo everything they have done to their victims with equal vividity. The imaginations found among human beings at this point in our history typically present us with imaginings less vivid

[18] As noted earlier, this is often under the label of "empathy" (e.g., Decety, Jackson, and Brunet 2007).

than immediate sensations. However, Maysa Khedr points out that the progress of technology may give us more vivid imaginations.[19] Virtual reality may be contributing to this already, and direct brain interfaces may contribute more. If we have moral reasons to harness such tools to put us in the place of people with whom we ought to sympathize more vividly, then perhaps criminals ought to use them to experience what they have done to their victims just as vividly as the victims. This would yield an emotional proportionality of equality, in a kind of non-retributive parallel with the *lex talionis*. My point here is not to endorse this equal proportionality, only to show that it follows from a natural line of thought about Kant's ideal, to show that the concept of sympathy does not imply that sympathy is a weak emotional experience. Skeptics might be reluctant to endorse equal proportionality. It might seem to prescribe more pain than necessary for a fitting response to wrongs. This would give us a reason to constrain Kant's ideal when it prescribes pain. But skeptics could endorse equal proportionality without importing retributivist premises.

Conclusion

To conclude, let me comment on what readers may feel is a dissonance between these justifications of punishment and remorse. The former aims at making criminals' imprisonment as painless as possible, while the latter says that criminals should suffer along with their victims. The latter view may seem inhumane relative to the former. But care is typically understood as an essential feature of humaneness, and sometimes it hurts to care. Here is another point that may diminish dissonance. Principles of punishment chosen in OPD correspond to what Kant calls

[19] Class discussion, "Philosophy of Law," City College of New York, Fall 2021.

principles of right in that they justify coercion, while care-based reasons for remorse correspond to what Kant calls *duties of virtue*, with which we cannot legitimately coerce compliance. It is up to each of us, on our own, to fulfill our duties of virtue. Thus, measures such as coercively strapping sympathy helmets onto criminals would violate Kantian ethics. Further, the value of wrongdoers' sympathy for their victims may be reciprocal, such that victims have a duty of virtue to sympathize with wrongdoers too.[20] Victims' sympathy for wrongdoers' sympathetic pain can prompt them to offer wrongdoers opportunities to make amends, which can in turn promote reunification of moral communities. Adopting this attitude may seem to be a lot to ask of victims. But skepticism about action-based desert may make it easier to adopt. This concluding thought may be a first step toward an account of forgiveness for free will skeptics.[21]

References

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

Caruso, Gregg D., and Derk Pereboom. 2020. "A Non-Punitive Alternative to Punishment." In *The Routledge Handbook on the Philosophy and Science of Punishment*, edited by Farah Focquaert, Elizabeth Shaw, and Bruce N. Waller, 355-65. New York: Routledge.

Clark, Michael. 2004. "A Non-Retributive Kantian Approach to Punishment." *Ratio* 17, no. 1 (February): 12-27.

[20] I read Kant (1996b, 6:459-61 [pp. 577-78]) as supporting this view.

[21] Thanks to Jeffrey Blustein, Gregg Caruso, John Lemos, Neil Levy, Derk Pereboom, Saul Smilansky, Jens Timmermann, Bruce Waller, and Allen Wood for helpful advice at various points in the history of this project.

Decety, Jean, Philip L. Jackson, and Eric Brunet. 2007. “The Cognitive Neuropsychology of Empathy.” In *Empathy in Mental Illness*, edited Tom F. D. Farrow and Peter W. R. Woodruff, 240-60. Cambridge: Cambridge University Press.

Dolovich, Sharon. 2004. “Legitimate Punishment in Liberal Democracy.” *Buffalo Law Review* 7, no. 2 (January), 307-442.

Fahmy, Melissa Seymour. 2009. “Active Sympathetic Participation: Reconsidering Kant’s Duty of Sympathy.” *Kantian Review* 14, no. 1 (March): 31-52.

Harsanyi, John. 1975. “Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls’s Theory.” *American Political Science Review* 69, no. 2 (June): 594-606.

Jeppson, Sofia. 2021. “Retributivism, Justification and Credence: The Epistemic Argument Revisited.” *Neuroethics* 14, no. 2 (July): 177-90.

Kant, Immanuel. 1996a. *Critique of Practical Reason*. In *Practical Philosophy*, translated and edited by Mary J. Gregor, 137-271. Cambridge: Cambridge University Press, 1996.

———. 1996b. *The Metaphysics of Morals*. In *Practical Philosophy*, translated and edited by Mary J. Gregor, 363-602. Cambridge: Cambridge University Press, 1996.

———. 1996c. *Toward Perpetual Peace*. In *Practical Philosophy*, translated and edited by Mary J. Gregor, 315-51. Cambridge: Cambridge University Press, 1996.

———. 1997. *Lectures on Ethics.* Translated by Peter Heath. Edited by Peter Heath and J. B. Schneewind. Cambridge: Cambridge University Press.

———. 2007. *Anthropology from a Pragmatic Point of View*. Translated by Robert B. Louden. In *Anthropology, History, and Education*, translated and edited by Günther Zöller and Robert B. Louden, 231-429. Cambridge: Cambridge University Press.

———. 2012. *Lectures on Anthropology*. Translated by Robert R. Clewis, Robert B. Louden, G. Felicitas Munzel, and Allen W. Wood. Edited by Allen W. Wood and Robert B. Louden. Cambridge: Cambridge University Press.

Lemos, John. 2016. “Moral Concerns about Responsibility Denial and the Quarantine of Violent Criminals.” *Law and Philosophy* 35, no. 5 (October): 461-83.

Maslen, Hannah. 2015. *Remorse, Penal Theory, and Sentencing.* Oxford: Hart.

McLeod, Owen. 2013. “Desert.” *Stanford Encyclopedia of Philosophy* (Winter 2013 edition), edited by Edward N. Zalta. plato.stanford.edu/archives/win2020/entries/desert/.

Murphy, Jeffrie G. 1973. “Marxism and Retribution.” *Philosophy and Public Affairs* 2, no. 3 (Spring): 217-43.

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

———. 2014. *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

———. 2021. *Wrongdoing and the Moral Emotions.* Oxford: Oxford University Press.

Pereboom, Derk, and Gregg D. Caruso. 2018. “Hard-Incompatibilist Existentialism: Neuroscience, Punishment, and Meaning in Life.” In *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, edited by Gregg D. Caruso and Owen Flanagan, 193-222. Oxford: Oxford University Press.

Rawls, John. 1999. *A Theory of Justice*. Rev. ed. Cambridge, MA: Harvard University Press.

Shaw, Elizabeth. 2019. “Justice without Moral Responsibility?” *Journal of Information Ethics* 28, no. 1 (Spring): 95-130.

Smilansky, Saul. 2000. *Free Will and Illusion*. Oxford: Clarendon.

———. 2011. “Hard Determinism and Punishment: A Practical *Reductio*.” *Law and Philosophy* 30, no. 3 (May): 353-67.

———. 2019. “Free Will Skepticism and Deontological Constraints.” In *Free Will Skepticism in Law and Society: Challenging Retributive Justice*, edited by Elizabeth Shaw, Derk Pereboom, and Gregg D. Caruso, 29-42. Cambridge: Cambridge University Press.

Sterba, James P. 1977. “Retributive Justice.” *Political Theory* 5, no. 3 (August): 349-62.

Strawson, P. F. 1962. “Freedom and Resentment.” *Proceedings of the British Academy* 48, no. 1: 187-211.

Stueber, Karsten. 2006. *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences.* Cambridge, MA: MIT Press.

Vilhauer, Benjamin. 2004. “Hard Determinism, Remorse, and Virtue Ethics.” *Southern Journal of Philosophy* 42, no. 4 (Winter): 547-64.

———. 2009a. “Free Will Skepticism and Personhood as a Desert Base.” *Canadian Journal of Philosophy* 39, no. 3 (September): 489-511.

———. 2009b. “Free Will and Reasonable Doubt.” *American Philosophical Quarterly* 46, no. 2 (April): 131-40.

———. 2012. “Taking Free Will Skepticism Seriously.” *Philosophical Quarterly* 62, no. 249 (October): 833-52.

———. 2013. “Persons, Punishment, and Free Will Skepticism.” *Philosophical Studies* 162, no. 2 (January): 143-63.

———. 2015. “Free Will and the Asymmetrical Justifiability of Holding Morally Responsible.” *Philosophical Quarterly* 65, no. 261 (October): 772-89.

———. 2017. “Kant’s Mature Theory of Punishment and a First *Critique* Ideal Abolitionist Alternative.” In *The Palgrave Kant Handbook*, edited by Matthew C. Altman, 617-42. London: Palgrave Macmillan.

———. 2019. “Deontology and Deterrence for Free Will Deniers.” In *Free Will Skepticism in Law and Society: Challenging Retributive Justice*, edited by Elizabeth Shaw, Derk Pereboom, and Gregg D. Caruso, 116-138. Cambridge: Cambridge University Press.

———. 2021a. “‘Reason’s Sympathy’ and its Foundations in Productive Imagination.” *Kantian Review* 26, no. 3 (September): 455-74.

———. 2021b. “‘Reason’s Sympathy’ and Others’ Ends in Kant.” *European Journal of Philosophy*. doi:10.1111/ejop.12658.

———. Forthcoming a. “Sages, Sympathy, and Suffering in Kant’s Theory of Friendship.” *Canadian Journal of Philosophy*.

———. Forthcoming b. “Kantian Remorse with and without Self-Retribution.” *Kantian Review*.

Waller, Bruce N. 2015a. *The Stubborn System of Moral Responsibility*. Cambridge, MA: MIT Press.

——— . 2015b. *Restorative Free Will: Back to the Biological Base*. New York: Lexington.