4
Crime, Protection, and Compassion

Much seriously harmful wrongdoing perpetrated by human beings is criminal, and criminal acts and omissions are those that violate criminal law in some state and may as a result be prosecuted by that state.[1] A prominent justification for criminal punishment is retributive, which, in its classical form, invokes basic desert. The position on treatment of criminals I've proposed (Pereboom 2001, 158–86, 2013b, 2014, 153–74, 2020) rejects such retributive justification, since it presupposes that criminals are typically basically deserving of punishment that inflicts pain or harm, which I reject. Instead my position seeks to satisfy two aims: that we be protected from criminal wrongdoing, and that the well-being of the criminal be taken into serious consideration. These aims require balancing two sets of emotional attitudes: those that motivate protection against agents who pose serious threats, and those, like compassion, that are directed toward the well-being of the offender.[2]

The general skepticism I endorse about the control in action required for blame and punishment that involves basically deserved pain or harm has a role in justifying this position. But independently, our increased understanding of the neural and genetic bases for criminal behavior provides a reason to question whether criminals of certain common types have such control. The link between criminal behavior and psychopathy discussed in Chapter 3 is a case in point. On the genetic front, to cite one famous study, Avshalom Caspi and his research team analyzed data from 442 New Zealand male adults involved in a long-term study (Caspi et al. 2002). The researchers identified 154 subjects who were abused or maltreated as children, including thirty-three who were severely abused. The researchers then evaluated the influence of a particular gene on the abused children's outcomes as adults. A 'low-activity' variant of this gene which affects levels of monoamine oxidase A (MAOA), an enzyme that metabolizes the brain chemicals serotonin, dopamine, and norepinephrine, had previously been linked to

[1] This discussion is limited to criminal laws that are morally justified.
[2] This chapter is a revision of Pereboom (2020).

Wrongdoing and the Moral Emotions. Derk Pereboom, Oxford University Press. © Derk Pereboom 2021.
DOI: 10.1093/oso/9780192846006.003.0004

abnormal aggression. Caspi and his associates discovered that 85 percent of severely abused subjects with the low-activity variant of the MAOA gene developed some form of antisocial behavior. In contrast, study participants with the high-activity variant only rarely exhibited aggressive or criminal behavior in adulthood even if they had been severely abused as children. "Although individuals having the combination of low-activity MAOA genotype and maltreatment were only 12 percent of the male birth cohort," the researchers say, "they accounted for 44 percent of the cohort's violent convictions" (Caspi et al. 2002, 1–2). Adrian Raine provides a systematic treatment of our understanding of conditions of this sort in his landmark book *The Anatomy of Violence: The Biological Roots of Crime* (2013). We have reason to believe that certain common dispositions to criminal behavior are due to genetic and neural conditions, which, at least as matters currently stand, the criminal cannot control.[3]

I've argued that the optimal theory for criminal jurisprudence invokes incapacitation justified by the right to self-defense and defense of others (Pereboom 2013b, 2014, 2020), and as Gregg Caruso (2016, 2017, 2021) has proposed, embedded in a public health model.[4] In addition, I defend the claim that limited general deterrence can be justified when it is restricted by the Kantian injunction never to treat agents merely as means but always as ends in themselves (Kant 1785/1981, and that such general deterrence is likely required for a workable criminology.

Blame as Protest with Forward-Looking Aims

In Chapters 1 and 2 I argued that the main thread of the historical free will debate issues a challenge to one aspect of our practice of holding morally responsible, the sense of moral responsibility set apart by the notion of basic desert. The practice of holding morally responsible is complex, and it features several distinct aims and justifications. One widespread justification for punishing is that the agent who has knowingly committed a crime deserves it. One conception of desert is basic in the sense that it is not grounded in distinct and more basic moral considerations, such as maximization of

[3] This type of consideration is consistent with those affected by these conditions having or having the potential to develop capacities to counteract their effects in relevant circumstances.

[4] Caruso (2020a) also argues that this conception of how to deal with criminal behavior fits best with the Buddhist ethical view on which compassion is a central virtue.

good consequences. In another conception, desert is non-basic. Daniel Dennett (1984, 2003) and Manuel Vargas (2013, 2015) advocate versions of a view in which the practice-level justifications for punishment invoke desert, but that desert is not basic, because at the higher level the practice is justified by its anticipated good consequences.

In Chapter 2 I argued that there is also a largely forward-looking component to our practice of holding morally responsible, which aims at goods such as moral formation of character, reconciliation in relationships, protection from harm, and retention of integrity on the part of victims, which is not challenged by arguments in the free will debate. When a child misbehaves, a parent might blame and punish him because she believes that this is the best way to form good character, and not, or not only, to give him what he deserves. Blame in relationships that have been impaired due to bad behavior may have the aim of reconciliation. A victim of bullying might overtly blame the bully as a means to retaining his sense of integrity and as a way to protect potential future targets. One may object that blame essentially invokes desert, but in Chapter 2 I set out a notion of blame as moral protest (Hieronymi 2001; Talbert 2012; Smith 2013; Pereboom 2017a), but, distinctively, one that does not involve desert or the attendant negative reactive attitudes. Such moral protest might indeed have the aims of character formation, reconciliation in relationships, retention of integrity, and protection.

When one adopts the stance of moral protest against criminal behavior, it is appropriately accompanied by compassion. Buddhists traditionally argue that one right general attitude toward humanity generally is compassion (Goodman 2009). I agree. In her *Upheavals of Thought*, Martha Nussbaum proposes an analysis of compassion that takes Aristotle's account as its starting point. She begins by defining compassion as he did, as pain caused by the perception that someone has undeservedly suffered a misfortune that one is liable to suffer oneself.[5] On Nussbaum's view, by contrast with Aristotle's, emotions are appraisals with cognitive content, and she proposes this Aristotelian perceptual content as the basic cognitive content of compassion. She then embellishes this basis with three further specifications. First, when a subject has the emotion of compassion, she believes that the target agent's misfortune is seriously damaging to his well-being, and not merely minor. Second, she believes that the agent did not deserve this misfortune—that it was not his fault and not due to actions for which he is to blame.

[5] But see Rachana Kamtekar (2020) for an argument that for Aristotle the claim that the suffering must be undeserved is qualified.

Third, she believes that she herself is similarly vulnerable, that she could be the subject of the same misfortune (Nussbaum 2001; Deigh 2004).

In response to questions for this account raised by John Deigh (2004), Nussbaum (2004) comes close to retracting the third of these specifications. There are cases of compassion, such as those that concern animal suffering, in which one does not regard oneself as similarly vulnerable. Moreover, even with respect to beings relevantly different from oneself, imagination can arouse compassion. At this point only the first two provisions, that the misfortune be serious and that it be undeserved, remain in place. But what if there is no deserved pain or harm, at least not of the basic sort? What if we, with Śāntideva, see wrongdoing as issuing from causes beyond the agent's control, and as a result come to hold that wrongdoers never basically deserve to be harmed due to what they've done? Moreover, to be a wrongdoer is itself to be subject to a kind of misfortune. Plato maintained that it's intrinsically and non-instrumentally worse for a person to perpetrate than to suffer injustice (Gorgias 468d–79e; Kamtekar 2020), and Victor Tadros (2020) reports that in informal surveys in which the options presented are one's child being a serious wrongdoer and correctly convicted, and one's child being innocent and mistakenly convicted of the same wrongdoing, most respondents see the former as the greater misfortune. Accordingly, compassion is plausibly appropriate as an attitude for us to take toward criminals, as Plato and Tadros also maintain, and this provides motivation to see to their reform, reconciliation, and reintegration.

Retributivist Theories of Punishment

On the classical retributivist theory for the justification of punishment, the good to be achieved by punishment is that a wrongdoer receive the pain or harm he deserves just because of his having acted wrongly or wrongly omitted to act, given requisite cognitive sensitivity.[6] Classical retributivism as a theory of the justification of punishment is distinct from retributivism as a penalty schedule, roughly captured by the "eye for an eye" adage, although they are related; here I focus on the theory of justification.

Classical retributivism would be undermined if the free will skeptic is right about basic desert, since this view aims to justify punishment solely on

[6] For examples of classical retributivism, see Kant (1797/2017), Moore (1987, 1998), Husak (2000), Kershnar (2000), Morse (2004, 2013), Berman (2008), and Alexander and Ferzan (2009).

the grounds of basic desert, and the skeptical position contends that we lack the control in action required for punishment so conceived.[7] Compatibilists and libertarians reject this reason. But there is a powerful epistemic objection against compatibilists and libertarians who propose to justify criminal punishment on retributivist grounds. If the retributivist justification of punishment featured by our actual practice requires the rationality of the belief that compatibilism or libertarianism is true, while at the same time there are serious and unanswered objections to these positions, we cannot legitimately respond to a challenge to this part of the practice just by saying that it is supported by one of these views (contra Stephen Morse 2004, 2013). Punishment inflicts harm, and in general, justification for harm must meet a high epistemic standard. If it is significantly probable that one's justification for harming another is unsound, then, *prima facie* that behavior is seriously wrong, and one must refrain from engaging in it (Pereboom 2001, 161, 2013a, 2014, 158; Vilhauer 2009b; Caruso 2020b). A strong and credible response to the objections to compatibilism or libertarianism is required to meet this standard.

Another objection to classical retributivism derives from a generally accepted conception of the limited justifiable purview of the state. Would the legitimate role of the state include inflicting on people the pain or harm they basically deserve? Supposing that the requisite capacity for control in action is in place, and that basic desert could be secured as good or right, we nevertheless have reason to question whether the state has the right to invoke it in justifying punishment. The legitimate functions of the state are generally agreed to include protecting its citizens from significant harm, and providing a framework for constructive human interaction. These functions arguably underwrite justification that in the first instance appeals to prevention of crime. But they have no immediate connection to the aim of apportioning punishment in accord with basic desert. The concern can be made vivid by considering the proposal that the state set up institutions devoted to fairly distributing rewards on the grounds of basic desert. Wouldn't classical retributivism generalize so that the state would have as much reason to fund rewarding morally exemplary action as to fund criminal punishment (Pereboom 2013a, 2014, 159–60)?[8]

[7] There are views of how punishment is justified that are classified as retributivism that do not invoke basic desert (e.g., Morris 1968). I use the term 'classical retributivism' to distinguish the view under scrutiny from these alternative retributivisms.

[8] For further discussion of this issue, see Victor Tadros (2011, 69–83).

A further reason to doubt classical retributivism is that retributivist sentiments might well have their genesis in vengeful desires, and if so, retribution may be on no better footing than vengeance as a reason for punishing (Pereboom 2001, 160–1, 2013b, 2014, 158–9; Singer 2005; Greene 2008). Acting on vengeful desires may be wrong for the following reason. Although acting on vengeful desires can bring about pleasure or satisfaction, no more of a moral case can be made for the permissibility of acting on them than can be made for acting on sadistic desires. In each case, acting on the desire aims at the harm of the one to whom the action is directed, and in neither case does acting on the desire essentially aim at any good other than the pleasure of its satisfaction. But then, if retributivist motivations have their genesis in vengeful desires, acting for the sake of retribution, like acting on sadistic desires, stands to be morally wrong.

In response to this type of concern, Shaun Nichols (2013/2015) correctly points out that classical retributivist recommendations for state punishment differ importantly from practices such as the blood feud, a more direct expression of vengeful sentiments. Still, retributivist justifications, in his view, plausibly derive from such sentiments. But in defense of retributivism, Nichols writes that:

> the vast bulk of our ordinary ethical worldview likely derives from fundamentally arational emotional processes (Blair 1995, Prinz 2007, Gill and Nichols 2008). For instance, if we did not find human suffering aversive, we would likely not have the moral revulsion we do at killing. Nor would we have the moral norm of helping strangers . . . But notice how dramatic it would be to cast these norms out of morality. To limit our ethics to norms that have some ultimate rational justification would leave us with an ethics more barren than almost anyone would be willing to accept.
>
> (Nichols 2015, 133)

Nichols then considers the fact that retribution faces a competing consideration: "*ceteris paribus*, it's wrong to harm others" (Nichols 2015, 136). Relevant here is the point, made just now, that justification for harm must meet a high epistemic standard, and that if it is significantly probable that one's justification for harming another is unsound, then, *prima facie* that behavior is seriously wrong. Nichols contends that the competing consideration that it is wrong to harm others is not sufficient to overturn retributive norms generally, such as setting back cheaters at games by infliction of minor harms not fully justified by non-retributive justifications. He does affirm, however, that this competing consideration proscribes applying the retributive norm in support of the serious harming involved in criminal incarceration or the death penalty (Nichols 2015, 139), and on this point, the relevant one in this context, I agree with him. T. M. Scanlon (2013) endorses a similar position: there is basic desert for wrongdoing, but the most it can justify is the withdrawal of good will, and not measures as severe as paradigmatic criminal punishment, and on this last point again I agree.[9]

Michael McKenna (2020, 2021) advocates a retributive position on criminal punishment weaker than that of paradigmatic classical retributivists, but in which retributive considerations are stronger than they are for Nichols and Scanlon. McKenna's (2021) "wimpy retributivism" features, first of all, basically deserved blame and punishment, where such punishment involves an intention to harm or to set back the interests of the culpable, and the harm is conceived as a non-instrumental good. But then, to temper the harm-justifying force of this basic retributivism, McKenna cites the need to meet the high epistemic standard demanded by justifications of harming in the face of arguments for free will skepticism, the concern that many criminals are mentally ill, and the countervailing value of compassion. I differ from McKenna in that I believe the argument for free will skepticism is stronger than he thinks it is, whereupon the demanding epistemic standard for justifying harm, together with the concerns about the legitimate role of the state and about retributive sentiments being rooted in vengeance, rules it out as a legitimate justification for punishment. At the same time, tempering the force of retributive considerations as McKenna suggests stands to result in a punishment practice much closer to what I advocate than that justified by retributivism not similarly tempered.

Deterrence Theories of Punishment

These concerns for classical retributivism suggest we turn to the prospects for justifying criminal punishment by an appeal to its deterrent effect. On deterrence theories, it is the prevention of criminal wrongdoing that serves as the good by means of which punishment is justified. Initially, it would seem that no feature of free will skepticism renders deterrence theories less

[9] For additional general discussions of objections to classical retributivism, see C. L. Ten (1987, 38–65), John Braithwaite and Philip Pettit (1990, 156–201), and Philip Montague (1995, 11–23, 80–90).

acceptable to it than to libertarianism or to compatibilism. But at least some deterrence theories are not immune to the skeptic's challenge, since they presuppose a retributivist justification. Furthermore, like classical retributivism, deterrence justifications of paradigmatic sorts of punishment face difficult objections independent of skepticism about free will.

One paradigmatic deterrence theory is Jeremy Bentham's (1823/1948). In his conception, the state's policy toward criminal behavior should aim at maximizing utility, and punishment should be administered if and only if it does so. The pain or unhappiness produced by punishment results from the restriction on freedom that ensues from the threat of punishment, the anticipation of punishment by the person who has been sentenced, the pain of actual punishment, and the sympathetic pain felt by others such as the friends and family of the criminal. The most significant value that results from punishment derives from the security of those who benefit from its capacity to deter both the criminal himself as well as other potential criminals.

Arguably the most serious misgiving raised against utilitarian deterrence theory is the *use* objection. A general problem for utilitarianism is that it allows people to be used merely as means, that is, harmed severely, without their consent, to benefit others, and this is often intuitively wrong (Kant 1785/1981). Punishing criminals for the sake of society's security would appear to be just such a practice. At this point Dana Nelkin (2019b) suggests that we combine deterrence theory with a measure of retributivism. On her proposal, the criminal's basically deserving harm functions as counterweight to the use objection, and thus we can appeal to such desert to justify treating criminals in ways that subserve general deterrence. Perhaps many who believe the point of punishment is deterrence are implicitly relying on such a retributivist assumption. But if the free will skeptic is right, this proposal's reliance on basically deserved harm is undercut, and the concern about the grounding of retributivism in vengeance counts against it as well.

The deterrence theory developed by Daniel Farrell (1985); cf., Quinn 1985; Kelly 2009) potentially avoids the use objection by justifying criminal punishment not by consequentialist considerations but by the right of self-defense (and defense of others). Farrell's theory is impressive in part because it justifies punishment on grounds that are widely accepted, and which meet a plausible epistemic standard for justifying harm. Because free will skeptics can also endorse the right to harm in self-defense, this justification of punishment is available to them as well.

Special Deterrence and Self-Defense

Farrell's deterrence theory highlights the distinction between special deterrence—punishment aimed at preventing the criminal, specifically, from engaging in future criminal behavior—and general deterrence—punishment aimed at preventing agents other than the targeted criminal from doing so. In his view, special deterrence is significantly easier to ground in the right to harm in self-defense than is general deterrence. In broad outline, Farrell's justification of punishment as special deterrence is as follows. Each of us has the right of direct self-defense—your right to harm an unjust aggressor to prevent him from harming you or someone else; and the right of indirect self-defense—your right to threaten an unjust aggressor with harm to prevent him from harming you or someone else. The right of direct self-defense is limited in the following way: it is the right to inflict the minimum harm on an unjust aggressor required to prevent him from harming you or someone else. The right of indirect self-defense is the right to threaten to inflict this minimum harm on a potential unjust aggressor on the condition that he attacks. The right of direct self-defense permits you to carry out this threat against the aggressor once he has violated the condition of the threat, that is, once he attacks. But furthermore, because each of us has these rights, the state, acting as proxy for us, can legitimately issue corresponding general threats to harm potential unjust aggressors, and can also legitimately carry out such threats once their conditions have been violated. In this way, the right to self-defense can justify the state's practice of criminal punishment.

This special deterrence theory avoids some, and perhaps all, of the objections to its utilitarian counterpart. On the concern for justifying punishment that is intuitively too severe, one may not, on grounds of indirect self-defense, issue a threat to inflict harm more severe than the minimum required to effectively deter the targeted crime. So, if a threat of a month in prison would be sufficient to deter auto theft, the state may not issue a threat of a two-year term. On the concern for punishing the innocent, the right to self-defense justifies harming only the unjust aggressors themselves.

Harming an unjust aggressor in self-defense does involve harming him, without his consent, for the benefit of persons other than himself, and this arguably would count as an instance of using him merely as a means to the benefit of others. But as Tadros (2017) points out, this is a case of the *use of threat elimination*, intuitively justified by the right of self-defense, by contrast with the more controversial *manipulative use*, which is not justified on the basis of this right. Farrell points out that the theory he proposes will not extend to full-fledged general deterrence, for this would involve harming someone to prevent not just his aggression but also the potential aggression of others, and that would involve use of the manipulative kind. Farrell does contend, however, that some general deterrence can be justified on the basis of his principle of distributive justice. When an agent wrongs you in such a way as to make you more vulnerable than you would otherwise be to the aggression of others, then you are justified in countering just this degree of additional vulnerability by harming him. Since this use is justified on the basis of the right of self-defense, it qualifies as the use of threat elimination, and is not an instance of manipulative use.

A concern I've raised for Farrell's line of reasoning is whether it can justify punishment, that is, treatment that involves an intention to harm, by contrast with incapacitation, such as preventative detention in the case of violent criminals who continue to pose a threat (Pereboom 2001, 172–4, 2013b, 2014, 168–9, 2020). What makes it appear as if punishment can be justified as Farrell proposes, I've argued, is the model of an unjust aggressor in circumstances in which state law enforcement and criminal justice agencies have no role—a 'state of nature' situation. A state of nature situation in which an aggressor poses immediate danger is relevantly different from the circumstances of criminals in our society who are subjected to state punishment. When the state sentences them to be punished, they are in the custody of the law. Moreover, the harms that the right of self-defense justifies in the case of aggressors in a state of nature situation are often more severe than those that this right would justify for those in custody. Suppose you confront a late-night intruder, and he clearly aims to kill you. To prevent him from killing you, the right to self-defense justifies knocking him out with the golf club you've armed yourself with. It would then be also permissible, prior to knocking him out, to threaten him with this amount of harm. Suppose he attacks anyway, but in the process, he trips over your kids' electric train set, which allows you to pin him to the ground and tie him up with an extension cord. At this point is it still legitimate for you to knock him out with the golf club? To do so would be wrong, and not be justified by the right to harm in self-defense. This right justifies only what one would reasonably believe to be the minimum harm required to prevent the aggression. Or suppose an aggressor clearly aims to kill your friend, and to protect her it is legitimate for you to knock him out with your golf club and to threaten to do so. Suppose that despite your efforts, he kills her, but that subsequently he loses his balance, falls, and you tie him up. Is it then permissible for you to knock him out? Not on the basis of the right to harm in self-defense and defense of others—he no longer poses an immediate threat. You retain the right to protect yourself and others against him, but not by carrying out a threat designed to prevent a harm that now has already occurred. So, then, it may be that a threat one might justifiably make and carry out to protect against an aggressor in a state of nature situation is not legitimately carried out in a situation in which the aggressor is in custody.

But what is the minimum harm required to protect against a violent criminal in custody? It seems evident that nothing more severe would be required than isolating him from those to whom he poses a threat. Thus, it would appear that Farrell's reasoning cannot justify the imposition of *punishment* on criminals, exactly, such as the imposition of serious physical or psychological suffering. Rather in the case of violent criminals who continue to pose a threat, this reasoning would at best justify only preventative detention. I've developed this non-punitive alternative by an analogy between the treatment of criminals and the treatment of carriers of dangerous diseases. Ferdinand Schoeman (1979) argues that if we have the right to quarantine carriers of serious communicable diseases to protect people, then for the same reason we also have the right to isolate the criminally dangerous. Quarantining a person can be justified when she is not morally responsible—in any sense—for posing a threat to others. If a child is infected with a deadly contagious virus that was transmitted to her before she was born, quarantine can still be legitimate. Imagine that a serial killer poses a grave threat to a community. Even if due to mental incapacity he is not morally responsible for his crimes or for posing a threat, the justification for preventatively detaining him is at least as strong as is quarantining a non-responsible carrier of a serious communicable disease (Pereboom 2001, 174–7, 2013b, 2014, 169–73, 2020).[10]

It would be morally wrong to treat carriers of communicable diseases more severely than is required to protect from the threat they pose. Similarly, on the self-defense justification, it would be morally wrong to treat criminals more harshly than is required to protect against the threats they pose. Just as moderately dangerous diseases may only justify measures less intrusive than quarantine, so moderately serious criminal tendencies

[10] Perhaps the justification for preventative detention is stronger than it is for quarantine, for the reason that it is worse for a person to be a victim of injustice than to be a victim of a natural threat. For discussion of this issue, see Derek Parfit (1984, 47), Victor Tadros (2016, 162–6), and Zofia Stemplowska (2018).

may only justify responses less intrusive than detention. The self-defense justification motivates a degree of concern for the rehabilitation and well-being of criminals that would reform current practice, and here compassion is an appropriate emotion (cf., Menninger 1968). Just as society should seek to cure the diseased it quarantines, so it should prepare criminals for reintegration.[11] Different sorts of rehabilitation programs, including some with therapeutic components, have proven to be effective, including cognitive and behavioral therapies (Pereboom 2001, 178–86), and benign biological intervention such as Omega-3 therapy and non-invasive brain stimulation (Raine 2013; Focquaert 2019; Choy et al. 2020). These achievements provide grounds for hope for higher levels of success in the near future.

Gregg Caruso (2016, 2017, 2021) embeds the account just set out within a public health model, and I welcome this development. A primary aim of the public health system is prevention of disease. In the case of dangerous communicative diseases, it is only when prevention fails that quarantine is required. Similarly, the public health approach to criminal behavior would make prevention of crime a primary aim. This approach shifts the focus to identifying and addressing the social determinants of crime, which include poverty, low social-economic status, racism, systematic disadvantage, mental illness, homelessness, educational inequity, and abuse, which would reduce the need for incapacitation. Quarantine is only needed when the public health system fails to prevent dangerous communicable diseases. Similarly, a public health approach to crime would foreground prevention, and incapacitation would be used only when we fall short of that primary aim. In Caruso's conception, the social determinants of illness and of criminal behavior are interrelated, and we should adopt a broad public health approach to address the causal factors in each case. As in the case of the social determinants of illness, it is important to identify and take action on the social determinants of criminal behavior to enhance societal well-being.

How Much General Deterrence?

Incapacitation, and preventative detention in particular, may nevertheless involve serious harm—such as loss of liberty, personal relationships, and

[11] *Prima facie* duties to cure and rehabilitate are generally in place for those appropriately positioned. One might plausibly suggest that such duties are enhanced in cases in which those to be cured or rehabilitated are quarantined or preventatively detained without deserving such treatment.

potential for career development—even if it does not qualify as punishment. In addition, plausibly the state should not conceal the fact that it detains violent criminals on such grounds, but instead make this information publicly available.[12] So even though preventative detention is justified as special deterrence, such a policy, together with a publicity provision, would yield, as a side-effect, general deterrence; it would deter others who are tempted to commit crimes. This general deterrent effect comes for free, so to speak, since it is a side-effect of the state's satisfying a publicity provision on a legitimate policy of special deterrence, justified on the basis of the right of self-defense. I call general deterrent effects justified as special deterrence by the right of self-defense *free general deterrence* (Pereboom 2020).

Free general deterrence comes with a significant limitation on how much harm can legitimately be inflicted—as I've emphasized, only the minimum harm required to protect against an aggressor is licensed. One might propose, however, that the free sort isn't enough to protect against certain sorts of wrongdoing, such as manipulation of financial markets, large-scale embezzlement, and illegal use of political influence for gain in personal wealth and power. Those who commit such crimes are typically not poor or from disadvantaged backgrounds, and the public health model, as Caruso sets it out, is not conceived to prevent crimes of this sort. Instead, many of those who commit such crimes are wealthy and well educated, but willing to free-ride for reasons of self-interest. They are often good at calculating risk, at weighing the probability of the wrongdoing being detected against the probability of significant personal gain. Free general deterrence would arguably involve the threat of loss of one's professional or political position, or say of a license to trade in financial instruments. Whether such threats are sufficient to deter the crimes at issue is an empirical matter, but, in disagreement with Caruso (2021, chapter 9), I would wager that they are not. The general deterrence in place in the United States, for example, is already much stronger than what free general deterrence would allow, and yet the incidence of such financial and political wrongdoing is fairly high. Reducing

[12] Kant advocates a strict publicity requirement: "All actions that affect the rights of other human beings are wrong if their maxim is not consistent with publicity" (Kant 1793/1983, 135). But there is good reason to deny the general claim. Governments are not required to publicize how their computer security systems work, even though this relates to the right of other human beings. For an overview, see Gosseries and Parr (2018).

Publicizing preventative detention, in particular in ways that don't reveal particular identities, may involve use, but involve little or any additional harm. Tadros argues (in conversation) that the legitimacy of making such measures public can be grounded in duties wrongdoers have. His more general view is discussed below.

the strength of the deterrents is thus apt to increase the incidence of such wrongdoing. As noted, public health measures that aim to reduce poverty and environmental degradation, and improve access to health care and education, are mismatched for crime of this kind. These considerations motivate an attempt to justify a stronger sort of general deterrence than the free sort already defended.

One way of justifying a stronger sort of general deterrence is on grounds of basic desert, as Nelkin (2019b) proposes. The state's function includes deterring crime, but punishment justified on general deterrence grounds is subject to the manipulative use objection. Yet as long as criminals basically deserve punishment of a particular severity, in Nelkin's view it is legitimate to recruit that punishment to the service of general deterrence. But again, this line of justification is not open to a skeptic about negative basic desert. Tadros (2017) concurs in rejecting basic desert, but aims to justify stronger general-deterrence-subserving penalties on the basis of claims about duties. In the proposal he develops, the manipulative use objection can be answered by invoking duties that wrongdoers owe to victims. Like Nelkin (2019b), I have concerns for the view as he sets it out, but I believe that the kinds of considerations he invokes serve to justify some stronger general deterrence (Pereboom 2020).

Tadros begins by arguing that wrongdoers who are not deserving of harm may sometimes be manipulatively used for the purposes of general deterrence. He does affirm that it is often intuitively wrong severely to harm one person without her consent to benefit others, a claim he illustrates with the following example:

Bridge: Dorabella is on a bridge with Fiordiligi. A trolley is heading on a track under the bridge towards five people who will be killed if Dorabella does nothing. Dorabella can save the five only by throwing Fiordiligi from the bridge onto the tracks. Fiordiligi's body will stop the trolley, saving the five, but Fiordiligi will be killed. (Tadros 2016, 84)

It is wrong for Dorabella to throw Fiordiligi off the bridge, knowing that he will die as a result. But Tadros contends that manipulatively using a person for a greater good is not always wrong. Consider:

Wrongdoer on the Bridge: As *Bridge* except Fiordiligi has wrongly started the trolley in order to kill the five, simply because he will enjoy seeing them die. (Tadros 2016, 84)

Tadros judges that it seems permissible for Dorabella to use Fiordiligi in the way specified to save the five. But he acknowledges that the intuition might be due to the sense that Fiordiligi deserves to be harmed due to his wrongdoing. To correct for this Tadros proposes that the intuition that Fiordiligi is permissibly used withstands his being intentionally manipulated to act, as in my manipulation cases (e.g., Pereboom 2014, and set out in Chapter 1) which Tadros and I agree would rule out his deserving to be harmed:

Manipulated Wrongdoer on the Bridge: As *Wrongdoer on the Bridge*, except that scientists have manipulated Fiordiligi's brain to ensure that he acts wrongly. However, Fiordiligi fulfils all plausible compatibilist conditions of responsibility—his effective first-order desire to kill the five conforms to his second-order desires; his process of deliberation from which the decision results is reason-responsive, in that it would have resulted in his refraining from posing this threat were his reasons different; his reasoning is consistent with his character, because he is egoistic; but he sometimes regulates his behavior by moral reasons; he is not constrained to act as he does, and he does not act out of an irresistible desire. (Tadros 2016, 85)

Tadros has the intuition that this use is permissible, and about it he says: "if this intuition is sound, it is plausibly sound in virtue of the fact that responsibility for wrongdoing, in the compatibilist sense, makes a difference to a person's liability to be used, even when the wrongdoing is secured through manipulation." He then provides the following diagnosis:

> The manipulated wrongdoer on the bridge is heavily involved in the threat that the five face. He has a powerful reason to ensure that he is not the author of their deaths; much more powerful than the reason that innocent bystanders have to do so. If he could save their lives at some moderate cost to herself, he is required to do so. If he is thrown from the bridge to save the five, the cost that is inflicted on him is no greater than the cost that he would be required to bear in service of the end that he is used to serve. In that case, his complaint against being used in this way seems weak. (2016, 86)

I (now) agree with Tadros that Fiordiligi is liable to defensive killing.[13] The reason is that, as I argued in Chapter 3, wrongfully posing a lethal threat

[13] In earlier publications I reported that in this case it was my strong sense that it is wrong to throw the manipulated man off the bridge (Pereboom 2017b, 2020). But the position on

makes one thus liable, and Fiordiligi has in fact wrongfully posed a lethal threat. True, in *Manipulated Wrongdoer on the Bridge* the lethal threat, the trolley-in-motion, is in process in a way that it is not in typical examples employed to illustrate the right to defensively kill, in which killing prevents the lethal process—e.g., the shooting or the knifing—from being activated in the first place. But this is plausibly not a morally relevant difference.

However, while manipulative use by killing is sometimes justified, it may be justified largely in cases in which doing so is required to prevent another killing. We might now ask: what are the limits on manipulative use in which the right to defensively kill to prevent another killing isn't at issue? I've proposed that it is the human rights to life, liberty, and physical security of the person that have a key role in making the manipulative use objection to general deterrence intuitive (Pereboom 2020). Those rights are grounded in the more fundamental right to a life in which one's capacity for flourishing is not compromised in the long term. I've argued that there is a heavily weighted presumption (but not an absolute prohibition) against punishment as manipulative use when such use involves intentional killing, long-term confinement, and infliction of severe physical or psychological harm. But what if the proposed penalties are significantly less extreme, such as monetary penalties (Pereboom 2001, 177, 2017b, 2020)? Would it then be impermissible to use undeserving wrongdoers in ways that involve such penalties to subserve general deterrence?

As I've suggested, there may be circumstances in which effective general deterrence would require penalties more severe than can be justified on special deterrence grounds, and I cited manipulation of financial markets, large-scale embezzlement, and illegal use of political influence for gain in wealth and power. Plausibly this may also be so for less serious wrongdoing. Suppose preventing a shoplifter from future theft requires only monitoring with use of an ankle bracelet. The probability of shoplifters without monitoring devices being caught is low, and as a result, for quite a few people the expected net utility of shoplifting is relatively high. Now imagine that increasing the severity of the penalty for shoplifting to a substantial but not overly burdensome monetary penalty would reduce the incidence of shoplifting significantly relative to the threat of monitoring. Suppose also that it would reduce the cost of deterrence substantially relative to the monitoring policy. Would increasing the severity of the penalty be permissible in these

defensive killing that I've recently developed, in Chapter 3, changed my mind. Thanks to Carolina Sartorio for discussion of this issue.

circumstances? Note that such a fine, by contrast with the death penalty and long-term imprisonment, need not hinder the prospects for a life lived at reasonable level of flourishing.

Moreover, if manipulative use involving fines is within bounds, should we say the same for short prison sentences, say of several months? Mark Kleiman (2016) argues that short prison sentences are often especially effective deterrents, especially in combination with a high expectation of being apprehended. This suggests that short prison sentences should also be within bounds as penalty extensions justified on general deterrence grounds. This provision would also solve a problem Tadros (in conversation) raises: what if people refuse to pay the fines they've been assessed? Here it would be helpful to have a short prison sentence as a backup, in particular given their effectiveness as deterrents.

We can add that effective general deterrence involving threats of manipulative use may sometimes require treatment less severe than what effective incapacitation would demand. Imagine someone who is guilty of insider trading, displaying a disposition to flout the regulations when it is to her advantage and the probability of getting caught is sufficiently low. Suppose that our insider trader is in fact a self-interested expected utility maximizer. What would be justified by way of incapacitation grounded in the right to self-defense? Arguably, exclusion from arrangements in which self-interested expected utility reasoning would lead to law violation of the sort at issue, such as loss of trading license and exclusion from this type of job. However, here manipulative use designed to deter such law violation might well be less harmful to her. The state might, for example, threaten and impose a substantial fine on general deterrence grounds, which on balance might well be less harmful to such offenders than the exclusion.

Schematically, the proposal is as follows:

General Deterrence Prerogative: If imposing a penalty on an offender on special deterrence grounds can be justified, imposing a somewhat more exacting penalty, not justified on special deterrence grounds, is justified if it (i) substantially increases general deterrence value, and/or (ii) substantially lowers the cost of deterrence, provided that the more exacting penalty doesn't hinder the prospects for a life lived at a reasonable level of flourishing. (Pereboom 2020, 94)

The rationale for this proposal has several components. First, one need not be a consequentialist to agree that consequences have weight when deciding moral and legal issues. The general deterrence prerogative specifies only that they have modest additional weight when special deterrence justification is already accounted for. Policies we all accept that would have to be justified in this way are already in place. We all accept that it's legitimate for the police to apprehend suspects of crime when there is adequate but nevertheless insufficient reason to believe that they are in fact criminals. This is a significant cost that we not infrequently impose on people who are in fact innocent, and this cost would be difficult to justify on other than consequentialist grounds.

Second, this account sets a credible standard for a weighted presumption against manipulative use: what justification might it have? Both the criminal who is given the short prison sentence for reasons of general deterrence, and the one who is made to serve life in prison or executed for this reason are being used as means for the safety of society. A pertinent question is: is each being used merely as a means? In the Kantian conception (Kant 1785/1981, this depends on whether he is also being treated as an end in himself. There are a number of accounts as to what this comes to. An attractive option is an elaboration of the idea that to treat a person as an end is to treat her in such a way as to facilitate her capacities and opportunities for developing herself as an autonomous, rational being. On capacities, we needn't privilege rationality: we can add other characteristics we value, such as the capacity for fulfilling personal relationships, the capacity to create and appreciate artistic products of culture, and the ability to excel in and value activities such as sports and physical labor. To treat someone as an end is to treat her in such a way as to allow her to flourish by developing such capacities in accordance with her preferences. Executing someone is clearly in violation of treating someone as an end in this sense, as is serving a life sentence in a standard American maximum-security prison. But a month in prison, with provision for education while confined and effective reintegration upon release, need not violate this standard. While such a short prison term is a violation of the liberty right, it is only a moderately serious violation, and will not in many cases preclude a life lived at a reasonable level of flourishing in the way that long prison terms typically do.

Might there be a non-consequentialist and non-desert-based justification for the General Deterrence Prerogative—for inflicting penalties on criminals for reasons of general deterrence that are somewhat more exacting than those justified on special deterrence grounds? As I noted above, Tadros (2016) develops a view of this sort that crucially invokes duties criminals have to their victims. In one of his examples, Dave, a lorry driver, involuntarily and non-culpably injures Veronica. Tadros contends, plausibly, that Dave has a more stringent duty to assist Veronica than does Xavier, a bystander. Suppose that instead Dave voluntarily injured her while satisfying the compatibilist conditions on moral responsibility. Tadros maintains that now Dave incurs even more stringent and extensive duties of this sort, even if factors beyond his control causally determine him to act, and even if he therefore doesn't basically deserve to have the cost imposed on him that carrying out these duties involves (Tadros 2016, 77–9). By analogy, it's plausible that those who commit crimes have a collective duty to compensate society that non-criminals lack, even if they don't deserve to suffer the harm involved in making this compensation. By virtue of committing crimes, criminals, as a sector of society, collectively make a costly criminal justice system necessary.

The right to self-defense justifies free general deterrence, in accord with Farrell's view. We can think of this, metaphorically, as a fence the state sets up against criminal behavior. Suppose Zoë builds a fence around her garden to prevent rampant plant-trampling. One night, Alice and Bob tear down part of the fence, enter the garden, and trample the plants. Would it be legitimate to require Alice and Bob to reconstruct the fence, on grounds other than basic desert, and not require innocent Chloë and Dan, who are also available, to help? As David Boonin (2008) and Tadros (2016) argue, duties of compensation can plausibly be supported on grounds other than desert. Suppose I accidentally break my aunt Ellen's vase, but I wasn't culpably negligent. It nevertheless seems reasonable to expect that I compensate by, for instance, replacing the vase, despite not deserving to bear this cost. By analogy, it's credible that we can reasonably expect criminals collectively to compensate, in part, for the expense of the criminal justice system, even if they don't deserve to bear that cost or to suffer the harm that such compensation may involve. As Tadros (2017) argues regarding one specific type of compensation he regards as permissibly imposed on wrongdoers: "The fact that wrongdoers wrongly lead us to be vulnerable to attack by others by undermining the credibility of our threats may be sufficient to render it permissible to use them" (Tadros 2017, 615), and this, in his view, is so regardless of considerations of desert. Note again that the objective is only to justify penalties somewhat more severe than those justifiable on special deterrence grounds alone. For this reason, such considerations needn't be especially weighty.

A Comparison with Dennett's Position

As noted earlier, Daniel Dennett (especially in Dennett and Caruso 2020) advocates a position, like Manuel Vargas's (2013), in which the practice-level justifications for blame and punishment cite backward-looking considerations of desert, while such desert is not conceived as basic because at a higher level the practice is justified by its forward-looking aims. These aims include enhancing the ability to recognize and respond to moral considerations and protecting people from the dangers wrongdoers pose. On Dennett's account, our actual practice insofar as it involves punishing criminals because they deserve it should be retained since doing so has the best overall consequences relative to alternative practices. His view has a contractualist element: we tacitly consent to rules for behavior and for penalties imposed for crime that we, as idealized consequentialist reasoners, would formulate and endorse. Note that Dennett is also a revisionist about punishment relative to actual practice in the United States; he believes that much of it is unjust and requires reform.

As discussed in Chapter 2, Dennett employs sports analogies to confirm that his non-basic desert is genuine desert (Dennett and Caruso 2020; cf., Doris 2015; Vargas 2015). It seems legitimate to say that someone who commits a foul in basketball deserves the penalty for that foul. But such sports desert isn't basic—it's instead founded in considerations about how basketball works best as a sport. Similarly, suppose penalties for criminal behavior are justified on forward-looking, deterrence grounds, in virtue of the anticipated effect of safety. Then it similarly makes sense to say that penalties are deserved. From my perspective the crucial point of agreement with my view, and also Caruso's (2021), is that the fundamental justifications are ultimately forward looking, and justifications that appeals to basic desert are ruled out. Dennett in fact claims that 'basic desert' is an incoherent notion; in my view, attributions of basic desert to human beings are instead coherent but false. Dennett's more general view is that there is no basic desert moral responsibility and no libertarian free will, but there is deserved punishment and we do have free will. The way he puts the point is that we have all the desert and free will worth wanting, while the stronger notions are incoherent or at least clearly don't apply to us.

On deserved punishment, Dennett's view is compatible with the main claims of mine and with Caruso's. Caruso and I can allow that the player who hands the ball in soccer deserves to have the specified penalty imposed.

Similarly, we can affirm that the insider trader deserves to have his trading license suspended, on the supposition that license suspension is the penalty that idealized forward-looking reasoners would specify for insider trading. Part of Dennett's conception is that the moral game needs knowable rules with specified penalties so that players can anticipate what will happen if they violate the rules. I agree. I don't see violent crime exactly in these terms. I say: if agents manifest a disposition to extreme violence, it is legitimate to preventatively detain them, justified on analogy with quarantine. Perhaps the meaning of 'desert' is sufficiently unrestricted for preventative detention to then count as deserved, but in my view not much depends on whether we use the term 'desert' in this context. What's key is that such measures are not justified on grounds of basic desert, and Dennett concurs.

Objections

Let us now consider several objections, each of which is a good challenge that occasions clarification of the position I'm proposing. First, Saul Smilansky (2017) objects that the justified detention of the criminally dangerous on the quarantine analogy will yield insufficient and inadequate deterrence. He contends that on this model, those who are detained would need to be compensated for their confinement by what he calls *funishment*, a paradigm of which he once specified as equivalent to a stay in a five-star hotel (Smilanksy 2011). Neil Levy (2011) and I (Pereboom 2014, 172–3; cf., Pereboom and Caruso 2018) disagreed, and I argued that less opulent accommodations and programs for rehabilitation and reintegration would be in order. Smilansky (2017) replied that two-star accommodation would also not yield adequate deterrence, and that therefore a harsher environment, justified on retributive grounds, would be required instead.

But in addition to detention justified by analogy to quarantine, further sorts of monitoring, and programs for rehabilitation and reintegration, the model I advocate includes general deterrence by monetary penalties and short-term prison sentences. This yields a response to one example Smilansky provides, greedy relatives who murder in order to secure an inheritance. Their motive is financial gain, and it stands to reason that they would be deterred by a credible threat of dispossession. Such monetary penalties can also serve as a deterrent for the spousal killer who poses no other genuine threat, although credible examples of this phenomenon may be extremely rare. Here limited prison sentences may also be effective, in particular in combination with a high expectation of being caught.

Smilansky's inadequacy claim is empirical, and there is empirical evidence that bears on the issue. Currently there is widespread discussion of the difference between the American model for criminal justice and those that we find in countries such as Norway, Sweden, Finland, Denmark, and the Netherlands. In Norway, for example, the aim of the criminal justice system is at least largely protection and reintegration, and famously, prisons are indeed the equivalent of two-star hotels. But in these countries crime and recidivism rates are much lower than they are in the United States, whose criminal justice system is closer to what Smilansky envisions. The reasons for differential success in deterrence and prevention between these countries and the United States are undoubtedly complex, and some argue that the policy is not feasible in the American context. But the success of such a policy counsels against ready acceptance of the claim that harsher prison conditions of the sort that Smilansky advocates should generally be preferred to alternative measures.

Second, Michael Corrado (2016), John Lemos (2016), and Smilansky (2017) object that implementation of this account would draw too many people into the criminal justice system. In particular, it would lead to incapacitating those who pose threats but have not yet committed crimes. As a remedy, Smilanksy maintains that retributivism can have the role of limiting incapacitation to an intuitively plausible degree (cf., Hodgson 2012). In response, I doubt that retributivism can effectively play this role (cf., Caruso 2021), and I contend that this account has other resources to safeguard the right people from the criminal justice system.

A concern for preventative detention that I've emphasized in the past (e.g., Pereboom 2014, 170–1), is that, for example, neural tests for determining whether someone is likely to commit a crime are invasive and may seriously conflict with the right to liberty, and current neural tests are not especially reliable and frequently yield false positives (Nadelhoffer and Sinnott-Armstrong 2012; Nadelhoffer et al. 2012). But still, better tests are being developed (Nadelhoffer et al. 2012). In *Free Will, Agency, and Meaning in Life* (2014, 170) I present an example in which an agent has been given a drug without his knowledge, and we can determine that as a result he will almost certainly commit a crime within a week. After a week the effect of the drug wears off. I suggested that the state is entitled to detain him for that week. Corrado (in correspondence) allows that if the drug impairs his reasons responsiveness, preventative detention may be permissible. Smilansky might be attracted to this kind of position: if an agent is dangerous but not sufficiently reasons responsive, he may be detained. But what if someone is dangerous and sufficiently reasons responsive? Corrado's (1996) position is that then he may not be detained unless it can be shown that he has a current intention to cause harm. As Corrado indicates, delineating the particular features of intention (e.g., how specific does it need to be?) is a delicate and difficult issue. But this general sort of position seems reasonable to me.

Corrado (in correspondence) suggests that a test for the demonstrable intention model is the landmark legal case *Tarasoff* v *The Regents of the University of California* (1974). The case involves Prosenjit Poddar, who confided his desire to kill a young woman, Tanya Tarasoff, to his therapist. The therapist believed the threat to be serious enough to have Poddar preventatively detained, but the therapist was overruled by his supervisor. Prior to these events, Poddar had been civilly committed as a dangerous person but was then released when he appeared rational. Poddar then killed Tarasoff. Subsequently, the doctor and his employer, the University of California, were sued by Tarasoff's family. The Supreme Court of California decided that the defendants were liable to the family, not because they hadn't detained Tarasoff, but for the reason that they hadn't warned her of the threat that Poddar posed to her. The case established a duty on the part of therapists to warn, but only where a specific victim was targeted. A general prediction that some unspecified person would be harmed would not justify a duty to warn.

In other jurisdictions, such as Ontario, Canada, the state has the right to detain the dangerous when rationally competent, albeit under mental health legislation.[14] It seems to me that the Ontario policy, supplemented with the demonstrable intention requirement Corrado proposes, is preferable (Corrado 1996; Pereboom 2017c, 2020). Tarasoff should not have been subjected to the burden of protecting herself against someone with a demonstrable intention to kill her. Would Smilansky agree? If not, would he have allowed Tarasoff to be subjected to the burden of self-protection? But if he does agree, then retributivism cannot play the detention-limiting role he advocates for it.

A more general concern of Smilansky's is that we not treat criminals unjustly, and if they don't deserve to be harmed, it's unjust to harm them. I share this concern. He and I both believe that the arguments for free will skepticism—that we lack the control in action required to ground desert (or

[14] Thanks to Jennifer Chandler for this information.

at least basic desert)—are strong (Smilansky 2000; Pereboom 2001, 2014). But in his response, by contrast with mine, Smilansky (2000) advocates retaining the illusion of free will and desert, even when justifying criminal punishment. However, this involves treating people unjustly by his own standard, since he believes that arguments for free will skepticism, and thus against the view that wrongdoers deserve to be treated harshly, remain unanswered. I agree that dangerous criminals don't basically deserve to be incapacitated, but that our right to defend ourselves provides an alternative reason for incapacitation, and in a limited respect, consequentialist considerations and duties owed by wrongdoers also count. So even if harming criminals is in an important sense unjust because undeserved, and should concern us for this reason, doing so in the limited ways I've specified is nevertheless justified.

Smilansky (2017, together with others such as David Hodgson (2012), cite as a reason for adopting retributivism that it can ensure that only the guilty, and not the innocent, are punished. On this suggestion, the best way to secure this good is by way of a certain legal practice—by lawyers, judges, and juries justifying their decisions at least in part on grounds invoking desert. Notice, however, that on this conception the desert invoked won't be basic, since the practice is justified at least partly on the ground that it's the best way to secure a good consequence—that the innocent not be punished. By contrast, we might imagine someone who does think that basic desert justifications are in place and who cites this benefit only as a side-effect. But this exact view isn't available to Smilansky, since he maintains that we don't have free will, that we lack the control in action required for basic desert attributions. Again, he is, by contrast, an illusionist about free will and about basic desert. In effect, Smilansky is contending that we must maintain the illusion about free will and basic desert for the sake of a good consequence, that the innocent not be punished. Thus he is in fact invoking non-basic desert, and not basic desert, in his account. Hodgson (2012), who also cites protection of the innocent from punishment as a benefit of belief in desert, is not an illusionist about free will. He can, by contrast, consistently invoke basic desert in his proposal.

Are Hodgson and Smilansky right to think that commitment to desert-based legal justifications would have the effect of protecting the innocent from punishment? For this to be so, it must be that someone's not deserving to be harmed is sufficient reason for the state not to harm that person for the sake of a further state interest. The problem is that there is another competing state interest that justifies harm: protection from threats (Caruso

(2021) argues similarly). Those who pose threats may uncontroversially not be deserving of harm, such as the mentally ill lethal threats we considered in Chapter 3. Or imagine someone who has been given a drug, without his knowledge, that makes him prone to extreme violence for a short time, and that the only way to stop him from killing someone is to incapacitate him with a painful taser. This is clearly legitimate. Or suppose that the drugged person is about to shoot as many students in school as he can, and the only way the police can stop him is to kill him. This is also legitimate. Thus, not being deserving of harm does not insulate a person from being justifiably harmed by the state on the basis of its interest in protection. Furthermore, and particularly troublesome for Hodgson and Smilansky's proposal, when innocent people are presumed to be threats, the belief that only the guilty should be detained or killed is often ineffectual. As Caruso (2021) points out, in the United States belief in retributivism is strong in regions in which convictions of the innocent who are believed to be threats, often unjustifiably, are also prevalent, particularly when the innocent are African-American and Latino men. This is consistent with belief in desert and in retributive justification for punishment reducing the incidence of such convictions, but it's not clear that these beliefs actually have this effect.[15]

Summary and Conclusion

My aim was to set out a theory for treatment of criminals that rejects the retributive justification for punishment, does not fall afoul of a plausible prohibition on using people merely as means, and can actually work in the real world. The proposal is largely justified as special deterrence by the right to self-defense and defense of others, as in Farrell's (1985) theory. My account adds the quarantine analogy-based rationale for preventatively detaining criminals together with provisions for rehabilitation and reintegration, and a justification for somewhat more exacting penalties to secure effective general deterrence, measures that cannot be justified as special deterrence by the self-defense right. Here consequentialist considerations and duties of compensation have a modest, but to my mind plausible, justificatory role.

[15] This observation also casts doubt on the related claim that belief in retributivism has the effect of limiting the severity of punishment. For a response to this claim, see Victoria McGeer (2013, 187–8).