A Defense of Free Will Skepticism: Replies to Commentary by Victor Tadros, Saul Smilansky, Michael McKenna, and Alfred R. Mele on *Free Will, Agency, and Meaning in Life*

Derk Pereboom, Cornell University

Penultimate draft

Abstract: This paper features my replies to commentaries by Victor Tadros and Saul Smilansky on his non-retributive, incapacitation-focused proposal for treatment of dangerous criminals; by Michael McKenna on his manipulation argument against compatibilism about basic desert and causal determination; and by Alfred R. Mele on his disappearing agent argument against event-causal libertarianism.

Let me begin by thanking my four commentators, Victor Tadros, Saul Smilansky, Michael McKenna, and Al Mele for their excellent, thoughtful, and constructive comments, and for valuable subsequent discussion. Each prompted me to refine and clarify my views.

Tadros and Smilansky on Criminal Justice

Victor Tadros (2016) and Saul Smilansky (2016) raise objections to the account of criminal justice I advocate in (2001) and (2014). In short, I defend a proposal for treatment of dangerous criminals that invokes our right to protect ourselves and to secure our safety, and employs the analogy to quarantine (Pereboom 2011, 2014; cf. Schoeman 1979; Caruso 2012, 2016; Pereboom and Caruso, forthcoming). This incapacitation view draws on a comparison between treatment of dangerous criminals and treatment of carriers of dangerous diseases. The free will skeptic claims that criminals are not morally responsible for their actions in the basic-desert sense (Feinberg 1970; Pereboom 2001, 2014; Scanlon 2013). Most carriers of dangerous diseases are not responsible in this or in any other sense for having contracted these diseases. Yet we generally agree that it is sometimes permissible to quarantine them, and the justification for doing so is the right to self-protection and the prevention of harm to others. For similar justificatory reasons, even if a dangerous criminal is not morally responsible for his crimes in the basic-desert sense, it could be as legitimate to preventatively detain him as to quarantine the non-responsible carrier of a serious communicable disease.

It is important to see that this analogy places several constraints on the treatment of criminals. First, the view is founded in the right to defend oneself and others against danger, and this right allows one to inflict only the minimum harm required for protection. Second, as less dangerous diseases justify only preventative measures less restrictive than quarantine, so less dangerous criminal tendencies justify only more moderate restraints. For instance, for certain minor crimes perhaps only some degree of monitoring could be defended. Third, the incapacitation account that results from this analogy demands a degree of concern for the rehabilitation and reintegration of the criminal that would alter much of current practice. Just as fairness recommends that we seek to cure the diseased people we quarantine, so fairness would counsel that we attempt to rehabilitate and reintegrate the criminals we detain. If a criminal cannot be rehabilitated, and our safety requires his indefinite confinement, this account provides no justification for making his life more miserable than would be required to guard against the danger he poses.

Both Tadros and Smilansky focus on the implications of my view for use of punishment as a general deterrent. Tadros contends that free will skepticism can justifiably accept more punishment justified on the ground of general deterrence than I allow. Smilansky argues that what I do allow by way of treatment of criminals will be insufficient in view of the need to deter criminal behavior, and that this amounts to a practical *reductio* of the position. In response, I first set out in more detail my justification of the detention of dangerous criminals on the ground of special deterrence, on analogy with the quarantine of carriers of dangerous diseases. Then, in response to Tadros, I develop in more detail my view (2001: 177) that a general deterrence system involving monetary penalties is consistent with free will skepticism and with the kind of prohibition on using people as means that I endorse.

The classic deterrence theory is the utilitarian version, advocated, for example, by Jeremy Bentham (1823/1948). In his conception, the state's policy on criminal behavior should aim at maximizing utility, and punishment is legitimately 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 pleasure or happiness that results from punishment derives from the security of those who benefit from its capacity to deter.

Of the objections raised against Bentham's position the one that has a prominent role in what follows is the use objection, which is a problem for utilitarianism more generally. In certain possible scenarios, utilitarianism requires people to be harmed severely, without their consent, in order to benefit others, and this is often intuitively wrong. The best option for justifying a policy for treatment of criminals invokes not utilitarianism, but instead the right to self-defense and defense of others (Pereboom 2001, 2014). Several deterrence theorists have argued that criminal punishment can be justified on such grounds (Farrell 1985; cf. Quinn 1985; Kelly 2009).

Daniel Farrell's account highlights the distinction between special deterrence – punishment aimed at preventing the criminal himself from engaging in criminal behavior – and general deterrence – punishment aimed at preventing agents other than the targeted criminal from doing so. In his view, special deterrence is significantly easier to ground in the right to harm in self-defense or defense of others than is general deterrence. His view also invokes a distinction between the right to direct self-defense, your right to harm an unjust aggressor to prevent him from harming you or someone else, and the right to indirect self-defense, your right to threaten an unjust aggressor with a reasonable amount of harm to prevent him from harming you or someone else. In broad outline, Farrell's justification of punishment as special deterrence is this. Each of us has the right to direct self-defense, and each of us also has the right to indirect self-defense. Because we have the right to direct self-defense, we have the right to inflict a reasonable amount of harm on a potential unjust aggressor to prevent him from harming us. Because we have the right to indirect self-defense, we also have the right to threaten to inflict this amount of harm. Our right to direct self-defense permits us to carry out this threat against him once the condition of the threat has been violated. But also, because we have these rights, the state, acting as proxy for us, may issue appropriate general threats to harm unjust aggressors, and may carry out such threats once their conditions have been violated. In this way, the right to self-defense can ground a legitimate state institution of punishment as special deterrence.

This special deterrence theory avoids some of the key objections to its utilitarian counterpart. On the concern for justifying punishment that is intuitively too severe, one may not, on the ground of indirect self-defense, issue a threat to inflict harm that is more severe than the minimum required to effectively deter the crime at issue. So, if a threat of one year in prison would be sufficient to deter auto theft, the state may not issue a threat of a 10 year term. On the concern for punishing the innocent, the right to self-defense justifies harming only unjust aggressors themselves. For instance, the right does not justify harming an unjust aggressor's innocent children even if this would deter him.

Still, harming an unjust aggressor in self-defense or defense of others does involve harming him, without his consent, for the benefit of persons other than himself, and this arguably would count as an instance of using him as a means to the benefit of others. Perhaps this is a legitimate type of use because its target brings it upon himself by his unjust aggression. But this proposal may invoke the notion of basic desert. Farrell in fact argues that the right to self-defense assumes a form of retributivism, albeit a weak kind. Underlying the right to direct and indirect self-defense, and hence also special deterrence on his account, is a "weakly retributive" principle of distributive justice, that if an aggressor forces me to make a choice between harming the aggressor or allowing myself or others to be harmed, then I am permitted to harm the aggressor to the degree that preventing the harm to myself or others requires (Farrell 1985: 385). If this principle is in fact retributive, and thereby presupposes basic desert, and it does in fact underlie the right to self-defense, then the arguments for the skeptical position imperil this right and a theory of punishment on which it is based. However, it is incorrect, I think, to call this principle "retributive" if in doing so basic desert is invoked. For this principle, and the right to harm in self-defense more generally, very plausibly apply to aggressors who are not morally responsible in the basic-desert sense, such as people who have been brainwashed, the significantly mentally impaired, and animals.

However, the use problem now arises again. Consider the aggressor who threatens to seriously injure me, and who I then harm in self-defense. Am I not using him merely as means to secure my own safety? I specified (2014: 167) that my answer is affirmative, but that, in such cases, if the harm inflicted is the minimal amount reasonably required to prevent the serious injury, the force of the use objection is outweighed by the right to harm in self-defense. The use objection has more force against general deterrence theories of punishment. Farrell contends that the type of theory he proposes will not extend to full-fledged general deterrence, for this would involve harming someone not just to prevent his aggression, but also the potential aggression of others, and this gives rise to a convincing use objection. But he argues that some general deterrence can be justified on the basis of his principle of distributive justice. When an agent wrongs you in such a way as to make you more vulnerable than you would otherwise be to the aggression of others, then you are justified in countering just this degree of added vulnerability by harming him. My sense is that in such cases the force of the use concern is also plausibly outweighed by the right to harm in self-defense, by contrast with a practice of full-fledged general deterrence.

So far, it appears that Farrell has proposed a justification for criminal punishment that the free will skeptic can endorse. And I do accept some of its core features. But I think (2014) that Farrell's line of reasoning can justify, in the case of a dangerous criminal, preventative detention but not imprisonment conceived of as punishment, where punishment is the intentional imposition of harm on the criminal for the reason that he has done wrong (2001: 159; in his comment McKenna suggests the intentionality requirement, which I accept). What makes it appear that punishment can be justified in this way is the model of an unjust aggressor in a situation in which state law enforcement and criminal justice agencies have no role – let's call this a "state of nature" situation. A state of nature situation in which an aggressor poses an immediate danger is very different from the circumstances of criminals in our society in which state punishment is carried out. Criminals are then in the custody of the law. Crucially, the kinds of harms that the right to self-defense and defense of others justifies in the case of an aggressor in a state of nature situation differ from those that this right justifies when he is in the custody of the law (Pereboom 2001: 172-4; 2014: 168-9). If one proposes to harm him more severely, for instance to provide credibility for a system of threats, the right to harm in self-defense would not supply the requisite justification.

What is the minimum amount of harm required to protect ourselves from a dangerous criminal in custody? It seems evident that nothing more severe would be required than isolating him from those to whom he poses a threat. Thus Farrell's reasoning cannot justify punishment of criminals by imprisonment or other intentional infliction of severe physical or psychological pain. Rather, in the case of violent and dangerous criminals, this reasoning would at best justify only incapacitation by preventative detention.

Tadros (2016) challenges my reasoning against Farrell's view. He begins by citing the fact that I specify that both self-defense and deterrent harming involve using the person. In the case of self-defense, I contended that the use objection is outweighed because using the person is necessary to avert a threat that he poses. But, Tadros argues, Farrell-style deterrence has the same features:

> The wrongdoer poses a threat – if we do nothing to respond to her wrongdoing, she will have caused others to cause us wrongful harm. We can avert this threat in which she is causally involved only by harming the wrongdoer. If we do the minimum harm to the wrongdoer that is necessary to avert this threat, what objection does Pereboom have to the fact that the person is being used? It is no good for Pereboom simply to point to the fact that she is being used, for he believes that is also true in standard cases of self-defence, yet he thinks that the objection is outweighed in that case. He must find some other difference between self-defence and Farrell-style deterrence, and show that this difference rules out the latter, even once the former is accepted. But he has found no such difference.

Tadros then argues that I should not have agreed that harming in self-defense involves use in the familiar sense at issue in the debate; there it is just eliminating a threat. But Farrell-style deterrence does not involve threat-elimination, but as he puts it, “manipulative using.” The core question at this point, Tadros contends, is whether wrongdoing can justify the manipulative use of the wrongdoer to avert a threat, supposing the absence of desert.

Tadros agrees that it is often intuitively wrong to harm severely one person without her consent to benefit others. He illustrates with this example:

> *Bridge*: X is on a bridge with Y, an innocent bystander. A trolley is heading on a track under the bridge towards five people who will be killed if X does nothing. X can save the

> five only by throwing Y from the bridge onto the tracks. Y’s body will stop the trolley, saving the five, but Y will be killed.

It seems wrong for X to kill Y. But Tadros proposes that, while manipulatively using a person for the greater good often seems wrong, it is not always wrong. Consider:

> *Wrongdoer on the Bridge*: As *Bridge* except Y has wrongly started the trolley in order to kill the five, simply because she will enjoy seeing them die.

Tadros judges that it seems permissible for X to use Y to save the five. But he acknowledges that the intuition might be due to the sense that Y deserves to be harmed due to his wrongdoing. To correct for this, he proposes that the intuition that Y is permissibly used withstands Y’s being intentionally manipulated to act, which he and I agree would rule out Y’s deserving to be harmed:

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

Tadros correctly predicts that I don’t have the intuition that this use is permissible. But he does. He adds:

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

I think this is a good challenge. But still, my strong sense is that it's wrong to throw the man off the bridge. However, in my view, it's the right to life, liberty, and physical security of the person that has the key role in the use objection to general deterrence. Those rights are grounded in the more general right to a life in which one's capacity for flourishing is not compromised in the long term. Thus there is a heavily weighted presumption against punishment as manipulative use, where that involves intentional killing, confining, or infliction of severe physical or psychological harm. But consider monetary fines, when they don't preclude living a life at a reasonable level of flourishing. I've argued (2001: 177) that the general deterrence involving monetary fines may be in the clear. So suppose that Rich is guilty of insider trading in the stock market, and he knows that insider trading is illegal. Now add that, while satisfying all of the prominent compatibilist conditions on moral responsibility, he is causally determined by Diana, as in Mele's (2006) zygote argument, to act as he does. (Or, deterministically manipulated by neuroscientists as in my Case 2 (Pereboom 2014: 77); in the Case 1 scenario (Pereboom 2014: 76-7), we should prefer to prevent wrongdoing by targeting the neuroscientists rather than the manipulated agents.) Suppose that a $50,000 fine would deter him and others from this sort of financial crime. My sense that it would be illegitimate to fine him is weaker than my intuition that it would be illegitimate to imprison him for the sake of general deterrence.

Imagine we all know that everyone has been causally determined by Diana from the outset, on the model of Mele's (2006) zygote argument, to act as they do throughout their lives, but that all of the prominent compatibilist conditions on moral responsibility are satisfied. In particular, agents have the capacity to shape their behavior and dispositions to behave through appreciation of reasons. Serious efforts at training in business ethics have been made, but insider trading is still rife. Imagine also that technology that allows for prevention of insider trading through effective monitoring is unavailable, and that as a result a policy that invokes only special deterrence would be ineffective. Under those circumstances, a system of general deterrence with monetary fines as threats would appear justified. Other options for effectively preventing insider trading seem worse morally, and wrongdoers are being manipulatively used only with regard to property rights, in ways that don't preclude living a life at a reasonable degree of flourishing. Thus, while I resist Tadros argument for manipulative use for generally deterrence when it involves killing, I do not resist this sort of argument when the manipulative use involves monetary fines, within limits.

One might contend that if manipulative use involving monetary fines is within bounds, so are short prison sentences, say of several months. Tadros and Richard Arneson (in conversation) argue that the difference between the two is insignificant, while short prison sentences are often especially effective deterrents, in particular in combination with a high expectation of being caught (Kleiman 2009), so they should be treated roughly the same way. In support, one might contend that, while a short prison term is a violation of the liberty right, it is only a moderately serious violation, and does not preclude living a life with a reasonable level of flourishing in the way that long prison terms typically do. This provision would help with 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, especially given its effectiveness as a deterrent, and in combination with a high expectation of being caught.

Smilansky objects that the policy for treatment of criminals that I endorse will fall far short of the requirement to adequately deter crime. His main concern is that the detention of the dangerous on the quarantine model I advocate will yield insufficient and inadequate deterrence. On this model, those who are detained would need to be compensated for their confinement by what he calls "funishment," which in an earlier version of the criticism he specified as equivalent to a five-star hotel (Smilansky 2011; cf. Corrado 1996). Neil Levy (2012) and I (2014: 172-3) disagreed, and I argued that reasonable accommodations and programs for rehabilitation and reintegration would be in order. Smilansky replies that two-star accommodation would also not yield adequate deterrence, and that therefore a harsher environment, justified on retributive grounds, would be required instead. I now want to emphasize, first, that, in addition to detention justified by analogy to quarantine, additional sorts of monitoring, and programs for rehabilitation and reintegration, the model I advocate includes general deterrence by monetary penalties and short prison terms. This yields a response to Smilansky's example of the greedy relatives who kill the source of their inheritance. Their motive is financial gain, and it stands to reason that they would be deterred by a credible threat of dispossession. Such penalties can also serve as a deterrent for the spousal murderer who poses no other genuine threat, although credible examples of this phenomenon may be extremely rare.

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, their prisons are indeed two-star hotels. And their crime and recidivism rates are much lower than they are in the United States, whose criminal justice system is closer to what Smilansky envisions (see e.g., Ward et al. 2015, and its references). There well may be important differences between American and Norwegian societies that account for this state of affairs. Norway features markedly less inequality of wealth than does the US. But policies to reduce inequality may well be preferable to maintaining a harsh prison system. Guns are more plentiful in the US than they are in Norway. But gun ownership restriction may well be preferable to the harsh prison system. The reasons for differential success in deterrence and prevention between Norway and the US are undoubtedly complex, but this counsels against ready acceptance of the claim that harsher prison conditions of the sort that Smilansky advocates would generally be more effective and are to be preferred to alternative measures. At least, those who argue as he does should consider the evidence of the success of criminal justice systems such as Norway's, and provide reasons to think that such an approach doesn't generalize.

Smilansky also objects that on this account intuitively too many people will be drawn into the criminal justice system. First, he intimates that many more people would be detained than is the case currently. Second, there is the issue of incapacitating those who pose threats but have not yet committed crimes. Smilansky is reasonably concerned about the prospects of such a policy, and he believes that retributivism has the effect of limiting detention to an intuitively plausible degree, in particular, I'm assuming, with regard to the rationally competent. Michael Corrado (2016) has raised similar concerns, and Gregg Caruso and I (forthcoming) have made an effort to respond to them. Let me reiterate what we say. On the first issue, the principle of least infringement, which derives from the grounding of incapacitation in the right to self-defense and defense of others, specifies that the least restrictive measures should be taken to protect public safety, and is highly relevant. While we should indefinitely detain mass murderers who cannot be rehabilitated and remain threats, non-violent shoplifters who remain threats and cannot be rehabilitated should not be preventatively detained, by contrast with being monitored, for example. The view does not prescribe that all dangerous people be detained until they are no longer dangerous. Moreover, other behavior that is now considered criminal would not require incapacitation. Our view is consistent, for example, with the decriminalization of non-violent behavior, such as recreational drug use, and thus is consistent with many fewer people being detained than in the US currently (Pereboom and Caruso, forthcoming).

On Smilansky's second issue, the incapacitation of the dangerous who haven't committed crimes, there are several moral reasons that count against such a policy. As Ferdinand Schoeman (1979) has also argued, the right to liberty must carry weight in this context, as should the concern for using people merely as means. In addition, the risk posed by a state policy that allows for preventative detention of non-offenders needs to be taken into serious consideration. In a broad range of societies, allowing the state this option stands to result in much more harm than good, because misuse would be likely. Schoeman also points out that, while the kinds of testing required to determine whether someone is a carrier of a communicable disease may often not be unacceptably invasive, the type of screening necessary for determining whether someone has violent criminal tendencies might well be invasive in respects that raise serious moral issues. Moreover, available psychiatric methods for discerning whether an agent is likely to be a violent criminal are not especially reliable, and as Stephen Morse points out, detaining someone on the basis of a screening method that frequently yields false positives is seriously morally objectionable (Morse 1999; Nadelhoffer et al. 2012).

But in (2014) I present an example in which an agent has been fed a drug without his knowledge that makes it predictable that he will cause a criminally prohibited harm within a week. After the week the effect of the drug wears off. I suggested that the state is entitled to detain him for that week. Corrado (in correspondence) argues that, if the drug renders him not reasons-responsive and thus not morally responsible, preventative detention is permissible. Smilansky might be attracted to this kind of position: if agents are dangerous but not reasons-responsive (cf. Fischer and Ravizza 1998), they may be preventatively detained. But what if someone is dangerous and reasons-responsive? Corrado's (1996) view is that such an agent may not be detained unless it can be shown that he has a current intention to cause harm. As Corrado indicates, specifying the particular features of intention that would be sufficient for preventative detention is a delicate and difficult issue (e.g., how specific must the intention be?). But this general sort of position seems reasonable to me.

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

However, in other jurisdictions, such as Ontario, Canada, the state may detain the dangerous even when rationally competent, albeit under mental health legislation (thanks to Jennifer Chandler for this information). It seems to me that the Ontario policy is preferable (arguably supplemented with the demonstrable intention requirement). Tarasoff should not have been subjected to the burden of protecting herself against someone with a demonstrated intention to kill her. Here I agree with Corrado that a demonstrable intention standard is legally feasible, and that the agent's being reasons-responsive should not all by itself rule out preventative detention. Would Smilansky agree? If not, he would allow the woman to be subjected to the burden of self-protection. If so, then retributivism alone won't have the detention-limiting role he advocates for it.

Lastly, Smilansky contends that retributivism is practically valuable, and indeed indispensable, despite the fact that his own view on free will undercuts its metaphysical basis, since, if we lack free will, we don't have the kind of control in action that retributivism demands. On the classical retributivist view, the good to be achieved by punishment, by means of which retributivism justifies punishment, is that an agent receive what he deserves just because of his having knowingly done wrong (Kant 1797/1963; Moore 1987, 1998; Morse 1999, 2004). This position would be undermined if the free will skeptic is right, since, if agents do not deserve to be blamed just because they have knowingly done wrong, neither do they deserve to be punished just because they have knowingly done wrong. Retributivism justifies punishment solely on the ground of basic desert, and because the skeptical position claims this notion does not apply to us, it is incompatible with retributivism.

Smilansky contends that, despite the sound arguments against our having the requisite kind of free will, we should, for practical reasons, accept that this kind of free will is compatible with causal determination (Morse (2004) defends a similar view). However, if the retributivist justification of punishment featured by our actual practice requires the rationality of the belief that compatibilism is true, while at the same time there are serious and unanswered objections to this position, we cannot legitimately respond to a challenge to this part of the practice just by saying that it is supported by compatibilism. Punishment inflicts harm and, in general, justification for harm must meet a high epistemic standard. If it is significantly probable that one’s justification for the harming another is unsound, then, *prima facie* that behavior is seriously wrong and one must refrain from engaging in it. A strong and credible response to the objections to compatibilism is required to meet this standard.

Moreover, there are substantial arguments for the claim that retributivism turns out to be unacceptable, even disregarding the free will skeptic’s considerations. First, retributivist sentiments may be grounded in vengeful desires, and therefore retribution has little more plausibility than vengeance as a morally sound policy for action. Acting on vengeful desires might be wrong for the following sort of reason. Although acting on such desires can bring about pleasure or satisfaction, no more of a moral case can be made for acting on them than can be made for acting on sadistic desires, for example. Acting on sadistic desires can bring about pleasure, but in both cases acting on the desire aims at the harm of the one to whom the action is directed, and in neither case does acting on the desire essentially aim at any good other than the pleasure of its satisfaction. But, then, since retributivist motivations are disguised vengeful desires, acting for the sake of retribution is also morally wrong.

Second, there are no strong positive reasons for maintaining that basic desert is a moral good, or, on a view in which the right is prior to the good, that it reflects a defensible principle of right. When Immanuel Kant introduces his retributivist justification of punishment in the *Metaphysical Elements of Justice* (1797/1963), no apparent link is forged with the comprehensive moral theory expressed in the various formulations of the categorical imperative (Kant 1785/1981). Does treating humanity in persons as an end in itself require that we regard persons as susceptible to punishment retributivistically justified? Will legislation for a kingdom of ends by a community of rational beings invoke retributivistic justification? Add to this the fact that purely consequentialist or contractualist views will not accommodate this sort of justification, and at best can only secure a weaker facsimile.

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

Are there viable ways to confront wrongdoing without invoking desert? I have argued in the affirmative. Instead of invoking desert, blame can be largely forward-looking, recast as appropriate moral protest, and aiming at protection, moral formation, and reconciliation. Treatment of criminals can be similarly forward-looking, justified on grounds of protection and moral formation, and featuring, as policy, incapacitation and rehabilitation. Revision of moral practice poses risks, but the elimination of negative desert promises, in these respects, a more compassionate and humane result. Here there is no practical need to resist the desert-denying implication of the skeptical view and to embrace Smilansky's (2000) illusionism instead.

McKenna on Manipulation and Basic Desert

On the position I propose, blame is to be justified in the largely forward-looking way, in view of the aims of protection, reconciliation, and moral formation. The immediate target of blame is a past action, and in this respect such blaming will have a backward-looking aspect: the badness of the past act is part of what makes blame appropriate. As McKenna points out, I agree that wrongdoers are often harmed upon being blamed, but I contend that this harm appropriately has only instrumental value. On his view, by contrast, harm that is suffered by a wrongdoer upon being blamed can be a non-instrumental good. As he recounts, I've objected that McKenna's defense of his position appears to indicate that for him the harm of blame is only instrumentally good, for example, good in virtue of promotion of membership in the moral community. He stands firm, however, in defending the following axiological desert thesis:

> AD: It is a better world that one who is guilty of wrongfully harming another be made worse off for having done so, as in comparison with a world in which only the victim of the wrong is made worse off.

McKenna supports this claim with the analogy of grief. Grief upon the death of one's mother is an expression of one's love for her; "It counts as a kind of harm that involves the overcoming of something bad, and so as an expression of positive value. Hence, it is something that is good." Dana Nelkin responds (2013) by arguing that the pain or harm of grief is not best counted as good. Rather, the pain of grief is a bad component of a state that is good and morally appropriate overall. So one might argue that, if grief could be experienced without pain, which may be psychologically impossible for us, it would be better all things considered. On this view, the pain of grief is bad in itself, but is nonetheless instrumentally valuable insofar as it, given our psychology, is required for grief, which is a good and morally appropriate state overall.

One might similarly argue that the pain of regret is not good in itself, but yet a component of the state of contrition, which is good overall. If we were capable of recognizing that what we did was wrong and engaging in reconciliation, compensation, and moral reform without the pain, the overall state would be better. Analogously, with respect to blame, suppose that it's psychologically impossible for a particular wrongdoer to undergo moral reform without feeling the pain that results from confronting him with reasons not to act as he has. The pain of regret induced in this way would be morally appropriate, but again due to its instrumental value.

McKenna raises a number of concerns about my four-case manipulation argument. The core idea of a manipulation argument against compatibilism is that an action's being produced by a deterministic process that traces back to factors beyond the agent's control, even when she satisfies all the conditions on moral responsibility specified by the contending compatibilist theories, presents no less of a challenge to basic-desert responsibility than deterministic manipulation by other agents. An argument of this kind acquires more force by virtue of setting out several such cases. In my version (Pereboom 1995, 2001, 2014), the first features the most radical sort of manipulation consistent with the proposed compatibilist conditions and with intuitive conditions on agency. The cases culminate in a fourth, which the compatibilist might envision to be ordinary and realistic, in which the action is causally determined in a natural way. An additional challenge for the compatibilist is to point out a principled difference between any two adjacent cases that would show why the agent might be morally responsible in the later example but not in the earlier one. I argue that this can't be done, and that the agent's non-responsibility thus generalizes from the first of the manipulation examples to the ordinary case.

In the set-up, in each of the four cases, Plum decides to kill White for the sake of some personal advantage, and succeeds in doing so. Two important features of the set-up are these:

1. The first case poses a direct challenge to the sufficiency of the various prominent compatibilist conditions, since in that case it is clear that Plum is not morally responsible, and yet he satisfies the compatibilist conditions.
2. It is impossible to draw a principled line between any two adjacent cases that would explain why Plum would not be responsible (and we're assuming the basic-desert sense) in the first, but would be in the second, largely because all of the prominent compatibilist conditions are satisfied in each – this is the "no difference" part of the argument.

A key concern was raised by a number of critics (Mele 2005; Baker 2006; Fischer 2004; Demetriou 2010)) about my former first Case 1 (1995, 2001). In that story, Plum was created by a team of neuroscientists, who can manipulate him directly through radio-like technology, but he is as much like an ordinary human being as is possible, given his history. The scientists locally manipulate him from moment to moment, by pushing buttons on their device. In this particular situation, by doing so, they cause him to undertake an egoistic process of reasoning that deterministically leads to his killing of White (2001: 112-3). The concern raised by the critics is that in this first case Plum's mental states lack the causal cohesion required for him to be a genuine agent. The specific worry raised by Kristin Demetriou is that, when the scientists manipulate Plum at some particular instant, they suppress the causal efficacy of his prior mental states, thereby making it the case that Plum's mental states are causally isolated from one another (Demetriou 2010: 608; Matheson 2016: 1972). So while Plum in Case 1 and Plum in

Case 4 have the same mental states, these states differ radically with regard to how they are causally connected. And this difference yields a difference in agency and, potentially, in moral responsibility.

In response, in (2014) I revised the case so that the neuroscientists manipulate him only sporadically. In this particular case, they (the team of neuroscientists) do so on a single occasion by pressing a particular button just before he begins to reason about his situation, which they know will produce in him a neural state that realizes a strongly egoistic reasoning process, which the neuroscientist know will deterministically result in his decision to kill White (as suggested by Seth Shabo 2010: 376). I argue that external influences, such as finding out that the home team lost, have a similar sort of effect, while there is no question that agency is preserved. But, in this case, while agency is preserved, Plum is plausibly not morally responsible since (I specify that) had the neuroscientists not intervened, he would not have killed White (2014: 76).

McKenna criticizes this new case on the ground that it is much easier to accept that Plum is morally responsible here than it was in my former Case 1. This stands to reason, but my thought was that the new case was strong enough. Imagine that Plum is on trial by jury, and that it was revealed that the neuroscientists intervened as they did, and that he would not have killed White had it not been for their intervention. What jury would convict him? But McKenna points out that in his (2008) article he proposed another solution, which retains the local manipulation while ensuring the causal integration of Plum’s states. Recently, Benjamin Matheson (2016) developed a detailed account of how this might work, specifically with Demetriou’s critique in mind. For me, the key ingredient in his account is this: sometimes we are motivated to perform an action, but not sufficiently strongly so as to perform it without some badgering by someone else. I may want to reorganize the basement, but I won't do it unless my wife keeps me on track. Here my earlier mental states are not suppressed, but enhanced, and agency is clearly retained under such external influences. So similarly, the neuroscientists, rather than completely suppressing the causal efficacy of all of Plum's prior mental states, strengthen some of them and diminish the efficacy of others, and control all of his actions in this way. I'm happy with this resuscitation of my former Case 1, and I'm grateful to McKenna and Matheson for their valuable efforts.

McKenna subsequently presents a valuable recap of the discussion we've been having over the past decade about the dialectic of manipulation arguments. One aspect of my argument is Spinozistic: in ordinary cases we're ignorant of the causes of our actions and tend to assume we're free as a result. Manipulation cases serve to make salient hidden causes, and setting up manipulation cases that feature no responsibility-relevant differences from ordinary compatibilist-friendly deterministic cases is key to the force of the argument. McKenna grants this, but then injects a new element into the discussion: the real life manipulation cases. He contends that they are "friendlier to a compatibilist diagnosis." By contrast with artificial cases, "[o]ur intuitions about freedom and responsibility are more stable if applied in contexts with which we are familiar and in which they were learnt and have evolved."

In his response, McKenna refers us to Nomy Arpaly's (2006: 111-3) real life cases. Arpaly first presents four pairs of cases of changes in action-relevant beliefs and desires. In the first of each pair, the change is natural; in the second, it is induced by another human being. Here is one pair:

> After seven days trapped in a desert cave, thirsty and cold to the point of delirium, I had an epiphany and became a theist. A clever cult leader, wanting me to become a theist, made sure I was trapped in a desert cave. Having an unusual intuition for such things, he achieved his goal and I became a theist.

Arpaly's point is that we react differently to the agent-manipulation than to the natural manipulation cases – "we greatly fear being under the control of another human" – but this difference "has nothing to do with moral responsibility." Agreed, and this "no difference" principle is a key element in manipulation arguments. But Arpaly's examples involve only acquisition of beliefs and desires, for which moral responsibility is generally in doubt. Let's extend Arpaly's example to action. Suppose, in the non-agential manipulation case, I subsequently donate $1,000 to a cult, and that I would have donated the $1,000 to Bernie Sanders's campaign had I not spent the time in the cave. My intuition is that I might be responsible for this. But the set-up, so far, does not specify that the action is causally determined. Suppose it did: the time in the cave causally determined a brain change that in turn deterministically caused me to donate the money to the cult, despite the compatibilist conditions being met. But now my sense is that I'm not responsible, although McKenna has a contrary intuition. But it's more important to tracking the dialectic to point out that my Case 3 is a natural manipulation case: there Plum is manipulated in his upbringing to act as he does. In my set-up, this case is in third and not in first place in the order of presentation, and the reason for this is that I expect Case 1 to provide a stronger hook to my target audience. If McKenna's suggestion is that I put Case 3 in first place, I'd resist it, since it would make the argument weaker.

A final point about the dialectic of the manipulation argument. In (2014: 99-100), I made a case for the claim that manipulation arguments should not be conceived as coercive. Let me explain. The incompatibilist believes that the compatibilist is mistaken in her judgments of moral responsibility in deterministic scenarios. The incompatibilist's scheme is to go on a fishing expedition – to hook those susceptible to compatibilism with a manipulation case, and then to point out that there are no responsibility-relevant differences between the manipulation cases and an ordinary deterministic counterpart. But suppose that someone isn't hooked – that she doesn't have the non-responsibility intuition in the manipulation case. At this point, the incompatibilist may suspect theoretical intransigence – her compatibilist commitments are determining her responses. Here a request can be made for setting theoretical commitments aside. But our ability to do that is limited. Relatedly, either side might point to what a neutral enquirer would think (Haji and McKenna 2004). But our access to what a neutral enquirer would think is limited by our theoretical commitments. Beyond this, I believe rational coercion is impossible. If the compatibilist lacks the no-responsibility intuition in the manipulation case, even when attempting to set her theoretical commitments aside, the incompatibilist has no justification for claiming that she is irrational. Another way to discuss these disagreements would need to be found.

Mele on the Disappearing Agent Argument

Al Mele focuses on the argument against event-causal libertarianism that I highlight and develop in Chapter 2. Here it is:

> *The disappearing agent objection*: Consider a decision that occurs in a context in which the agent's moral motivations favor that decision, and her prudential motivations favor her refraining from making it, and the strengths of these motivations are in equipoise. On an event-causal libertarian picture, the relevant causal conditions antecedent to the decision, i.e., the occurrence of certain agent-involving events, do not settle whether the decision will occur, but only render the occurrence of the decision about 50% probable. In fact, because no occurrence of antecedent events settles whether the decision will occur, and only antecedent events are causally relevant, *nothing* settles whether the decision will occur. Thus it can't be that the agent or anything about the agent settles whether the decision will occur, and she therefore will lack the control required for basic-desert moral responsibility for it.[1]

Mele develops a series of criticisms of this argument. First, he states that he does not have a grip on or understand what it means to settle whether a decision occurs. In reply, in a quotation of his own work in his contribution, he himself provides material that helps us understand this notion. In *Springs of Action* (1992: 158-9), he writes: "in deciding to A, one settles upon A-ing (or upon trying to A), and one enters a state—a decision state—of being settled upon A-ing (or upon trying to A)". In this formulation, Mele uses the notion of settling with respect to options for action, by contrast with options for decision. Mele does not provide a characterization of "settling upon A" in other, distinct terms, which I think is fine, given that it's common in ordinary language. But let me suggest a synonymous formulation for the case of action. To

[1] I say that this argument targets basic-desert moral responsibility rather than agency [2014: 32]. The reason is that this argument doesn't show that the occurrence of S's decision to A absent S's settling whether the decision to A occurs cannot count as an instance of S's agency.

settle which action to perform is to *determine*, not necessarily causally, *which of one's options for action to perform*. And now, as Mele's exposition specifies, one settles (or at least one can settle) whether *the decision* to A (conceived of as a decision state) occurs by settling or determining whether to A or not-A. This is at least typically the case: one settles whether a decision to A occurs by settling upon A.

As Mele points out, in some extraordinary possible cases (e.g., Kavka 1983), one's focus in deliberating is on whether to decide to A by contrast with whether to A. If the evangelist says: "The time to decide has come – will you decide to commit your life to Christ?" and you believe that your eternal fate depends on your decision, specifically, you may focus on whether to decide to commit your life to Christ rather than on whether to commit your life to Christ. But Mele is right to think that this would be an unusual type of case. As he suggests, settling as pertaining to decisions is a kind of direct control insofar as it contrasts with control exercised through information gathering and reasoning processes, as long as it's understood that, as just discussed, this sort of settling typically targets action and not decision directly.

There is arguably more to the notion of settling than action-determination. I agree with Helen Steward (2012) that settling also involves difference-making – the agent making the difference as to which option for action occurs. Putting the two together, I propose the following biconditional characterization:

> An agent settles which of one's options for action occurs just in case she determines, not necessarily causally, which action occurs, and she makes the difference as to which action occurs.

One might ask: is such difference-making compatible with causal determination? Steward thinks not. But a case can be made against her claim. Carolina Sartorio (2013) argues that causes of action in Frankfurt cases, despite the agent being unable to do otherwise, still make a difference to their effects in that the effects wouldn't have been caused by the absence of their causes.[2] Following this lead, we can propose the following difference-making-capturing sufficient condition for settling in the context of a Frankfurt example:

> An agent settles whether an action occurs if it is caused by certain reasons of hers, where the absence of those reasons would not have caused that action

Note that this characterization is compatible with the agent being causally determined to act as she does. Adapted to an agent-causal context, we get the following sufficient condition:

> An agent settles whether an action occurs if she agent-causes it for certain reasons, where the absence of her agent-causing the action for those reasons would not have caused that action,

which can also be satisfied if the agent is causally determined to act.[3]

Mele provides a valuable discussion of an important related issue: whether the relevant sort of settling or control in deciding should be interpreted as involving *complete* control. Let me note first that I never say that it does, and I don't in fact think it does. But I did not discuss this issue in (2014), and so Mele's question about whether this is what I did mean is on the table. I think that his examples – in which an agent's control in deciding allows for a smallish

[2] If one is averse to absence causation, one might invoke Phil Dowe's (2001) notion of "quasi-causation" in its place, where (roughly) the absence of watering the plant quasi-causes it to die just in case watering it would have caused it to continue to live. Sartorio (2016) suggests this alternative.

[3] This treatment of difference-making corrects for problems for the earlier account I provided in (2015). Thanks to Richard Holton and Carolina Sartorio for valuable discussion.

probability that whatever an agent does to settle whether a decision occurs does not issue in the decision's actually occurring – are convincing. That is, I think that a smallish probability of such a failure is consistent with settling, and with the control in deciding that's at issue in this debate – in my view, the control required for the agent's basic-desert moral responsibility for the decision.

But suppose I also maintained that a 50% probability of such a failure renders the agent's control insufficient. Do we now have in place a barrier to whether the notion of settling in play can be understood? An analogous concern has familiarly been raised for the notion of knowledge. Just as one might propose that settling involves complete control, one might propose that knowledge involves certainty. But at least in some contexts there being some smallish probability that one's belief is mistaken seems consistent with knowledge attribution, just as a smallish probability that what one does in settling fails to issue in the relevant decision is consistent with the attribution of settling. But a 50% probability that one's belief is mistaken is always inconsistent with knowledge attribution. Yet these findings would not seem to warrant a claim as strong as "we don't understand what is meant by 'knowing.'" So the analogous findings would not warrant a claim as strong as "we don't understand what is meant by 'settling' as it pertains to decisions." Or perhaps the concern is focused at a higher level: until we understand precisely the range of application of 'settling.' as it pertains to decisions, we don't understand what 'settling' means here. But this also appears mistaken. It's implausible that unless we understand precisely the range of application of 'knowing' we don't understand what 'knowing' means.

Some philosophers have argued that the concerns raised about ‘knowing’ show that the notion of knowing isn’t precise enough for use in epistemology, and they recommend that we instead use the notion of credence, that is, precise degree of belief, exclusively. But this is a minority position. Most others are happy to continue to use ‘knowing’ in epistemological discussions, despite the lack of clarity concerning degree of belief this notion features. By analogy, we could recommend dispensing with ‘settling’ in the philosophical discussion and employ instead determination-of-decision-to-a-precise-degree. However, in epistemology we have effective methods for discovering credences, for example, by betting preferences. But we have no analogous methods for adjudicating precise degrees of determination of decision. So it looks like we’re stuck with ‘settling,’ or analogous notions such as ‘decision-and action-determination’ that are similarly imprecise. But, again, we shouldn’t conclude from such imprecision that we don’t know what’s meant by ‘settling’ as it pertains to decisions. Even if ‘knowing’ is too imprecise to use in epistemology, it’s not that we don’t know what ‘knowing’ means.

On my construal of event-causal libertarianism, control is a causal matter; since settling which decision occurs would be the key control relation for securing moral responsibility in paradigmatic cases, such settling must be a causal matter. In addition, the event-causal libertarian claims that all causation is causation by events. My claim is that, given these commitment of event-causal libertarianism, the theory lacks the resources to explain how the agent can settle which decision occurs in the equipoise situation. I contend that we understand what it means to make such a claim about settling, and that Mele himself provides us with an account that provides us with this understanding. It may be that in certain cases in which what the agent does to settle which decision occurs is compatible with a smallish probability that it does not occur, the agent nevertheless settles that the decision occurs as required by moral responsibility for it in the basic-desert sense. I claim that the relevant antecedently occurring agent-involving events don't settle which decision occurs in cases in which the probabilities are not in equipoise but are significant on both sides, where the relevant boundary between smallish and significant is vague.

At this point, the event-causal libertarian might contend that causation by appropriate agent-involving events in the equipoise case is sufficient for settling which decision occurs. After all, the agent is involved in each of the two competing sets of antecedent events, so, no matter which decision results, she herself will have settled which one occurs. Mele would agree that which decision occurs is a matter of luck, but would question the claim that the action is not free in the sense required for attribution of basic-desert responsibility. Just as for the hard line response against the manipulation argument, I doubt that there is a rationally coercive response to this line. Like the manipulation argument, the disappearing agent argument involves fishing for an intuition. In the manipulation argument, it's for the intuition of non-responsibility in a manipulation case; in the disappearing agent argument, it's for the intuition that the agent does not settle which decision occurs in the equipoise case, supposing that settling is a causal matter, and thus the agent is not morally responsible in the basic-desert sense. If the opponent does not have this intuition, then the argument fails as a way to engage her, and another medium for negotiating the dispute would have to be found. Mele may be in this camp, and, if so, I have no rationally coercive way of showing him that he's mistaken. Still, others have the non-settling intuition, as I do, and this is required for the argument to engage them.

Suppose that you have this non-settling intuition in the equipoise case. The natural remedy is to look for some other way that the agent can settle which decision occurs, and here agent causation counts as an alternative. To my mind, the agent-involvement independently of events counts as an advantage, as does the idea of the agent as cause, given the plausibility of the claim that control in action is a causal matter (2014: 55), and that settling involves control. At the same time, as I explain in (2014), the agent-causal view is not in the clear, and is subject to a number of concerns.

Mele intimates that it is mistaken to contend that the event-causal libertarian can allow only events *antecedent* to a decision to settle whether a decision occurs. He objects to my formulation of the concern insofar as it employs the future tense 'will occur':

> On an event-causal libertarian picture, the relevant causal conditions antecedent to the decision, i.e., the occurrence of certain agent-involving events, do not settle whether the decision will occur (2014: 32).

This formulation suggests that settling would result from events that occur at least in part antecedently to the decision, and not by events that occur only simultaneously with the decision, which Mele thinks should be permitted for the event-causal libertarian.

As I said, in my discussion of event-causal libertarians, I'm assuming that they are committed to the claim that control in action is a causal matter. It's the non-causalists who hold that control in action is non-causal. If the locus of moral responsibility is an agent's decision, and if control is a causal matter, then it would seem that control in settling which decision occurs would be a function of how the decision is caused. On event-causal libertarianism, only events can be causes, so control in settling which decision occurs would, then, be a matter of the causing of the decision by events, and some of these would be agent-involving events. The agent-involving events would at least typically be belief-events and desire-events. And, in the usual case, if a decision is caused by beliefs and desires, the beliefs and the desires would exist for at least a short time prior to the decision. Thus it would make sense to describe such a case as one in which the occurrence of such events do or do not settle whether the decision *will occur*. At the same time, I have no convincing argument against the claim that it's possible for a decision to be caused by belief-events and desire-events that occur precisely at the time of the decision and not at earlier times. Such simultaneous event causation is controversial but sometimes endorsed. But I'm happy to accommodate Mele's recommendation and eliminate the apparent futurity requirement.[4]

Nevertheless, a word of caution. I argue in Chapter 2 of (2014) that alleged event-causal libertarians have a tendency to slip into non-causalism when it comes to an account of settling which decision occurs or the equivalent, and non-causalism is suggested by at least some simultaneous-determination views of settling, such as Mark Balaguer's (2010). When it comes to specifying what settles which of two possible decisions in motivational equipoise is made, Balaguer (2010: 93) specifies that it's the agent; "it's *Ralph* who does the choosing," and Ralph's choosing would indeed be simultaneous with the relevant decision state. But I argue (2014: 36-

[4] Note that, on one variety of the agent-causal view, decisions are agent *causings*, which in turn are analyzed as activations of the agent-causal power (e.g., DeRose 1993). On such a position, a decision to A will be simultaneous with an agent causing of A by virtue of the decision's being identical with the agent causing.

9)) that the claim that Ralph does the choosing can't be cast in terms of event-causal relations without reintroducing the disappearing agent objection.

More recently, Balaguer (2014) proposes a response to the disappearing agent argument that features a combination of event-causal and non-causal relations. Resisting the agent-causal resolution, he says: "in this scenario, the event that settles which option is chosen is the conscious decision – i.e., the event with a me-consciously-choosing-now-phenomenology" (2014: 83). First, suppose that the event *S's deciding to A at t* is caused by S-involving events, say (desire) E1 and (belief) E2. Supposing equipoise – or in Balaguer's terms, the decision's "being torn" – there will be other events, E3 and E4, poised to cause the alternative action-event, *S's deciding to not-A at t*. However, given the event-causal libertarian picture, now there seems to be nothing left to settle whether *S's deciding to A at t* by contrast with *S's deciding to not-A at t* occurs, at least on the supposition that the settling role is causal and only events can play it. Only E1-E2 and E3-E4 are candidates for this role, but they don't settle which decision-event occurs.

But Balaguer clearly intends the non-causal part that S plays in the event *S's deciding to A at t* to have the settling role. That's the point of his specifying that this event has a me-consciously-choosing-now-phenomenology. S's making the decision to A features a non-causal being-the-subject-of relation between S and A with this phenomenology. My arguments against non-causal options for settling are distinct from the disappearing agent argument, and so the concern that the disappearing agent argument has no leverage against what is in fact a non-causal view won't count against it. My main objection to non-causalism is that its proponents, such as Carl Ginet (1990, 1997, 2007), claim that in the case of a basic action the agent makes the action happen and makes the difference as to when it will happen, and making-happen and this sort of difference-making would seem to be causal relations. The worry is that, despite the claim of non-causalism, the settling relation is causal after all (Pereboom 2014: 39-47). A genuinely resolute non-causal view, on the other hand, faces the problem of accounting for the control in action required for responsibility in a non-causal way.

Mele argues, plausibly, that on event-causal libertarian view it remains correct to say that *agents*, and not events, decide and make decisions. “[Sam] has and exercises the ability or power to sink free throws, and he sinks many of them. His intentions, beliefs, skills, and the like do not sink free throws – alone or in combination with one another. And that is no surprise, because they are not able to sink free throws.” But this yields no leverage in the disputes at issue. On the assumption that event-causal libertarianism is committed to a broadly causal theory of action, and to the claim that all causes of actions are events, this position is committed to the thesis that all causal influences on action are most fundamentally event causal, and thus explicable in exclusively event-causal terms. So what is Mele claiming when he says that Sam’s “intentions, beliefs, skills, and the like do not sink free throws”? He can’t mean that in the sinking of a free throw *Sam* is a causal influence distinct from Sam-involving events and states. It of course sounds odd to say that a collection of events or states sinks a free throw, and we don’t talk this way. But the event-causal libertarian must affirm that what grounds the truth of the claim “Sam sank the free throw” is that a collection of events, a number of them Sam-involving, caused a further event, the ball’s going through the hoop.

What's at issue here isn't how we're used to speaking, but the metaphysical commitments of an event-causal theory of action.[5]

I maintain that the disappearing agent argument successfully targets the semantic position, according to which claims such as "Sam sank the free throw" are true, but the causal relations that make them true are solely event-causal relations. I also believe that it is successful against the metaphysical position that all substance-causal relations reduce to event-causal relations. But let me note that I do not claim that the argument successfully targets a position according to which agents as substances are causes, and such substance causation does not reduce to causation by events, but is instead (non-reductively) constituted by or grounded in causation solely by events at more basic levels. (I tentatively affirm a deterministic version of an agent-causal picture of this non-reductive sort (2015)). Such a view endorses the metaphysical thesis that there is unreduced causation by agents fundamentally as substances, and thus denies that all causation is event-causation, a claim essential to the views the disappearing agent argument is designed to undercut.[6]

[5] Near the end of his comment, Mele outlines a position according to which an agent's moral responsibility on the event-causal view can result from the shaping of her values by past free decisions. I develop an objection to this type of view in (2001: 49).

[6] Thus by "causation fundamentally as a substance" I do not mean to exclude substance causation wholly grounded in or constituted by event causation, as long as it does not reduce to event-causation. For my views on constitution, grounding, and reduction, see Chapter 7 of (2011). Thanks to Randy Clarke for prodding me to clarify my position on this issue.

Thanks to Gregg Caruso, Jennifer Chandler, Michael Corrado, Richard Holton, John Lemos, Dana Nelkin, and Carolina Sartorio for valuable comments and discussion.

References

Arpaly, N. 2006. *Meaning, Merit, and Human Bondage.* Princeton: Princeton University Press.

Baker, L.R. 2006. "Moral Responsibility without Libertarianism." *Noûs* 40(2): 307-30.

Balaguer, M. 2010. *Free Will as an Open Scientific Problem.* Cambridge, MA: MIT Press.

Balaguer, M. 2014. "Replies to McKenna, Pereboom, and Kane." *Philosophical Studies* 169(1): 71-92.

Bentham, J. 1823/1948. *An Introduction to the Principles of Morals and Legislation*. New York: Macmillan.

Caruso, G.D. 2012. *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lanham, MD: Lexington Books.

Caruso, G.D. 2016. "Free Will Skepticism and Criminal Behavior: A Public Health-Quarantine Model." *Southwest Philosophy Review* 32(1): 25-48.

Corrado, M.L. 1996. "Punishment and the Wild Beast of Prey: The Problem of Preventative Detention." *Journal of Criminal Law and Criminology* 86(3): 778-814.

Corrado, M.L. 2016. "Two Models of Criminal Justice," ms. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757078.

Demetriou (Mickelson), K. 2010. "The Soft-Line Solution to Pereboom's Four-Case Argument." *Australasian Journal of Philoso*phy 88(4): 595-617.

DeRose, K. 1993. "Review of William Rowe's *Thomas Reid on Freedom and Morality*." *Philosophy and Phenomenological Research* 53(4): 945-9.

Dowe, P. 2001. "A Counterfactual Theory of Prevention and 'Causation' by Omission." *Australasian Journal of Philosophy* 79(2): 216-26.

Farrell, D.M. 1985. "The Justification of General Deterrence." *Philosophical Review* 94(3): 367-94.

Feinberg, J. 1970. *Doing and Deserving.* Princeton, NJ: Princeton University Press.

Fischer, J.M. 2004. "Responsibility and Manipulation." *Journal of Ethics* 8(2): 145-77.

Fischer, J.M., and M. Ravizza. 1998. *Responsibility and Control.* Cambridge: Cambridge University Press.

Ginet, C. 1990. *On Action*. Cambridge: Cambridge University Press.

Ginet, C. 1997. "Freedom, Responsibility, and Agency." *Journal of Ethics* 1(1): 85-98.

Ginet, C. 2007. "An Action Can Be Both Uncaused and Up to the Agent." In C. Lumer and S. Nannini, eds. *Intentionality, Deliberation, and Autonomy.* Farnhnam, UK: Ashgate, pp. 243-56.

Haji, I., and M. McKenna. 2004. "Dialectical Delicacies in the Debate about Freedom and Alternative Possibilities." *Journal of Philosophy* 101(6): 299-314.

Kant, I. 1785/1981. *Grounding for the Metaphysics of Morals*. J. Ellington, tr. Indianapolis: Hackett.

Kant, I. 1797/1963. *The Metaphysical Elements of Justice*. J. Ladd, tr. New York: Bobbs-Merrill.

Kavka, G. 1983. "The Toxin Puzzle." *Analysis* 43(1): 33-6.

Kelly, E. 2009. "Criminal Justice without Retribution." *Journal of Philosophy* 106(8): 440-62.

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

Levy, N. 2012. "Skepticism and Sanction: The Benefits of Rejecting Moral Responsibility." *Law and Philosophy* 31(5): 477-93.

Matheson, B. 2016. "In Defence of the Four-Case Argument." *Philosophical Studies* 173(7): 1963-1982.

McKenna, M. 2008. "A Hard-Line Reply to Pereboom's Four-Case Argument." *Philosophy and Phenomenological Research* 77(1): 142-59.

McKenna, M. 2016. "Manipulation Arguments, Basic Desert, and Moral Responsibility: Assessing Derk Pereboom's *Free Will, Agency, and Meaning in Life*." *Criminal Law and Philosophy*. DOI: 10.1007/s11572-015-9388-8.

Mele, A. 1992. *Springs of Action*. New York: Oxford University Press.

Mele, A. 2005. "A Critique of Pereboom's 'Four-Case Argument' for Incompatibilism." *Analysis* 65(1): 75-80.

Mele, A. 2006. *Free Will and Luck*. New York: Oxford University Press.

Mele, A. 2015. "On Pereboom's Disappearing Agent Argument." *Criminal Law and Philosophy*. DOI: 10.1007/s11572-015-9374-1.

Moore, M. 1987. "The Moral Worth of Retribution." In F. Schoeman, ed. *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press, pp. 179-219; reprinted in J.G. Murphy, ed. *Punishment and Rehabilitation*, third edition. Belmont, CA: Wadsworth Publishing Company, 1995, pp. 94-130.

Moore, M. 1998. *Placing Blame*. Oxford: Oxford University Press.

Morse, S. 1999. "Neither Desert nor Disease." *Legal Theory* 5(3): 265-309.

Morse, S. 2004. "Reasons, Results, and Criminal Responsibility." *University of Illinois Law Review* 2004(2): 363-444.

Nadelhoffer, T., S. Bibas, S. Grafton, K.A. Kiehl, A. Mansfield, W. Sinnott-Armstrong, and M. Gazzaniga. 2012. "Neuroprediction, Violence, and the Law: Setting the Stage." *Neuroethics* 5(1): 67-99.

Nelkin, D. 2014. "Responsibility, Conversation, and Desert: Comments on Michael McKenna's *Conversation and Responsibility*." *Philosophical Studies* 171(1): 63-72.

People v. Poddar, 518 P. 2d 342 – Cal. Supreme Court 1974.

Pereboom, D. 1995. "Determinism *al Dente*." *Noûs* 29(1): 21-45.

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

Pereboom, D. 2011. *Consciousness and the Prospects of Physicalism.* New York: Oxford University Press.

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

Pereboom, D. 2015. "The Phenomenology of Agency and Deterministic Agent Causation." In H. Pedersen and M. Altman, eds. *Horizons of Authenticity in Phenomenology, Existentialism, and Moral Psychology: Essays in Honor of Charles Guignon*. New York: Springer, pp. 277-94.

Pereboom, D., and G.D. Caruso. Forthcoming. "Hard-Incompatibilist Existentialism: Neuroscience, Punishment, and Meaning in Life." In G.D. Caruso and O. Flanagan, eds. *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*. New York: Oxford University Press.

Quinn, W. 1985. "The Right to Threaten and the Right to Punish." *Philosophy & Public Affairs* 14(4): 327-73.

Sartorio, C. 2013. “Making a Difference in a Deterministic World.” *Philosophical Review* 122(2): 189-214.

Sartorio, C. 2016. *Causation and Free Will*. Oxford: Oxford University Press.

Scanlon, T.M. 2013. “Giving Desert its Due.” *Philosophical Explorations* 16(2): 101-16.

Schoeman, F. 1979. “On Incapacitating the Dangerous.” *American Philosophical Quarterly* 16(1): 27-35.

Shabo, S. 2010. “Uncompromising Source Incompatibilism.” *Philosophy and Phenomenological Research* 80(2): 349-83.

Smilansky, S. 2000. *Free Will and Illusion*. Oxford: Oxford University Press.

Smilansky, S. 2011. “Hard Determinism and Punishment: A Practical *Reductio*.” *Law and Philosophy* 30(3): 353-67.

Smilanksy, S. 2016. “Pereboom on Punishment: Funishment, Innocence, Motivation, and Other Difficulties.” *Criminal Law and Philosophy*. DOI: 10.1007/s11572-016-9396-3.

Steward, H. 2012. *A Metaphysics for Fr*eedom. Oxford: Oxford University Press.

Tadros, V. 2016. “Doing Without Desert.” *Criminal Law and Philosophy*. DOI: 10.1007/s11572-016-9398-1.

Tarasoff v. the Regents of the University of California, 17 Cal.3d 425, 131 Cal.Rptr. 14, 551 P.2d 334 (1976) – Cal. Supreme Court, 1976.

Ward, K., A.J. Longaker, J. Williams, A. Naylor, C.A. Rose, and C.G. Simpson. 2015. “Incarceration Within American and Nordic Prisons: Comparison of National and International Policies.” *Engage: The International Journal of Research and Practice on Student*

Engagement 1(1): 46-47. http://www.dropoutprevention.org/engage/incarceration-within-american-and-nordic-prisons/.