The Epistemic Argument Against Retributivism

Elizabeth Shaw

Introduction

In his book, *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice,* Gregg Caruso has provided an exceptionally well-argued, thorough and original defence of a non-retributive approach to criminal justice, based on free will scepticism. The book is both of great philosophical interest and practical significance. Within the scope of this commentary, I will focus on one aspect of his argument: the epistemic objection against retributivism. Caruso is among a group of theorists who have recently developed the epistemic objection – an argument that represents a paradigm shift in the free will literature.[1] It is often suggested that the free will debate, which has raged for millennia, has reached a "stalemate", with plausible arguments for each of the main opposing positions and no prospect of resolution. The novel contribution the epistemic argument is that it invites one to take a step back from this debate and to consider the practical ethical implications of the stalemate itself. Specifically, the epistemic argument focuses on the moral implications of the uncertainty about whether we have the kind of free will required by the retributive theory of punishment. The epistemic argument is an approach to moral uncertainty (uncertainty about the soundness of ethical theories), which implies that theories of punishment should be held to a high standard of credibility. In this article, I disagree with Caruso's suggestion that the epistemic argument is not applicable to the Public Health-Quarantine (PHQ) model. However, I argue that this does not undermine Caruso's overall theory, because it is plausible that his theory can satisfy the standard of credibility required by the epistemic argument.

The PHQ Model and the Standard of Proof

The epistemic argument is based on the plausible idea that it is wrong (intentionally) to inflict serious harm on people, unless there are very strong grounds for believing that harming them is justified.[2] In other words, there should be a presumption against (intentionally) inflicting serious harm, the burden of proof should lie with the person who is in favour of inflicting serious harm, and the argument for inflicting serious harm must be established to a high standard of credibility. Like several other theorists who have applied this epistemic argument to punishment, Caruso persuasively argues that the appropriate standard of credibility for justifications of punishment is "beyond reasonable doubt".

---

[1] E.g. Pereboom D, Living without Free Will (CUP, Cambridge 2001); Double R, 'The Moral Hardness of Libertarians' (2002) 5 (2) Philo 226; Vilhauer B, 'Free Will and Reasonable Doubt' (2009) 46 (2) American Philosophical Quarterly 131; E Shaw, Free Will Punishment and Criminal Responsibility (PhD thesis, University of Edinburgh 2014); M Corrado, 'Punishment and the Burden of Proof' (UNC Legal Studies Research Paper) Available at: https://ssrn.com/abstract=2997654 or http://dx.doi.org/10.2139/ssrn .2997654.

[2] I will argue below that, although the intention/foresight distinction may carry *some* moral weight, there are reasons for thinking that that the epistemic argument still applies to a significant extent to *foreseeable*, unintended harm.

However, Caruso's case would be stronger if he modified his view about the non-applicability of the epistemic argument to his own PHQ model. Caruso argues that, while proponents of retributive punishment must bear a heavy burden of proof, the PHQ model need not be held to the beyond reasonable doubt standard, because the PHQ model is nonpunitive and does not involve harming people intentionally. However, it would be better for Caruso to reframe his defence of the PHQ model by acknowledging that he must bear the same burden of proof as other penal theorists but arguing that he can discharge this burden. Firstly, I will argue that, although the PHQ model can legitimately be regarded as nonpunitive, it should be held to the same epistemic standard as penal theories, because the PHQ model resembles punishment in certain relevant respects (and might be called quasi-punitive). Secondly, I will argue that the PHQ model does involve inflicting harm intentionally. Thirdly, I will point out that Caruso cites unintentional harms when justifying applying a high epistemic standard of proof to other penal theories, so it would be inconsistent for him to argue that his theory should be exempt from that standard on the basis that his approach (supposedly) does not involve intentional harm. Fourthly, I will suggest some reasons for thinking that the PHQ model meets the required standard of credibility, or, at least, that it could form a major component of an approach that would satisfy that standard.

Differences and Similarities Between the PHQ Model and Punishment

When arguing that the public-health quarantine model should be held to a lower epistemic standard than penal theories such as retributivism, Caruso emphasises that his favoured model is nonpunitive and does not involve the intentional infliction of harm. He writes:

"…there is an important difference between the state intentionally harming individuals (which is an essential component of punishment) and the state causing unintentional harm (which would apply to any harms caused by a nonpunitive system of incapacitation). Arguably, the former must meet a higher epistemic bar…[The] justification for intentionally harming wrongdoers should carry a higher burden of proof than adopting a nonpunitive approach [such as the PHQ model] that may or may not have unintended negative consequences."[3]

I agree with Caruso that it is important, for the sake of conceptual clarity, to distinguish between punitive and non-punitive responses to criminal behaviour, and I do not contest his claim that the PHQ model can legitimately be classed as non-punitive. However, the public-health quarantine model clearly *resembles* punishment in terms of the seriousness of the harms imposed on offenders and the restriction on the offender's liberty. The PHQ model would involve measures that are currently used in penal systems, such as detention in a secure setting, close monitoring in the community, or offering offenders the choice between detention and biomedical treatments (such as therapies for drug addiction or sexual offending).[4] Like punishment, measures imposed under the PHQ model could have negative impacts on the offender's friendships, family life and career and such measures (albeit unintentionally) could be stigmatic. These similarities seem relevant to one of the key intuitions that motivates holding theories of punishment to a high standard of credibility – the intuition that seriously harming people requires strong justification. In order to defend holding the public-health quarantine model to a significantly lower standard of credibility than that which penal theories should meet it would have to be shown that the differences

[3]G Caruso, *Rejecting Retributivism: Free Will, Punishment and Criminal Justice* (CUP 2021) 125-127.

[4] M Corrado, 'Criminal Quarantine and the Burden of Proof' (2019) 47 Philosophia 1095.

between this model and punishment are *relevant* to the standard of credibility and that these differences are *weightier* than the relevant similarities. Otherwise, to exempt arguments for seriously harming people from the high standard of credibility, just because the harm is classed as "nonpunitive", would look like a merely semantic manoeuvre.

From the offender's point of view, if the harm involved is equally serious, it may not matter that much whether the authorities view this harm as nonpunitive (and regrettable), or as fulfilling the goals of punishment. Nevertheless, Caruso maintains that it is morally relevant whether the harm is inflicted intentionally or unintentionally. While I agree that this distinction carries some moral weight, it is unclear whether it carries *enough* moral weight to justify significantly lowering the standard of credibility for theories that involve serious unintentional harm. Key analogies that lend intuitive support to the epistemic argument do not support the idea that there is a very weighty difference between intended and foreseen harms in this context. Caruso draws a useful analogy between the epistemic argument and the precautionary principle in the context of climate change – this principle could justify measures to combat climate change even if we were not 100% certain that manmade climate change exists.[5] However, it would not be very persuasive for an organisation to say that it was exempt from this precautionary principle because its carbon emissions were not *intended* to cause climate change but were just a foreseen consequence of a strategy that was intended to maximise financial profits.

Does the PHQ Model Involve Intentional harm?

One might challenge Caruso's claim that the public health quarantine-model involves only unintentional harm. Caruso claims that, on his model, "just as we do not seek to intentionally harm or punish those contagious individuals we are forced to quarantine, we do not seek to intentionally harm or punish those dangerous individuals we are forced to incapacitate."[6]

One might reply to this claim that the relationship between what the authorities intend to do to someone who is quarantined/incapacitated and the harm which that involves is a *constitutive relationship* not a causal one. Where one state of affairs is causally downstream from another state of affairs, it can be meaningful to speak of intending the former, while foreseeing the latter as an unintended side-effect. However, where one state of affairs constitutes another state of affairs, the intention side-effect distinction cannot apply. Consider an analogy from the literature. A group of explorers are trapped in a cave – one of their number, who is particularly fat, is stuck in the entrance and cannot be moved.[7] The only way to escape is to use dynamite to blow him to bits. Can the explorers claim that they did not intend to kill him; they only intended to blow him to bits? It seems not, because the relationship between those two states of affairs ("killing him" and "blowing him to bits") is too close. "Blowing him to bits" constitutes "killing him".[8] Similarly, when the state intentionally imposes severe burdens on an offender, such detaining him for a very long period in a secure facility, this constitutes harming him. Thus, one cannot claim that one intends to impose these severe burdens on the offender without intending to harm him.

[5] Caruso (n3) 117.
[6] Ibid 126.
[7] See e.g. W FitzPatrick, 'The Intend Foresee Distinction and the Problem of 'Closeness' 128 (2006) Philosophical Studies 585.
[8] Ibid.

Foreseen Harms Provide Intuitive Support for the Beyond Reasonable Doubt Standard

When arguing that retributivism should be held to the beyond reasonable doubt standard, Caruso invokes Tadros's observations that when a criminal is subjected to (retributive) punishment, "relationships are destroyed, jobs are lost, the risk of the offender being harmed by other offenders is increased, and all at great expense to the state."[9] Caruso then contends that, "given the practical importance of moral responsibility to legal punishment and given the gravity of harm caused by legal punishment (to the individuals punished as well as those family and friends who depend upon the imprisoned for income, love, support, and/or parenting), the proper epistemic standard to adopt is the prudential burden of proof beyond a reasonable doubt."[10] Many of the harms that are invoked to motivate applying the beyond reasonable doubt standard to retributivism are "causally downstream" of the punishment itself – such as the impact on the offender's family, friends and job – and therefore could be characterised as foreseen rather than intended. Retributivists would certainly contest the idea that when retributively punishing an offender the state *intends* to harm the offender's family. If a major part of the case for applying the high epistemic standard of proof to penal theories (e.g. retributivism) rests on the *foreseen* harms that flow from implementing these theories, it would be inconsistent to argue that the PHQ model should be exempt from that standard on the basis that this model merely involves foreseen harms.

Meeting the High Standard of Credibility

There are some reasons for thinking that the PHQ model meets the required standard of credibility, or, at least, that it could form a major component of an approach that would satisfy that standard. Although, a full defence of this position is outside the scope of this commentary, I will outline some considerations that suggest that decision-makers under conditions of "moral uncertainty"[11] would, firstly, favour some of the core claims of the PHQ model and, secondly, would favour the *outcomes* concerning who should/should not be subjected to state coercion and concerning the severity of coercive measures that are implied by the PHQ model.

The rationale for requiring that penal (or quasi-penal) theories are held to a high standard of credibility is the idea that, under conditions of moral uncertainty, we should adopt a cautious approach that prioritises avoiding inflicting serious, unjustified harm. This cautious approach rests on the idea that there is more certainty about the idea that we should avoid unjustifiably inflicting serious harm than there is about any substantive theory of punishment. Vilhauer has persuasively argued that we can have more certainty about propositions about which most mainstream theorists agree, than about principles concerning which they are divided.[12] In contrast to the entrenched division of opinion about penal theories, all mainstream ethical theories agree that we have a strong reason to avoid deliberately inflicting serious harm on others (even though they differ as to *why* this is so and disagree about which reasons can override this harm avoidance principle). [13] It seems plausible, that decision-makers under

[9] V Tadros, The Ends of Harm: The Moral Foundations of the Criminal Law (OUP, Oxford 2011) 1.
[10] Caruso (n3) 112.
[11] Defined in para 1.
[12] B Vilhauer, 'Taking Free Will Skepticism Seriously' (2012) 62 *The Philosophical Quarterly* 833
[13] Ibid.

conditions of moral uncertainty should be reluctant to override principles about which there is such a high degree of consensus. Specifically, it can be argued that other considerations should only be allowed to override the principle of harm avoidance if we have a similarly high degree of credence in these overriding considerations. The PHQ model identifies a reason for overriding the harm avoidance principle that seems hard to dispute. A core claim of the PHQ model is that society has a right, grounded in self-defence, to detain those whose criminal conduct demonstrates that they pose a serious threat to others, e.g., rapists and murderers. If a theorist, who had previously endorsed retributivism, came to believe that there is not enough certainty about the soundness of retributivism to allow retributivism to override the harm avoidance principle what should this theorist do? Surely, it would not be rational for such a theorist to say that, if retributivism can no longer provide a sufficient basis for interfering with the liberty of rapists and murderers, such offenders may not be interfered with at all.[14] True, a retributivist, *qua* retributivist, would not take dangerousness to be part of the justification for interference with an offender's liberty. But it seems hard to deny that such a theorist, *qua* ethical theorist under conditions of moral uncertainty, would agree with Caruso that some interference with the offender's liberty would be justified.

Endorsing a cautious approach when reasoning under conditions of moral uncertainty seems to imply that one should opt for a model of criminal justice which (compared with other theories) would recommend subjecting a relatively small number of people to state coercion and which would, overall, recommend relatively lenient responses to criminal behaviour.[15] Caruso's theory would fit this description. Caruso provides a detailed defence of a range of robust safeguards against unjustified punishment. For example, his account stresses the importance of liberty, which he argues should only be infringed in accordance with various principles, including the "principle of least infringement, which holds that the least restrictive measures should be taken to protect public health and safety".[16] The cumulative effect of these safeguards is that Caruso's theory is among the most lenient of mainstream approaches to dealing with criminal behaviour. Decision-makers under conditions of moral uncertainty would therefore have reasons to favour the outcomes implied by Caruso's theory, even if they disagreed about his rationale.

It is possible that there is more than one equally rational strategy for making decisions under conditions of moral uncertainty. An approach, which I have defended elsewhere, would also recommend ways of dealing with offenders that resemble the outcomes implied by the PHQ model.[17] I propose a "convergence requirement" according to which a person should only be punished if there is a sufficient degree of agreement, from the perspectives of each of the main penological theories, that punishing that person is appropriate. The convergence approach minimises the risk of inflicting unjust harm by providing more than one plausible rationale for a decision to impose hardship on an offender (e.g. positive retributivism and societal protection). This provides a 'theoretical safety net' – i.e. even if positive retributivism fails to deliver an adequate justification for imposing hardship, the decision may still be justified on the basis of societal protection and *vice versa*. The consensus approach minimises

[14] K. Murtagh, 'Free Will Denial and Punishment' (2013) 39 (2) Social Theory and Practice 223.
[15] Vilhauer (n1).
[16] Caruso (n3) 185.
[17] Shaw (n1).

the risk of inflicting unjust harm by only harming offenders if there is a reason for doing so in which all theorists can have a high degree of credence.

The convergence requirement is based on the idea that the probability of a disjunction as a whole being true is a function of the probability of the truth each of its disjuncts taken together with the number of those disjuncts. So, if retributivism recommends punishing a particular person (and retributivism has a certain probability of being true) and a theory based on societal protection also recommends punishing that person (and the societal protection view has a certain probability of being true); then the probability that this person ought to be punished is higher than if we relied on one of the disjuncts alone. The convergence requirement would generate similar outcomes to the PHQ model, because both approaches would rule out the most contentious cases of punishment, e.g., framing innocent people[18] and punishing people who are retributively blameworthy, but who are completely non-dangerous.

Conclusion

This commentary argued that the PHQ model (although non-punitive) should be held to the same epistemic standard as penal theories, because it resembles punishment in certain relevant respects. However, it suggested some reasons for thinking that the PHQ model meets the required standard of credibility, or, at least, that it could form a major component of an approach that would satisfy that standard.

[18] Caruso (n3) 192-194.