Diametros (2024), 79: 132–145
doi: 10.33392/diam.1931

Expanding The Scope of The Epistemic Argument to Cover Nonpunitive Incapacitation

– Elizabeth Shaw –

Abstract: A growing number of theorists have launched an epistemic challenge against retributive punishment. This challenge involves the core claim that *it is wrong (intentionally) to inflict serious harm on someone unless the moral argument for doing so has been established to a high standard of credibility.* Proponents of this challenge typically argue that retributivism fails to meet the required epistemic standard, because retributivism relies on a contentious conception of free will, about whose existence we cannot be sufficiently certain. However, the scope of the epistemic challenge should not be limited to doubts about free will or retributivism. In this article, I argue that the epistemic challenge should be expanded beyond the original focus on justifications of punishment. By "expanding the epistemic challenge" I mean demanding that other purported justifications for serious (intentional) harm be held to a high standard of credibility. To provide a focus for the argument, I will concentrate on the "Public Health Quarantine Model" defended by Gregg Caruso, but my arguments have wider implications beyond this model. A growing number of "abolitionist" theorists believe that punishment is wrong in principle. If retributive punishment, or punishment in general, were abandoned, we would need to ask, "how else should we respond to crime?". My arguments suggest that all such abolitionists will have to face the same epistemic standard as penal theorists if they wish to replace punishment with the intentional imposition of non-punitive severe coercive measures.

Key words: free will, skepticism, retributivism, criminal justice, punishment, epistemic argument, G. Caruso

Submitted: 13 July 2023
Accepted: 11 March 2024
Published online: 19 April 2024

1. Introduction

Broadly speaking, retributivism is the view that punishment is justified because criminals deserve to suffer in proportion to their moral blameworthiness. A growing number of theorists have launched an epistemic challenge against retributive punishment.[1] This

Dr Elizabeth Shaw
University of Aberdeen
School of Law
Taylor Building, Old Aberdeen
Aberdeen, AB24 3UB, UK
Email: eshaw@abdn.ac.uk

[1] E.g., Pereboom (2001), Double (2002), Waller (2011), Vilhauer (2009), Shaw (2014), Kolber (2018), Corrado (2019), Caruso (2020), Chiesa (2020).

challenge involves the core claim that *it is wrong (intentionally) to inflict serious harm on someone unless the moral argument for doing so has been established to a high standard of credibility.* Proponents of this challenge typically argue that retributivism fails to meet the required epistemic standard, because retributivism relies on a contentious conception of free will, about whose existence we cannot be sufficiently certain. However, the scope of the epistemic challenge should not be limited to doubts about free will or retributivism. Some theorists have already expanded the epistemic challenge to include other contentious aspects of retributivism and/or to cover non-retributive theories of punishment.[2] By "expanding the scope of the epistemic challenge" I mean demanding that other purported justifications for serious (intentional) harm be held to a high standard of credibility.[3] In this article, I will argue that the epistemic challenge should be expanded beyond the original focus on justifications of punishment. Those who wish to abolish punishment should still have to satisfy a high standard of credibility if they wish to replace punishment with the intentional imposition of non-punitive measures that are still coercive and severe.

My arguments apply to all non-punitive approaches to criminal behavior that have the following characteristics: they authorize the active, intentional imposition of coercive measures on (alleged) offenders, and these measures are severe in that they involve significant hardship and carry stigma (even if the authorities do not intend them to be stigmatic). To provide a focus for the argument, I will concentrate on the non-punitive approach defended by Gregg Caruso – the Public Health Quarantine (PHQ) model.[4] I will focus on the PHQ model, because, in my view, it is one of the most highly developed, persuasive, and original non-punitive accounts of criminal justice proposed in recent years. Caruso's model is both empirically well informed and philosophically sophisticated and is grounded in a wealth of comprehensive and up-to-date research into justifications for punishment and studies on the effectiveness of penal policies.

Caruso argues that the PHQ model is exempt from the epistemic argument because 1) the PHQ model is non-punitive and 2) the PHQ model merely involves foreseen, unintended harm. In section 2, I will briefly describe the PHQ model, agree that the PHQ model is non-punitive, but argue that this does not exempt it from the high epistemic standard because the measures imposed under the PHQ model resemble punishment in relevant ways. In section 3, I will reply to Caruso's second reason for exempting the PHQ model from the high epistemic standard. I will question his claim that that harm inflicted on offenders under the PHQ model is merely foreseen and unintended (in the sense that is relevant to the doctrine of double effect). It is often assumed that the epistemic challenge only targets punitive theories, and especially retributivism. This article contests this assumption, arguing that non-punitive approaches to criminal behavior should also be held to a high standard of credibility. Ultimately, I do not think the arguments presented in this article undermine non-punitive theories, like Caruso's

[2] E.g., Hanna (2023).

[3] The current article focuses specifically on justifications for imposing coercive criminal justice measures on offenders. The epistemic argument may also have relevance for other types of harm, such as certain civil measures, see fn. 26, but this is beyond the scope of this article.

[4] Caruso (2021).

PHQ model. Instead, I offer advice on the best strategy for defending such non-punitive approaches. Rather than attempting to exempt the PHQ model from the required standard of credibility, Caruso should turn his attention to establishing that his theory meets this standard. I have argued elsewhere that the latter strategy is likely to succeed.[5] The current article builds on my previous work,[6] but offers new reasons for thinking that the PHQ model resembles punishment, and that Caruso cannot defend his claim about justificatory standards by appealing to the intend/foresee distinction.

2. The PHQ Model is Nonpunitive but Resembles Punishment

2.1. What is the PHQ Model?

Key claims/characteristics of the PHQ model include:

a) Unlike retributivism, the PHQ model does not depend on the conceptions of "moral responsibility" or "free will" at issue in the free will debate. Instead, it is based on the following analogy: just as it is sometimes permissible for individuals to use force in self-defense and for the state to quarantine carriers of certain infectious diseases, even when the attacker or disease-carrier was not morally responsible for posing a threat, analogously the state may incapacitate dangerous offenders.

b) Coercive criminal justice measures should only be imposed on offenders who pose a serious threat and should be the least restrictive measures required to avert that threat.

c) Rather than focusing just on sanctions, the state should prevent crime by reducing the shared social determinants of ill-health and offending and by tackling unjust social inequalities.

d) The state must safeguard the dignity and welfare of offenders.

Proponents of the epistemic argument often claim that a theorist's *overall justification for punishment* must be held to an epistemic standard resembling the beyond reasonable doubt (BRD) standard. This article maintains that if that is true, then the PHQ model should be held to a similar standard. I argue that a sharp division cannot be drawn between justifications for punishment and the PHQ model regarding the *level* of the epistemic standard. Given my focus, it is not within the scope of this article to discuss in detail what Caruso would need to do to meet the relevant standard. I make (but do not defend) some assumptions about this. I assume that he would (at least) need to establish to a high standard of credibility that we have the kind of self-defensive rights sketched in a) above, and to establish that the risk posed by any given offender warrants the use of coercive measures. I make two assumptions about what Caruso does *not* have to prove to the BRD (or similar) standard. Firstly, although he is a free will skeptic, I assume he does not need to prove lack of free will, as he claims coercive measures can be justified *regardless* of whether free will exists. Secondly, I assume that his overall justification for a coercive measure could succeed even if it could not be proved (BRD) that the offender

[5] Shaw (2021).
[6] Ibidem.

would have caused serious harm but for the measure. The use of coercion to prevent a relatively *low risk* of sufficiently serious harm might be justified BRD. For example, if there is 10% risk that pressing a certain button would destroy the world, the use of coercion to prevent people from pressing the button would be justified BRD.

2.2. The PHQ is Nonpunitive

I accept that the PHQ model is non-punitive, even though I deny that it involves merely foreseen, unintended harm. I should note at this point that I will be using an "expansive" conception of "harm" which refers to the imposition of a loss or a *pro tanto* bad thing or the deprivation of a *pro tanto* good thing.[7] Examples of relevant harms include "loss of liberty and deprivation" of the "benefits" of living freely in one's community.[8] On this understanding of harm, a harm does not have to be bad for the person all-things-considered.[9] I will use harm in this sense because Caruso uses it that way. However, readers who prefer different terms such as "hardship" or "burden"[10] could substitute those terms when I use "harm." Regardless of whether they speak of "harm," theorists generally agree that intentionally imposing the kind of harms/hardships/burdens involved in punishment requires strong justification.

One should class the PHQ model as non-punitive if one accepts the plausible proposition that, for a measure to count as punishment, it is necessary for the measure's harmful/burdensome character to be a *motivating reason* for imposing it.[11] On retributive or deterrence theories of punishment the appropriate punishment is selected partly *because* it has the characteristic of being harmful/burdensome. Otherwise, it would not inflict enough deserved suffering or discourage (re)offending. In contrast, when an offender is deprived of his liberty under the PHQ model, or when a blameless carrier of an infectious disease is quarantined, the motivating reason is to protect society. The measure was not selected *because* it inflicts a *pro tanto* bad thing on the offender or carrier or *because* it deprives him of a *pro tanto* good thing. Hence detention of offenders under the PHQ model is not punishment, for the same reason that quarantine of disease-carriers is not punishment.[12]

However, this does not establish that the PHQ model merely involves *foreseen, unintended harm* in the sense that is relevant to whether it needs to meet a high justificatory standard. To support his claim that justifications of punishment should be held to a substantially higher standard than the PHQ model, Caruso cites "double effect" (DE) – the widely supported principle that intended harm is much harder to justify than foreseen harm. Yet, establishing that something was not the motivating reason for one's action does not establish that it was merely foreseen in the sense relevant to the principle of DE. A sentencing judge under the PHQ model, when intentionally depriving

[7] Hanna (2022).
[8] Caruso (2021): 111 and 13.
[9] Hanna (2022).
[10] Hoskins and Duff (2021).
[11] Hanna (2022).
[12] It is beyond the scope of this article whether the PHQ model does not satisfy various other necessary conditions for punishment.

an offender of his liberty, would not merely *foresee* that deprivation of liberty will *cause* a *pro tanto* bad thing or a loss to befall the offender or *cause* him to be deprived of a *pro tanto* good thing. The intended sentence – deprivation of liberty – *constitutes* a loss, a *pro tanto* bad thing or the deprivation of a *pro tanto* good thing (i.e., a "harm" in Caruso's sense). It has been persuasively argued that DE is only applicable or morally relevant when the relationship between the putative intended effect and the putative foreseen side-effect is causal rather than constitutive.[13] So, as I will further argue in section 3, it is not clear that Caruso can defend his claim about justificatory standards by appealing to DE. In the rest of section 2, I will focus on the PHQ model's resemblance to punishment in that they both pass a certain threshold of severity.

2.3. The PHQ Model Resembles Punishment

This section will argue that the PHQ model's non-punitive status should not exempt it from the high epistemic standard, because the measures imposed under the model resemble punishment by involving severe adverse consequences for offenders. Caruso has responded to my argument by highlighting a number of respects in which the PHQ model would be less severe (and more constructive) than practices in current legal systems. [14] For example, the PHQ model would a) prevent crime through addressing its social determinants; b) provide more alternatives to incarceration, e.g., mental health/drug treatment and social work interventions; c) criminals would be detained, where necessary, in much more humane, rehabilitative environments; d) life without parole would be abolished; e) three strikes laws would be repealed; f) voter disenfranchisement of felons would be ended; g) offenders' rights would only be limited to the extent necessary to remove the danger offenders pose, and they would be "placed in the lowest possible security regime."[15]

I agree with Caruso that these are great strengths of the PHQ model, which help to make it one of the most humane and defensible approaches to criminal behavior. The measures imposed under the PHQ model would generally be milder than those imposed by criminal justice systems in the US and UK. Nevertheless, on the PHQ model the offender would not simply be invited for a cup of tea and a chat. It is plausible that the hardships imposed as part of the PHQ model would pass the threshold of severity that makes it appropriate to apply a high epistemic standard to the argument for imposing such measures. The measures imposed on offenders would still be a) *coercive;* b) would often involve *involuntary detention;* c) in the case of very dangerous offenders the *detention could last for many years*; d) these measures would significantly interfere with *fundamental rights*, e.g., freedom of association and assembly, the right to private and family life, freedom of movement, and (especially in the case of rehabilitative medical treatments) the rights to bodily and mental integrity; e) the measures would often cause *psychological distress*; and f) they would also inevitably be s*tigmatic*, even if that was not the authorities' intention ("intention" will be discussed in section 3). In these important ways, the measures supported by the PHQ model pass the threshold of severity that is relevant to the epistemic argument.

[13] FitzPatrick (2006): 585.
[14] Caruso (2021[b]).
[15] Caruso (2021[b]): 211–212.

As I said in an earlier response to Caruso,

> 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 between this model and punishment are *relevant* to the standard of credibility and that these differences are *weightier* than the relevant similarities.[16]

Since the differences Caruso cites pertain to the severity of the hardship imposed on offenders, they are potentially relevant to the standard of credibility. However, I maintain that they are not weighty enough to show that the PHQ model should be held to a significantly lower standard of credibility than other approaches to criminal behavior. For one thing, Caruso would (rightly) demand that even the most humane versions of retributivism should be held to a high standard of credibility. Yet, these versions of retributivism would endorse many of the same policies that Caruso supports.[17] The hardships endorsed by the PHQ model are not so much more lenient than those endorsed by rival criminal justice theories to warrant holding the PHQ model to a significantly lower standard of credibility.

Secondly, some of the hardships currently imposed in the name of punishment (e.g., fines, suspended sentences, and some forms of community service) are less severe than some of the hardships that would be imposed in the name of the PHQ model. On the PHQ model, very dangerous offenders could be detained for decades and, although parole would always be a possibility, some offenders, in practice, could spend the remainder of their lives in detention. It would be implausible to demand that a retributive argument for a fine, community service or a suspended sentence should be held to the BRD standard while holding Caruso's argument for detaining someone for decades to a considerably lower standard of credibility. True, the retributivist's justification for the mild penalty (for a minimally culpable offender) would be intentionally punitive. Whereas Caruso's rationale for a lengthy period of detention for a dangerous offender would not be intentionally punitive; and I agree that intention can carry some moral weight. However, to play the role that Caruso requires of it, the concept of "intention" would have to do an implausibly huge amount of work. How could the intentionality consideration require a retributive argument for a very lenient punishment to be held to the beyond reasonable doubt standard, but allow Caruso to bear a much lighter epistemic burden when arguing for the imposition of a much more severe measure? It could only do this if, in this context, intention carried vastly more weight than the severity of the harm. However, this is implausible. Indeed, when defending the epistemic argument, Caruso himself spends much more space elaborating on the harms inflicted by retributive punishment than discussing the concept of intention. It seems plausible that harm/hardship is one of the *main* factors that motivates the core idea behind the epistemic argument – the idea that that it is wrong to inflict serious hardship on people unless we have very strong grounds to believe that doing so is justified.

[16] Shaw (2021): 157.
[17] Jeppsson (2021): 177.

Thirdly, Caruso's arguments for holding the PHQ model to a lower standard of credibility than that applicable to penal theories would have troubling wider implications that he would not accept. To say that the PHQ model should be held to a lower epistemic standard because it is more humane and lenient than rival approaches and because it is not intentionally punitive would raise an important question: why then should the beyond reasonable doubt standard apply to establishing that someone committed a crime in the first place? Now, Caruso argues that his model should retain the current rule that the prosecution must prove *actus reus* and *mens rea* to the BRD standard.[18] However, if the measures that would be imposed on the offender under the PHQ model would be lenient, humane, and non-punitive, and would not involve intentional harm, and if these very considerations warrant holding the PHQ model as whole to a lower epistemic standard, then, by parity of reasoning, should not the standard of proof at the trial stage also be lowered if society adopted the PHQ model?[19]

Caruso has also provided a further response, that the similarities in "the *methods* used in incapacitation" and "those used in punishment" are less important than the difference in the *justification* for the PHQ model compared with the justification for retributive punishment. He insists that

> the real thing that sets the public health model apart is that the justification it provides for incapacitation and other liberty-limiting restrictions appeals to the right of self-defense and defense of others and not a retributive justification…Eliminative harming is much easier to justify [than retributive harming].[20]

I will discuss the definition and moral significance of eliminative harming in section 3 below. For now, it is enough to say that I agree eliminative harming is "easier to justify" than retributive harming; but it does not follow that arguments for eliminative harming are exempt from being held to a high epistemic standard. Rather, "being easier to justify" could simply mean that it is easier to show that eliminative harming *meets* this epistemic standard. In fact, Caruso's reasons for saying that eliminative harming is easier to justify than retributive harming seem relevant to whether the high epistemic standard *has been met*, not to whether the high epistemic standard *is applicable*. Caruso states that "justifying retributive legal punishment is difficult because the claim that agents are morally responsible in the basic desert sense…. faces powerful and unresolved objections and *as a result falls far short of the high epistemic bar*"; whereas eliminative harming does not rely on this dubious conception of moral responsibility.[21] By saying that retributivism "falls far short of" the epistemic standard, Caruso himself has switched to talking about whether the standard has been met, not whether it is applicable. In general, the fact that there are many doubts surrounding a certain argument is more relevant to

[18] Caruso (2021): chapter 9.
[19] Caruso might attempt to distinguish between the standard of proof applicable to criminal justice theories and the standard of proof applicable at the trial stage. If Caruso attempted to make this maneuver, he would need to do so in a way that could not also be adopted by retributivists.
[20] Caruso (2021[b]): 210-211.
[21] Caruso (2021[b]): 211, emphasis added.

whether the argument meets the required standard of credibility than to the applicability or demandingness of the standard of credibility.

I maintain that the severity of the hardship to be imposed on someone is one of the main factors that determines whether a higher or lower epistemic standard applies to the argument for imposing hardship. Whereas the level of doubt about whether harm should be imposed is relevant to whether the epistemic standard has been met, not how demanding the standard should be. For an analogy, consider the difference in the standard of proof that applies when proving facts in a criminal case, versus a civil case. For civil cases the lower balance of probabilities standard applies.[22] For criminal cases the higher beyond reasonable doubt standard applies.[23] This is not because there are *more doubts* about the factual basis for criminal liability than about the factual basis for civil liability.[24] Rather, a more plausible moral rationale for the difference in standards of proof is based largely on differences in the nature and severity of the hardship that would be imposed on someone.[25] Criminal sanctions are generally more severe and stigmatic/condemnatory than civil ones.[26]

3. Does the PHQ Model Involve Intentional Harm?

In the previous section, I argued that severity of harm is one of the main considerations when deciding whether a purported justification for harm should be held to a high epistemic standard. One might agree with this while maintaining that whether harm was *intended* makes *some difference* to the epistemic standard (and this difference might be important in certain cases). So, I will now examine Caruso's claim that the PHQ model should be held to a lower epistemic standard than justifications for (retributive) punishment, because the former merely involves foreseen, unintended harm, whereas the latter involves intended harm.

I have argued that when the state intentionally imposes severe burdens on an offender, such as lengthy detention 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.[27] Caruso responds that

---

[22] See, e.g., *Miller* v *Minister of Pensions* [1947] 2 All ER 373.

[23] See, e.g., the following English cases: *Woolmington* v *DPP* [1935] AC 462; *R* v *Hunt* [1987] AC 352.

[24] If an allegation is unlikely, it is harder to meet the standard of proof, but "this does not mean… the standard of proof required is higher": *Re H (Minors)* [1996] AC 563, 586 (Lord Nicholls).

[25] See, e.g., the United States case of *In Re Winship* 397 US 358, 361-364 *per* Justice Brennan, "The requirement of proof beyond a reasonable doubt has this vital role in our criminal procedure for cogent reasons. The accused during a criminal prosecution has at stake interest of immense importance, both because of the possibility that he may lose his liberty upon conviction and because of the certainty that he would be stigmatized by the conviction. Accordingly, a society that values the good name and freedom of every individual should not condemn a man for commission of a crime when there is reasonable doubt about his guilt." See also Ball (2011).

[26] Exceptionally, some civil cases are so serious in terms of stigma and potential consequences that there is a moral argument for raising the standard of proof in those civil cases. See, e.g., Ragavan (2014). This would go against the current approach in England and Wales - see *Re H (Minors)* [1996] AC 563, *per* Lord Nicholls at 586. In the United States, for certain serious civil cases, the "clear and convincing evidence" standard applies, which is higher than the usual civil standard, but lower than beyond reasonable doubt - see, e.g., *Matthews* v *Eldridge* 424 US 319 (1976).

[27] Shaw (2021): 158.

> while the distinction between *foreseeable-but-unintended* harm and *intended harm* is sometimes difficult to draw, we should not dismiss the importance of the distinction to the epistemic argument or moral issues in general…[In] accordance with the principle of [DE], it is permissible…to restrict the liberty of those individuals who pose a significant threat to public health and safety, provided the harm caused by such restrictions is unintended… [This] kind of eliminative harming is significantly different than the kind of harming involved in retributive punishment which *intentionally* seeks to impose harm or harsh treatment…[28]

I agree with Caruso that it is *sometimes* possible to draw the intend/foresee distinction and that this distinction is *sometimes* morally relevant. However, I am not convinced that *in the specific context under consideration* the intend/foresee distinction can be clearly drawn.

Caruso states that the PHQ model involves "eliminative harm" and is grounded in the right to self-defense and defense of others. So, it would strengthen Caruso's argument if it could be shown that when people cause eliminative harm and act in self-defense they merely cause foreseen, unintended harm. However, this is not clearly the case. Caruso draws on the conception of eliminative harm developed by Victor Tadros who introduces this concept when analyzing self-defense. Tadros writes,

> "[t]here are different kinds of *intentional* harming. Here are two. We might intentionally harm a person as a means to a further end. Or we might *intentionally* harm them simply to negate the threat that they currently pose [i.e., eliminative harming]. The latter kind of *intentional* harming would not involve harming the person as a means…[29]

He goes on to explain that

> [s]ome people think that it may be permitted *intentionally* to harm another if harming that person eliminates a threat that they pose (eliminative harming) but that it is wrong intentionally to harm another person as a means to avert a threat that they do not pose (manipulative harming).[30]

It is clear that Tadros conceives of eliminative harming as a kind of intentional harming. Caruso does not defend an alternative definition of eliminative harming, but instead relies on Tadros's account. Tadros is far from alone in thinking that self-defense involves intentional harming.[31] According to Caruso, Aquinas argued that killing in

[28] Caruso (2021[b]): 212-213.
[29] Tadros (2011): 242, emphasis added.
[30] Tadros (2011): 267, emphasis added.
[31] See, e.g., Uniacke (1994), McIntyre (2001), Leverick (2006), Sangero (2006). One of the few contemporary theorists to base self-defense on DE is Kaufman (2009). Yet, even Kaufman acknowledges that this view "has become widely ignored or rejected in mainstream moral philosophy… as an explanation of self-defense."

self-defense can be permissible if the death is foreseen but not intended.[32] However, this is not the dominant view in the modern literature on self-defense, and it is even controversial how to interpret Aquinas on this point.[33] If the numerous doubts surrounding retributivism are grounds for rejecting that theory, then it would be inadvisable for the justification of the PHQ model to depend on the idea that justified self-defense always involves unintended harm, since that claim is also doubtful.

Some cases of permissibly causing harm in self-defense involve foreseen, unintended harm, but many involve intended harm. If the defender hides her face behind a metal shield to protect herself from an attacker who is trying to headbutt her, the defender may foresee but not intend that the attacker will ram his head against the metal shield and be harmed. In contrast, if the only way to prevent an unjust attacker from detonating a bomb killing many people is to shoot the attacker through the heart before they can detonate, then it seems permissible to shoot the attacker through the heart even though the defender who is trying their best to fire a lethal shot surely *intends* to do so. To say that permissible self-defense always involves unintended harm would have the counterintuitive implication that the harm involved in the shooting-through the heart example was no more intended than the harm in the metal shield example.[34]

It might be wondered whether this analysis of self-defense contradicts the idea that self-defense involves eliminative harming. Recall that eliminative harming involves harming someone intentionally but does not involve harming them to achieve some further goal. The two analyses can be reconciled as follows: harming in self-defense does, in a sense, harm the attacker as a means to saving the defender's life, but this does not involve the "manipulative harm" that is so hard to justify. Manipulative harm involves using someone as a means to a *further end*, beyond merely negating the unjust threat posed by the attacker. Manipulative harm involves co-opting someone into a plan of your own (a new causal chain that you initiate) rather than just cancelling out a threat initiated by the person you are harming. Manipulative harm can involve harming someone to avert a threat arising from a completely different source, e.g., using an innocent bystander as a human shield, punishing an innocent person to prevent a riot, or killing one patient in order to redistribute his organs to several other patients with organ failure.

---

[32] Caruso (2021[b]): 213.

[33] According to McIntyre (2001: 249) Aquinas used a sense of "intend" in a sense that "is appropriate to discussions of an agent's underlying motives," not in the sense that features in modern discussions of the DE. She notes that her interpretation of Aquinas is "in substantial agreement with" the following writers' interpretation: Montaldi (1986), Anscombe (1982), Levy (1986).

[34] One reason why Aquinas might not have drawn the modern intend/foresee distinction could be because this contemporary debate has been influenced by the development of modern weapons. Glazebrook writes, "The now familiar question of the permissibility of the inevitably fatal pre-emptive shot (or bomb) would not have arisen when Aquinas was writing. A killing in self-defense would then almost invariably have taken place in hand-to-hand combat, for only then would there have been an immediate threat to life. The killing could thus properly be seen as a further consequence of the permissible intention to disable the aggressor" (Glazebrook 1995: 210). They had bows and arrows in Aquinas's time (1225-1274), but these were primarily used in warfare, not private self-defense. Aquinas's justification for killing in warfare was different from his justification for killing in private self-defense cases, because soldiers are acting for the sake of others, not just to preserve their own lives (Aquinas: c. 1273).

Permitting manipulative harm would problematically expand the purposes for which harm may be imposed and would problematically widen the scope of those eligible to be harmed – potentially roping in anyone whenever that would serve the varied goals you may be pursuing. Eliminative harm is a *strictly limited reaction* to a threat initiated by the person you are harming – it is strictly limited to negating the threat the attacker himself poses.[35] Such factors, rather than the foresee/intend distinction explain the difference in permissibility between eliminative and manipulative harm.

So where does this leave Caruso's claim that the PHQ model's infliction of eliminative harm (grounded in the right to self-defense) merely involves unintended harm and so should not be held to the high epistemic standard applicable to penal theories? I agree that eliminative harming (to defend potential victims against unjust attacks) is easier to justify than harming people for reasons of retribution or general deterrence. However, this is not because eliminative harming is unintended. Nor should we hold eliminative harming to a lower epistemic standard than other kinds of harming. It is just easier for eliminative harming to meet the required standard. Why is eliminative harming easier to justify than harming someone for reasons of retribution or general deterrence? Eliminative harming is easier to justify than retributivism, because as Caruso says, eliminative harming does not rest on a highly contentious conception of moral responsibility, but instead depends on the widely accepted idea that we have a right to stop seriously dangerous individuals, such as rapists and murderers, from carrying on violating the rights of others (i.e., we have a right to try to eliminate the threat they pose by humane, necessary and proportionate means). Eliminative harming is easier to justify than general deterrence, because eliminative harming just stops the attacker from carrying on his harmful course of conduct. Whereas general deterrence arguably involves manipulative harm in the sense outlined above – the offender is harmed to make an example of him to deter others, i.e., he is harmed in order to avert threats stemming from other sources.[36]

I have suggested that causing eliminative harm in self-defense is (often) intended; and the justification for such harm does not seem to depend on its being unintended. So, the idea that the PHQ model only causes unintended harm cannot be defended simply

[35] Manipulating or instrumentalizing someone is typically condemned because it involves treating a person as if he were a mere "tool." The fact that there is a meaningful conceptual distinction between eliminating a threat and using as a tool can be demonstrated by considering the morally neutral example of using objects as actual tools. If a dead branch were about to poke you in the eye as you were pushing your way through undergrowth and you grabbed it and snapped it off, you would not have used the branch as a tool – you would just have eliminated the risk that it posed. But if you were to use the branch to dislodge some hard-to-reach object, then you would be using it as a tool, because you would be using it for some further purpose.

[36] Tadros (2011) argues that general deterrence is one of the rare cases when manipulative harm is justifiable. However, his account depends on controversial claims, e.g., that an offender cannot be culpable for pre-determined choices in the sense of culpability that can justify retributive harm, but he can be culpable in the sense that justifies manipulatively harming him; and the claim that the rights of a specific offender's victims to be protected from that offender can be transferred to potential victims of other offenders who might do similar crimes. Because manipulative harm is rarely justifiable, and because this way of defending manipulative harm is very controversial, it is harder for Tadros's theory to meet the high epistemic standard than a theory that is simply based on eliminative harm.

by claiming that the PHQ model involves eliminative harm and is based on the right to self-defense. Perhaps Caruso could point to another consideration to support his claim that the PHQ model should be held to a lower epistemic standard than retributivism. One consideration might be the intention to cause suffering. At one point, Caruso states that "retributivism… requires a much higher [epistemic standard than the PHQ model] since [retributivism] attempts to justify a set of punitive practices and policies that cause *a great deal of intentional pain and suffering*."[37] This formulation is different from his original justification for requiring a high epistemic standard of "those who want to justify legal punishment, since the harms caused in this case are often quite severe – including the loss of liberty [and] deprivation."[38] Harm, in Caruso's original formulation was expansive enough to include "withdrawal of a benefit."[39] However, if Caruso opts for the "great deal of intentional pain and suffering" formulation and abandons the expansive conception of harm, then he cannot justify holding milder versions of retributivism or many non-retributive justifications of "legal punishment" to the high epistemic standard.

Can his two formulations be reconciled? Perhaps he means that intending to deprive someone of liberty only amounts to intentional harm if there is an intention to cause suffering. However, this is implausible. The harm of loss of liberty *per se* is distinct from the suffering that it might cause. This can be seen if we imagine that the motive for depriving someone of liberty is bad. Imagine that Albert pays Brian to unjustly detain his rival, Charlie, for ten years in a secure facility. Brian's motive for detaining Charlie is simply to get the money. Now, Brian could plausibly say that he does not *intend Charlie to suffer*. Maybe Brian has ensured that the detention facilities are as pleasant as possible, and Brian very much regrets any suffering that Charlie might experience. However, Brian could not plausibly deny that he intends to deprive Charlie of his liberty for ten years, and since depriving someone of their liberty for ten years *is* a harm, he could not plausibly deny that he has intentionally harmed Charlie. Similarly, when the authorities under the PHQ model legitimately detain someone for ten years, they do not intend the offender to suffer, but they do intentionally impose on him the harm of loss of liberty. Furthermore, restricting "intentional harm" to "intentional suffering" could not be justified simply by appealing to the "widely accepted" principle of DE, since this principle states that intended harm is harder to justify than foreseen harm, not that intended suffering is harder to justify than other kinds of intended harm.

Conclusion

Gregg Caruso and other proponents of the epistemic argument against retributivism are right to demand that justifications of punishment be held to a high standard of credibility. If retributive punishment, or punishment in general, fail to meet this standard and were abandoned, we would need to ask, "how else should we respond to crime?" If the proposed replacement for punishment involves the intentional imposition of severe coercive measures, then, I have argued, these measures should also be held to a high

[37] Caruso (2021[b]): 213.
[38] Caruso (2021): 111.
[39] Caruso (2021): 13.

standard of credibility. Caruso attempted to exempt his PHQ model from this standard, because 1) it is non-punitive and 2) it supposedly involves merely foreseen, unintended harm. I argued that this attempt to avoid the high epistemic standard fails. The PHQ model is indeed non-punitive. However, the demandingness of the epistemic standard should not depend on punitiveness. It should largely depend on whether the proposed harm passes a certain threshold of severity, which the PHQ model indeed passes. I agree with Caruso that, in accordance with the principle of DE, whether harm was intended could make some difference to the demandingness of the epistemic standard. However, I rejected his claim that the PHQ model clearly involves "merely foreseen" harm in the sense relevant to the principle of DE. The arguments presented in this article are not an attack on the PHQ model. Rather, they are an invitation to defend this model further, to establish that it can meet the same epistemic standard as justifications of punishment are required to meet. My arguments have wider implications beyond the PHQ model. A growing number of theorists believe that punishment is wrong in principle, and, for these "abolitionist" theorists, "a central question is how the state should respond to the types of conduct for which one currently would be subject to punishment."[40] My arguments suggest that all such abolitionists will have to face the same epistemic standard as penal theorists if they wish to replace punishment with the intentional imposition of non-punitive severe coercive measures.

Acknowledgments: I am grateful to the anonymous reviewers for this journal and especially to Przemysław Zawadzki for providing helpful feedback during the writing of this article.

Funding: This paper was produced as part of a research project on "Moral Uncertainty and Criminal Justice" funded by the Royal Society of Edinburgh.

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

References

Anscombe E. (1982), "Medalist's Address 'Action, Intention and 'Double Effect'," *Proceedings of the American Catholic Philosophical Association* 56: 12-25.

Aquinas, Saint Thomas (c. 1273), "Whether It Is Lawful to Kill a Man in Self-Defense?" [in:] *Summa Theologica,* Second Part of the Second Part, Question 64, Article 7, URL = www.newadvent.org/summa/3064.htm#article7 [Accessed 16/04/23].

Ball D. (2011), "The Civil Case at the Heart of Criminal Procedure: In re Winship, Stigma, and the Civil-Criminal Distinction," *American Journal of Criminal* Law 38 (2): 117-180.

[40] Hoskins and Duff (2021), fn10.

Caruso G. (2020), "Justice without Retribution: An Epistemic Argument Against Retributive Criminal Punishment," *Neuroethics* 13(1): 13-28.

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

Caruso G. (2021[b]), "Retributivism Free Will Skepticism and the Public Health-Quarantine Model: Replies to Corrado, Kennedy, Sifferd, Walen, Pereboom and Shaw," *Journal of Legal Philosophy* 46(2): 161-215.

Chiesa L. (2020), "Selective Incompatibilism, Free Will, and the (Limited) Role of Retribution in Punishment Theory," *Rutgers University Law Review* 71: 977-1001.

Corrado M. (2019), "Criminal Quarantine and the Burden of Proof," *Philosophia* 47(4): 1095-1110.

FitzPatrick W. (2006), "The Intend/Foresee Distinction and the Problem of 'Closeness,'" *Philosophical Studies* 128: 585-617.

Glazebrook P. (1995), "'Permissible Killing: The Self-Defense Justification of Homicide' by Suzanne Uniacke," *The Cambridge Law Journal* 54(1): 210-211.

Hanna N. (2022), "Punitive Intent," *Philosophical Studies* 179: 655-669.

Hanna N. (2023), "Against Legal Punishment," [in:] *The Palgrave Handbook on the Philosophy of Punishment,* M. Altman (ed.), Palgrave Macmillan, London: 559-578.

Hoskins Z., Duff, A. (2021), "Legal Punishment," [in:] *The Stanford Encyclopedia of Philosophy*, E. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2021/entries/legal-punishment/ [Accessed 28/03/24].

Jeppsson S. (2021), "Retributivism, Justification and Credence: The Epistemic Argument Revisited," *Neuroethics* 14: 177-190.

Kaufman W. (2009), *Justified Killing: The Paradox of Self-Defense,* Rowman & Littlefield, Lanham.

Kolber A. (2018), "Punishment and Moral Risk," *University of Illinois Law Review* 2, 487-532.

Leverick F. (2006), *Killing in Self-Defense,* OUP, Oxford.

Levy S. (1986), "The Principle of Double Effect," *Journal of Value Inquiry* 20: 29-40.

McIntyre A. (2001), "Doing Away with Double Effect," *Ethics* 111(2): 219-255.

Montaldi D. (1986), "A Defense of St Thomas and the Principle of Double Effect," *Journal of Religious Ethics* 14: 296-332.

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

Ragavan S. (2014), "An Intermediate Standard of Proof in Serious Civil Cases in England and Wales," *Northern Ireland Legal Quarterly* 65: 81-100.

Richard D. (2002), "The Moral Hardness of Libertarianism," *Philo* 5: 226.

Sangero B. (2006), *Self-Defense in Criminal Law*, Bloomsbury Publishing, Oxford.

Shaw E. (2014) *Free Will Punishment and Criminal Responsibility* (PhD thesis), University of Edinburgh, URL = https://era.ed.ac.uk/bitstream/handle/1842/9590/Shaw2014.pdf?sequence=2&isAllowed=y.

Shaw E. (2021), "The Epistemic Argument Against Retributivism," *Journal of Legal Philosophy* 46(2): 155-160.

Tadros V. (2011), *The Ends of Harm: The Moral Foundations of Criminal Law,* OUP, Oxford.

Uniacke S. (1994), *Permissible Killing: The Self-Defense Justification of Homicide*, CUP, Cambridge.

Vilhauer B. (2009), "Free Will and Reasonable Doubt," *American Philosophical Quarterly* 46 (2): 131-140.

Waller B. (2011), *Against Moral Responsibility*, MIT Press, Cambridge.