From Journal of Information Ethics, Vol. 28, No. 1 (Spring 2019) © 2019 Edited by Robert Hauptman by permission of McFarland & Company, Inc., Box 611, Jefferson NC 28640. www.mcfarlandbooks.com

Justice without Moral Responsibility?

Introduction

In recent years, a growing number of theorists have argued that there is an important sense in which no-one is morally responsible for their actions (e.g., Pereboom 2006, Caruso 2019, Vilhauer 2013, Waller 2017, Corrado 2001, Levy 2011, Smilansky 2000, Greene and Cohen 2004). Broadly speaking, moral responsibility depends on two pre-requisites: firstly, an epistemic requirement (such as the ability to understand the nature of one's action and whether it was morally right or wrong) and, secondly, a control or "free will" requirement. The challenge to moral responsibility typically focuses on the second requirement and so theorists who make this challenge can be referred to as "moral responsibility sceptics" or "free will sceptics". There is a vast literature on the various reasons for being sceptical about free will and moral responsibility (for an overview see Caruso 2018). For example, some sceptics argue that if all our actions are ultimately determined by causal factors outside our control, such as our genes and upbringing, then we cannot be responsible for our actions, as these actions would be the inevitable consequence of causal factors for which we were not responsible. Many sceptics also claim that if, on the other hand, determinism is false, we would still not be free or morally responsible, as our actions would be a matter of luck – agents would be "unable to *settle* whether a decision/action occurs and hence [would not] have the control in action required for moral responsibility" (Caruso 2018). It has become increasingly common for sceptics to support traditional philosophical arguments against free will and moral responsibility with empirical evidence, such as findings from neuroscience.

This article will not focus on the *reasons* for rejecting (certain mainstream conceptions of) "free will" and "moral responsibility", but will instead concentrate on the *implications* of doing

so. Some writers fear that rejecting these ideas would have devastating social consequences (e.g. Smilansky 2000, 2001, 2011) . For example, it has been suggested that if the public came to disbelieve in free will and moral responsibility, people would be much more likely to behave immorally. As Nadelhoffer and Wright (2017) have put it, "…in this respect, the possible spectre of widespread scepticism about free will is as much a public policy issue as it is a metaphysical debate". If these fears were well-founded, this would raise the ethical question whether researchers or even governments have a duty to withhold/supress information that supports free will scepticism in order to protect the public from this "dangerous knowledge".[1] In contrast, certain free will sceptics - whom Gregg Caruso (e.g. Caruso 2018) calls "optimistic sceptics" - argue that if this information is communicated properly, it could have neutral or positive effects on people's behaviour (and there is some empirical evidence to support this optimism - Levy et al 2018). Optimistic sceptics may also advocate finding effective ways of communicating arguments about free will scepticism to policy-makers, as, they argue, policies based on free will scepticism could be beneficial for society.

This article will focus on one important social implication of scepticism about free will and moral responsibility: the implication that we should abandon retributivism, which posits that offenders deserve to be punished (in proportion to their wrongdoing) because they were morally responsible for committing crimes, and that those who were not morally responsible for committing crimes do not deserve punishment. Many moral responsibility sceptics argue that, instead of focusing, as retributivists do on making offenders suffer for their past moral wrongdoing, society's approach to criminal behaviour should seek to achieve forward-looking aims, such as preventing future crimes. Is the idea of abandoning retributivism something that we should worry about? In addition to the free will sceptics' claim that retribution is based on an unsound conception of moral responsibility, our current system of dealing with offenders – informed as it is by retributive thinking – faces a range of other serious criticisms, including evidence of its limited effectiveness in promoting important social goals such as rehabilitating and reforming offenders. However, this same criminal justice system, despite its flaws, also

[1] Similar issues have arisen in other contexts (e.g. Kozlowski and Sweanor 2016).

contains vitally important safeguards for human rights, including the rights of offenders. Hence, there is a strong temptation to resist calls from free will sceptics (and others) to alter the current system radically, for fear of "throwing out the baby with the bathwater".

Retributivists do not just aim to provide a positive reason in favour of punishment, and are not only concerned with safeguarding the rights of victims, or society. Retributivism, arguably, also provides a rationale for having important safeguards for the rights of *offenders and those accused of crimes*. Retributivists have claimed, that if society abandoned retribution, the rights of offenders and accused people would not be respected and, in this sense, they would be treated unjustly. Although there are also many other grounds on which retributivists might object to moral responsibility scepticism, the objection that sceptics' proposals would undermine the rights of offenders and accused people deserves particular attention. If successful, this objection would undercut one of the main practical advantages that optimistic sceptics claim would result from applying their theory to social practices and institutions. These sceptics claim that their approach to criminal justice would be more humane than the supposed vengefulness of traditional retributivism. However, if the above-mentioned retributive objection is sound, then far from being humane, moral responsibility scepticism could lead to grave injustices and ill-treatment including: 1) framing the innocent, 2) grossly disproportionately severe punishments and 3) the absence of due process safeguards such as a) placing the burden of proof on the state to provide strong evidence before coercive measures can be imposed on an offender and b) the right of accused people to challenge the case against them.

This article will argue that moral responsibility sceptics can justify safeguards against the three practices mentioned above and can explain why these practices are unjust by referring to non-retributive considerations. However, it will add the following caveats. Firstly, the nature of the safeguards recommended by this non-retributive account are not identical to the safeguards proposed by retributivists (particularly in regard to disproportionality). Secondly, the explanation provided in this article of why it would be "unjust" to fail to have these safeguards will not entirely satisfy a purely retributive conception of justice. Instead, the article aims to

identify types of non-retributive injustice that occur in cases of framing, disproportionately severe punishments, or due process violations and argues that these non-retributive considerations provide *sufficient* reasons to have safeguards against such practices (even if retributivists would not think these considerations constitute a *complete* account of what is unjust about these practices). If this argument is successful, it would weaken the case that moral responsibility scepticism would have dire implications for the criminal justice system.

The account provided here has some similarities to Derk Pereboom and Gregg Caruso's defence of their quarantine-public health model of criminal behaviour, which draws an analogy between offenders and carriers of dangerous diseases (e.g. Pereboom and Caruso 2002). This is one of the most impressive and well-developed models of non-retributive criminal justice and the account defended here is in agreement with their general approach. However, unlike the Pereboom-Caruso model, the argument put forward in this article focuses on an analogy with the treatment of offenders who are non-responsible due to mental disorders, which leads to somewhat different conclusions from those drawn by Pereboom and Caruso. Before developing this argument, the article begins by explaining the sense of "moral responsibility" that is challenged by sceptics and by further analysing the above-mentioned retributive objection to moral responsibility scepticism.

Moral Responsibility and Retribution

Pereboom (2018, p3) provides the following definition of the type of moral responsibility challenged by sceptics, which he calls the "basic desert" sense of moral responsibility:

"For an agent to be morally responsible for an action in the basic desert sense is for the action to be hers in such a way that she would deserve to be blamed if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve to be blamed or praised just because she has performed the action,

given sensitivity to its moral status; and not, for example, by virtue of consequentialist or contractualist considerations"

Moral responsibility sceptics can accept that there are other senses of "moral responsibility" that are not challenged by their arguments. For example Pereboom (2018, p3) endorses the following "moral protest" account of moral responsibility and blame:

"For B to blame A is for B to issue a moral protest against A for immoral conduct that B attributes (however accurately) to A. Such moral protest might indeed have the aims of character formation, reconciliation in relationships, retention of integrity, and protection."

Retributivism (as traditionally formulated) seems to depend on the "basic desert" sense of moral responsibility. For instance, according to the influential version of retributivism defended by Michael Moore (2010), wrongdoers should be punished simply because they deserve it, since "the suffering of the guilty is intrinsically good". On this view, punishment may happen to promote good consequences, such as protecting society from future crimes, but such consequences form no part of the justification for or the intended aim of punishment.[2] In contrast, sceptics, like Pereboom, justify punishment by appealing to the need to protect society. For Pereboom, it seems that moral responsibility in the sense of "moral protest" against wrongdoing can inform *how* offenders are treated (e.g. can form part of a rehabilitation programme), but does not provide the main *justification* for punishment. Pereboom's (2006) justification is based on the idea that, just as society has a right to quarantine carriers of certain

[2] In Moore's version of retributivism, the connection between "basic desert" and the justification of punishment is clear, as for Moore, deserving punishment (like deserving blame) just depends on the offender's wrongdoing, rather than on further consequences that might flow from punishing/blaming. In contrast, the relationship is more complex for those retributive theories that allow that punishment can have forward-looking aims. For "mixed" theories, establishing that someone was morally responsible in the basic desert sense might be seen as providing "permission" for the state to impose hardships on the offender (partly) in order to achieve forward looking aims (Lewis 1953, Strawson 1962). It might be said that basic desert is one element of these mixed theories. However, other retributivists (e.g. Duff 2001) might not seem to rely on *basic* desert, viewed as an element separate from punishment's forward looking aims, since, on Duff's view, forward- looking aims (e.g. reform) are meant to be *internally* connected to the backward looking aspect of punishment (holding offenders responsible for their crimes).

dangerous easily communicable diseases, it has a right to preventatively detain those offenders who pose a serious risk of harm to others.

Approaches that compare wrongdoing to illness and advocate social protection have long been opposed by advocates of basic desert moral responsibility. PF Strawson (1962), who had a profound influence on the free will literature, suggested that it would be psychologically impossible to give up the idea of moral responsibility (in something like the basic desert sense), but that even if this were possible it would be undesirable, as it would mean treating all wrongdoers in the same way that we treat the mentally ill. He claimed this would involve viewing them as things to be manipulated, rather than as persons who have rights. He wrote:

"In the extreme case of the mentally deranged, it is easy to see …the impossibility of what we understand by ordinary inter-personal relationships. Given this …impossibility, no other civilized attitude is available than that of viewing the deranged person simply as something to be understood and controlled in the most desirable fashion."

Similarly, the retributivist, CS Lewis, warned that if retribution were abandoned in favour of harm prevention, responsible offenders would be objectified and would no longer be protected by considerations of justice*:*

> "There is no sense in talking about a...'just cure'... We demand of a cure not whether it is just but whether it succeeds. Thus when we cease to consider what the criminal deserves and consider only what will cure him or deter others, we have tacitly removed him from the sphere of justice altogether; instead of a person, a subject of rights, we now have a mere object, a patient, a 'case'."[3]

[3]. PF Strawson (1962) also advocated drawing a sharp distinction between the norms governing sane and insane law-breakers on similar grounds.

In recent years, Lewis's article has still been cited approvingly (e.g. Morse, Vincent) and other writers (even if they do not endorse his retributivism) have raised related concerns (e.g. Smilansky 2011, Dennett 2011, 2018).

Retributivists maintain that they can provide an intuitively plausible account of justice. Their focus on basic desert provides a simple explanation of why it would be unjust to punish the innocent or to inflict disproportionately severe punishments – these individuals do not deserve such treatment, since they were not, or were not sufficiently, morally blameworthy. It has, however, been argued that retributivists' idea of "proportionality" is too vague or disputed to provide an adequate safeguard against ill-treatment of offenders (Caruso 2018). Retributivists also typically claim that their theory explains the importance of due process rights, e.g. that the state must prove the offender's guilt beyond reasonable doubt. This claim has also been challenged, since retributivism *per se* does not state whether punishing the innocent is worse than not punishing the guilty (Tadros 2012), but such critiques of retributivism will not be explored here. Instead, this article will contest the claim that social-protection approaches to criminal justice, or the sort defended by moral responsibility sceptics like Pereboom, remove offenders from the "sphere of justice" and "rights" altogether. It will argue that offenders would still be protected by non-retributive considerations of justice, in virtue of the fact that they are persons. It will focus on an analogy with those who are non-responsible due to mental illness, as this example has been invoked by critics of moral responsibility scepticism, but, in fact, supports the claim that moral and legal rights should (and to some extent already are) accorded to non-responsible individuals, based on their personhood.

Framing the Innocent

Here is an example that is frequently cited by retributivists to support the idea that retributivism is preferable to forward-looking approaches to punishment, such as consequentialist approaches:

Framing a moral agent

> A horrible act of violence is committed and the culprit cannot be found. A riot will ensue that will harm many innocent people, unless the mob is persuaded that the wrongdoer has been apprehended and punished. So the authorities frame and punish an innocent man. (McCloskey 1972)

Retributivists allege that consistent consequentialists must endorse this, since the authorities' actions promoted the best over-all consequences. Only retributivism, they claim, can adequately explain why the framed person has been treated unjustly. To understand the basis for a non-retributive prohibition on framing the innocent, consider the following example:

Framing someone who is not a morally responsible agent

> A horrible act of violence is committed by an attacker with severe learning disabilities. The attacker cannot be found. A riot will ensue that will harm many innocent people, unless the mob is persuaded that that the attacker has been apprehended and confined in a secure mental hospital. The authorities find a man, Timothy, who has severe learning disabilities, but who has never committed an act of violence before. Timothy is perfectly harmless and has until now enjoyed his freedom to move about the town and interact with the townspeople and wants to be liked by them. However, due to his mental condition, he cannot be considered a morally responsible agent. Because of various circumstances, the authorities are able to persuade the mob that Timothy was the attacker. So the authorities frame Timothy and shut him up in a secure mental hospital (despite knowing that he is perfectly safe and was not the attacker).

It seems intuitive to say that Timothy has been treated unjustly. Retributivists cannot explain this intuition with reference to retributive desert. Retributive desert does not come into it. The authorities do not claim that Timothy *deserves* to be locked up. Timothy is not a responsible

agent and so would not deserve retributive blame even if he had been the attacker. The actual attacker (due to the severity of his mental condition) does not deserve retribution either.

The authorities' actions can be criticised for the following reasons. Firstly, they have wronged Timothy by lying about him. The lie is particularly objectionable because it denies important good qualities that Timothy actually has (e.g. gentleness and friendliness), and falsely asserts that he has serious negative qualities (a propensity to kill innocent people). True, Timothy is not responsible for having these good qualities, nor do the authorities claim that Timothy is responsible for his alleged negative qualities. Nevertheless, it seriously wrongs a person to tell this kind of lie about him. Furthermore, Timothy is detained on the basis of such outrageous lies. This also wrongs him, because the grounds of his detention are illegitimate. Timothy's detention also treats him merely as a means to avert a threat from elsewhere. It does not seek to eliminate a threat that he himself poses.[4] Victor Tadros has persuasively argued that the objection against using someone merely as a means is best characterised as an objection against "*manipulative use*" – where someone is used in order to promote some further, independent goal. In contrast, harming someone to *eliminate* a threat they pose is much easier to justify, based on the right to self-defence. Tadros cites a wide range of examples where it is intuitively objectionable to use someone manipulatively (although he maintains that in exceptional cases the prohibition on manipulative use can be outweighed). Tadros's interpretation of the mere means principle as prohibiting "manipulative use" also seems to provide a plausible explanation of the intuition that framing Timothy is unjust. It will be assumed, for the purposes of this article, that Tadros's distinction between eliminating harm and manipulative use is valid (although the way he uses these ideas in his justification of punishment will not be endorsed). A moral responsibility sceptic, who argues that nobody is morally responsible in the basic desert sense, could still raise the above-mentioned objections against framing an innocent person (regardless of whether the person had a mental disorder), since it is plausible that all

[4] The mere means argument was famously put forward by Kant. However, Kant seems to have tied this argument to the idea of rational agency, so it is not clear whether non-rational agents are protected by the duty of respect for persons as he originally formulated it (Kant 1948, p85). However, it is submitted that the principle of respect for persons should extend to non-rational or partially rational people such as the mentally ill, learning disabled people and young children (See A Wood and O O'Neill 1998).

persons, regardless of their moral responsibility status, have the rights not to be lied about and not to be treated merely as a means (in the sense of manipulative use).

Pereboom and Caruso have also argued that free will sceptics can appeal to the prohibition against treating an offender merely as a means in order to explain why only genuinely dangerous offenders may be incapacitated.[5] Appealing to their quarantine analogy, they argue that the "mere means" prohibition can be outweighed or does not apply in cases where a disease carrier's freedom is restricted to prevent the spread of a dangerous disease, and so, similarly, this prohibition is outweighed or does not apply where offenders' freedom is restricted to prevent them harming others. But if the someone does not pose a direct threat (as Timothy does not in the example above), it would be illegitimate to restrict his/her freedom as a means of promoting the general good. It might be wondered whether the quarantine analogy provides a sufficiently strong basis for protecting non-offenders against state interference. After all, the state can isolate carriers of diseases who have not yet done harm (i.e. they have not actually infected anyone else yet) and can even quarantine those who have been exposed to the disease, *but might not even be carriers.* In response to this potential limitation of the quarantine comparison, Pereboom does invoke a comparison with individuals suffering from mental illness, arguing that it should be required that they have done harm or have expressed a clear intention to do harm before they can be legitimately detained, and that the same sort of requirement should be met before sane offenders can be detained.

However, Pereboom does not use the analogy with mentally ill individuals to rule out *framing* specifically and it is not clear that the way he uses the mental illness example, in the context of the rest of his account, could rule out all objectionable kinds of framing. Pereboom (2018)

[5] Pereboom previously argued that incapacitating dangerous offenders involved using them, but that the use objection was outweighed by the right to self-defence. Subsequently, he endorsed Tadros's analysis, according to which self-defensive harm does not count as "manipulative use" (compare Pereboom 2014 and 2018).

argues that it is "the right to life, liberty, and physical security of the person that have a key role in making the manipulative use objection …intuitive. Those rights are grounded in the more fundamental right to a life in which one's capacity for flourishing is not compromised in the long term." He claims that the presumption against manipulative use is particularly strong, where that use involves "intentional killing, long-term confinement and infliction of severe physical or psychological harm". Yet, framing an innocent person would still be unjust even if the punishment were not severe, or even if there were no punishment.

The retributivist would say that the injustice of framing could be explained in terms of the retributive principle that only wrongdoers should be convicted or subjected to (even mild) punishments. Yet, the example provided in this article suggests that there is another, non-retributive account of the injustice of framing, based on the prohibition against manipulative use, provided that this prohibition is conceived of more broadly than Pereboom suggests.[6] The *false declaration,* which the authorities make in order to quell the angry mob, that Timothy (albeit in a non-responsible state) committed the violent attack, is manipulative and unjust. Furthermore, detaining him in hospital, to satisfy the mob, might not greatly compromise his capacity for "flourishing" in the "long-term", as the conditions in hospital might not be much worse than the conditions he would face in ordinary life and it might not be for that long. But such detention would still be manipulative and unjust. Given that Timothy is non-responsible (and no-one is claiming that he is responsible), this suggests that even if we adopted scepticism about moral responsibility in general and abandoned retributivism, we could still justify a prohibition on framing non-offenders, based on the manipulative use objection (conceived broadly). Furthermore, the moral responsibility sceptic can appeal to a broader range of rights than Pereboom mentions, including reputational damage (which Pereboom does not discuss in

[6] Pereboom's narrow conception of the stringent prohibition on manipulative use, allows him to justify the imposition of sanctions as a limited form of general deterrence, but, as discussed under "Proportionality" (below) this manoeuvre seems questionable.

this context). Timothy's reputation would be damaged by the lie that he committed a violent attack, even if no-one alleged that he was morally responsible for it. This would be unjust even if it did not cause him "psychological harm". His learning disabilities might prevent him from appreciating the wrong that had been done to him and from being psychologically harmed by that knowledge, but he would still have been treated unjustly.

Now, the retributivist will object that the manipulative use prohibition and the reputational damage involved in falsely accusing someone of non-culpably doing harm do not capture everything that is unjust about framing a sane person for a crime they did not commit. For the retributivist, to capture this injustice, one must invoke the idea of moral responsibility. In response, it is conceded that the account defended here will not satisfy those who are already committed to a retributive conception of justice. This article just aims to identify one type of injustice that is involved in the framing of non-offenders that does not depend on the ideas of moral responsibility and retribution. If the argument succeeds, then it is not true that, as some retributivists have alleged, abandoning retributive moral responsibility would mean removing offenders from the sphere of justice altogether. It is submitted that the reason why this non-retributive account can claim to involve a conception of *justice* is because of its focus on individual rights. Unlike consequentialist objections to framing people, the reasons given here seem to capture the intuitive idea that framing is unjust, because the *framed person* has been victimised. The consequentialist rationale refers to some calculation of the general welfare and this fails to capture our intuitions about the injustice done to the individual.[7]

[7] Having said this, there may be some extreme situations in which framing non-offenders is permissible. If the authorities knew that the world would be destroyed unless an innocent person was framed, then framing that person seems permissible in this dire situation. Nevertheless, an injustice would still have been done to the individual, even though it would be permissible on balance to perpetrate this injustice. This is an outcome that most retributivists would accept. At precisely what point consequences can be said to be sufficiently serious to warrant inflicting injustice is a hard question. But it is no harder for the theory being defended here than for retributive theories. Recognising the tension between the need to do justice, and the need to avert bad consequences better captures the complexity of our moral experience, than a theory that claims to produce neat, conflict-free answers to such questions

Benjamin Vilhauer (2013) has proposed a different kind of personhood-based, non-retributive argument against framing the innocent. According to Vilhauer, respecting someone's personhood means treating them in a way that they would rationally consent to be treated. He does not rely on the person's actual consent, but on the notion of 'hypothetical consent'– i.e. they *would* consent to be treated this way *if* they were rational. He uses Rawls's idea of 'the original position' to model rational consent (Rawls 1999). The original position is a thought experiment in which people choose the rules that will govern a society. The rules are chosen behind a 'veil of ignorance': the choosers are unaware of certain facts about what their own position will be in the society and what personal characteristics (e.g. race, gender, wealth, strength, intelligence and industriousness) they will have. They are aware of the fundamental interests that they all have in common (e.g. security and the freedom to pursue one's goals) and they have knowledge of relevant scientific and sociological theories. The veil of ignorance is designed to describe a situation of fairness among the social contractors, to ensure their impartiality and to filter out factors that are just down to luck. Each deliberator must also imagine that he or she is just as likely to be harmed by any principle that is chosen as to benefit from it. Vilhauer, unlike Rawls, includes knowledge of whether one will be a wrongdoer as a factor that is hidden from the social contractors. This is because Vilhauer is a free will sceptic and believes that one's moral character is, like race and gender, a product of the genetic and environmental lottery. Vilhauer claims that respecting someone's personhood means treating them as they would rationally consent to be treated, i.e. in accordance with a principle that would have been agreed to by deliberators in the original position. He claims that no rational deliberator could have chosen the principle that the authorities may, when it is expedient, frame innocent individuals. Such a regime would involve the authorities systematically deceiving the members of this society. Otherwise, the scapegoating of innocent individuals would be ineffective. A deliberator in the original position must acknowledge that under this regime he could be one of those who are deceived about a basic principle governing that society. Consenting to systematic deception undermines one's status as a rational agent. Therefore,

according to Vilhauer, the idea that a rational deliberator would choose to be systematically deceived about something so important is self-contradictory.[8]

Vilhauer's argument is intriguing and could be invoked to supplement the position defended in this article. However, it does not seem to capture the *main* reason why the authorities' actions are wrongful in the two framing cases. Intuitively, the main injustice in both cases is the wrong that has been done *to the framed individual*. However, Vilhauer's explanation focuses on the wrong of deceiving the general public. On Vilhauer's account, the wrong that is done to the framed individual derives from the supposed logical problems with a principle that endorses deceiving the public. This seems too indirect.

Furthermore, it is not obvious that choosing to be deceived by the authorities is necessarily irrational. Imagine that the original position deliberator is considering whether to choose the principle that the authorities must never deceive the public even if that is the only way to prevent a riot. The deliberator must assume that she is equally likely to be harmed by that policy as to benefit from it. In other words, the deliberator must assume that, if the policy were implemented, she might well end up as one of the people harmed or killed in the riot. It is not obviously irrational for the deliberator to prefer the risk of being deceived by the authorities to the risk of being harmed or killed in the riot. It does not seem that Vilhauer's argument can support the strong claim that consenting to such deception is logically contradictory. However, it might support a weaker claim. There is a disturbing paradox in the idea of a rational agent choosing to be systematically deceived and the original position deliberator certainly has reason to hesitate before endorsing such deception. This would not necessarily lead to a complete prohibition on framing innocent individuals in all cases, but it does imply that these cases are always ethically troubling. Perhaps this better captures the conflicting intuitions that are evoked by cases of framing than a principle which categorically prohibits framing 'though the heavens may fall'. If this modification of Vilhauer's argument is successful, then this argument can provide an additional non-retributive explanation of our concerns about framing.

[8] This strategy of arguing is also inspired by Kant.

Proportionality

If desert were abandoned, some fear that the state's response to law–breaking would no longer be governed by principles of proportionality. For instance, Lewis maintained that a medical model of punishment would permit the authorities to interfere with the liberty of citizens, whenever the authorities found this convenient. They would simply label the citizens 'diseased'. He claimed that the authorities could impose on such unfortunate citizens any 'treatment', no matter how burdensome, and any period of confinement, no matter how lengthy. Ordinary people, he maintained, would have no basis for objecting to this on grounds of justice, since 'justice' is a retributive concept (Lewis 1953).

This line of argument is based on a misconception of the principles that should apply to the mentally ill. It is unjust to confine someone or force her to undergo treatment against her will merely because she has a mental illness. She must pose a threat to the safety of herself or others.[9] Furthermore, certain treatments are so risky or so devastating to the individual that it would be unjust to impose them on her, even if she is mentally ill and dangerous. It would also be unfair to impose a particularly lengthy or onerous treatment/confinement on someone if her behaviour only had a relatively minor impact on the welfare of any particular individual.[10] For instance it would be grossly unfair to lock up a mentally ill person for life in a secure institution, just because she made loud noises in the street, causing only minor irritation. This is a consideration of proportionality (though clearly of a non-retributive kind). It is not merely a question of whether the intervention is *necessary* in order to prevent the objectionable behaviour. It is conceivable that for some people, a measure almost as drastic as confinement in an institution might be required in order to prevent them from causing a nuisance. Imposing such a drastic measure would still be unjust. This proportionality constraint is not merely the result of utilitarian calculation. Classical utilitarianism is aggregative. On an aggregative approach if enough people were each caused a tiny bit of distress by the nuisance, then that could eventually outweigh the interests of mentally ill person and justify locking her up. In

[9] See e.g. Mental Health (Care and Treatment) (Scotland) Act 2003, ASP 13.
[10] In the context of a discussion of the punishment of sane offenders, this principle is defended in Honderich 1984, p78.

contrast the proportionality principle defended here states that the intervention must be proportionate to the impact that the harm to be prevented by the intervention would have on *any particular* victim. So a greater intervention, such as lengthy confinement, would be justified to prevent killing or a serious violent or sexual attack. Whereas a much more minor intervention, such as counselling, or supervision in the community would be justified to prevent nuisances. The proportionality principle is based on respect for the separateness of persons and on an ideal of equality – it is *prima facie* wrong to create a situation where people suffer grossly unequal levels of distress.[11]

If this principle of proportionality applies to insane law-breakers who are clearly not deserving of retribution, then an analogous principle of proportionality would also be available to sane offenders under a non-retributive system. It might be objected that the proportionality principle does not give very precise recommendations about the exact degree of burdensomeness that is appropriate in each case. However, this "vagueness" objection is arguably even more of a problem for retributive conceptions of proportionality (Caruso 2018).

Pereboom and Caruso have also defended something similar to the non-retributive proportionality principle outlined above, based on the quarantine analogy. They argue that the coercive measures imposed on offenders should be proportionate to the harm offenders pose and should constitute the least infringement of their rights necessary to protect society from their harmful conduct. However, Pereboom (2018), has recently departed somewhat from this model in response to criticisms from general deterrence theorists. Although some level of deterrence may result as a side-effect of a system based on incapacitation (which Pereboom "free deterrence"), Pereboom concedes that a greater level of deterrence might be desirable. He therefore argues that it can be justifiable to impose measures on an offender that are "somewhat" harsher than would be necessary to protect society from the harm the offender poses, provided that this would provide substantial benefits in terms of general deterrence or

[11] Like most of the principles of justice defended here, this is a strong presumption, but not necessarily an absolute prohibition in all cases. As noted earlier, retributivists themselves often admit that principles of justice can sometimes be outweighed if the consequences are serious enough.

cost-effectiveness. This would might seem to breach the prohibition on manipulative use, since part of the sanction is imposed in order to promote a further goal, independent of the need to eliminate the threat posed by the offender. Pereboom justifies this by invoking the idea (mentioned in the section on framing, above) that the prohibition on manipulative use applies most stringently to using people in a way that that would seriously compromise their capacity for to lead a life at a reasonable level of flourishing. He argues that short prison terms and large fines can be justified on the grounds of general deterrence or cost-effectiveness, as these sanctions would not seriously compromise their capacity for flourishing.

Does the analogy with mentally disordered offenders imply, contrary to Pereboom's analysis, imply that imposing sanctions to achieve general deterrence would violate non-retributive principles of proportionality (and the prohibition on manipulative use)? Well, it does not seem to rule out what Pereboom calls "free deterrence". A measure may be used as a deterrent, provided that it is *also* strictly necessary in order to incapacitate the dangerous person. It is possible that someone may be non-responsible, due to her mental condition, but also capable to a certain extent of being deterred. For instance, a person with severe learning disabilities may understand that some form of behaviour (e.g. running into the road, or being violent) will result in a negative consequence for her (e.g. she will have less freedom, and be subject to greater supervision). The thought of this negative consequence may help to restrain her from engaging in the dangerous behaviour. It is not wrong for her carers to explain to the person (in humane, non-inflammatory terms) that these negative consequences will occur as a result of such behaviour and have been imposed on others. They may explain this in the hope that this will affect the conduct of the person with learning disabilities. The knowledge that mentally ill offenders will still be confined, if dangerous, may also deter some sane offenders from trying to fake an insanity defence. The state does not wrong mentally ill law-breakers by publically pointing out that such law-breakers need to be confined if dangerous. Any deterrent effect such statements may have is no bad thing, provided that the authorities do not use unduly stigmatising and inflammatory language. Therefore, moral responsibility sceptics could also legitimately rely on this type of deterrence, when dealing with sane offenders.

However, the analogy with mentally disordered offenders suggests it would not be legitimate to impose *additional* periods of detention (over-and-above what would be needed for incapacitation) for the sake of general deterrence. It would be unjust to confine a non-dangerous mentally ill person in order to 'make an example' of him, even if this were just for a short period. However, it might be easier for the moral responsibility sceptic to justify imposing fines aimed at compensating victims (in addition to preventative detention, or where detention was unnecessary). In tort law, individuals with mental disorders (no matter how severe) who negligently harm others are still liable to compensate their victims.

How would consequentialist and retributive theories deal with the issue of deterrence? Many consequentialist theories would allow sane people (and possibly also mentally disordered people) to be harmed in order to deter others. However, these theories are vulnerable to the retributive challenge that they allow offenders and accused people to be treated in ways that strike many as intuitively unjust. Pure retributivists would argue that any additional sanction imposed purely for reasons of general deterrence would be disproportionate. It might be thought that mixed consequentialist-retributive theories could have the best of all worlds (although mixed theories face challenges of their own). Similarly, a non-retributive theory of justice might attempt to deal with the problem of deterrence, by allowing consequentialist considerations to play a role.

To conclude this section: There are grounds for opposing grossly disproportionate punishments, which do not depend on retributive moral responsibility. This non-retributive conception of proportionality will not satisfy the retributive conception of justice, but may appeal to the intuitions of those who are not already committed to retributivism. A potential problem with this non-retributive approach is that the sanctions recommended might not be severe enough to promote general deterrence and this might threaten social stability. This is an empirical claim, which has been challenged by non-retributivists (e.g. Caruso 2018). However, it might be argued that non-retributive proportionality principles, could justifiably be

overridden by consequentialist considerations, in cases where social stability seemed to be seriously threatened.

Due Process

Daniel Dennett (2011, 2018), though far from being a traditional retributivist, has recently argued that there will be 'totalitarianism', unless we have a system of punishment based on desert. However, this ignores the fact that important individual rights and rules of due process apply in contexts where desert is not an issue e.g. when the state wishes to restrict the liberty of non-responsible, mentally ill offenders. Such individuals cannot be detained at the mere whim of a totalitarian dictator.

For instance, article 5 (1) of the European Convention on Human Rights (ECHR) provides that such detention must be 'in accordance with a procedure prescribed by law'. Non-responsible individuals are also entitled to challenge the grounds for their detention. Article 5 (4) of the ECHR provides that 'everyone who is deprived of his liberty by arrest or detention shall be entitled to take proceedings by which the lawfulness of his detention shall be decided speedily by a court and his release ordered if the detention is not lawful.' This provision applies to sane people and to people of 'unsound mind'.

Domestic legislation also implements various safeguards which protect the rights of mentally ill persons against infringements by the authorities. The Mental Health (Care and Treatment)(Scotland) Act 2003 provides that a mentally ill person who may be subject to compulsory treatment or hospitalisation is entitled to have her interests defended by a 'named person'. Decisions about compulsory treatment/hospitalisation are made by a Mental Health Tribunal which is independent of the executive and which must consult with and provide information to the mentally ill person and her named person. The burden of proof is on the experts to demonstrate that the mentally ill person poses a 'significant risk' to the safety of herself or others and that compulsory treatment/hospitalisation is necessary. The Mental

Welfare Commission is a separate, independent body whose role is to protect the welfare of individuals who are vulnerable through mental disorder. The mentally ill person or her named person is also entitled to appeal against decisions to impose/continue compulsory treatment or hospitalisation.

Thus it can be seen that several important principles of due process do not depend on basic desert moral responsibility and are applicable to sane and mentally ill individuals. To summarise, these principles include the following: interventions are prescribed by law; the burden of proof is on those who wish to intervene; decisions are made by courts or tribunals that are independent of the executive; the person who may be subject to the intervention is entitled to participate in the process and to be fully informed and adequately represented; persons subject to interventions are entitled to initiate a review of the legitimacy of the interventions. Any non-retributive response to law-breaking should uphold these principles. However, there are further principles of due process that should apply specifically to sane offenders. These will be discussed in the final section.

Differences between Sane Law-Breakers and those with Mentally Disorders

So far, this article has focussed on similarities between the norms governing our response to sane offenders and people who are dangerous due to mental disorder. However, there are also important differences between these groups that cannot be ignored.

Different Methods of interacting with Sane Offenders and People with Mental Disorders

Different methods are appropriate for dealing with the behaviour of mentally disordered as opposed to sane law-breakers. Psychiatric counselling or treatment is typically the best approach for mentally disordered law-breakers. Sometimes it is justifiable to make such

counselling or treatment compulsory, if the ability of the individual to make decisions about her own treatment is compromised by mental illness.

However, the behaviour of sane offenders may change for the better if they come to see the force of the moral reasons against wrongdoing. It is widely accepted that rationality is compatible with determinism, even if retributive desert is not. Presenting offenders with moral reasons for reforming themselves shows respect for the offender's ability to grasp such reasons. As we saw in the example involving Timothy at the beginning of this article, it is important for the state to acknowledge and not deny positive qualities that citizens may have, even if the citizen is not retributively responsible for having those qualities. Rationality is a quality that sane offenders possess and which the state must recognise. Sane offenders might also benefit from certain limited kinds of psychological treatment or enhancement. However, such interventions should only be given to the offender if the offender consents (see e.g. Focquaert 2014, Shaw 2014).

The Trial Process

Restrictions may sometimes be placed on the liberty of mentally disordered people, without ever putting those people through a criminal trial before a jury. This is often the most humane and sensible approach, since the issue of what treatment or supervision such mentally disordered people require is best determined by medical experts.

However, as noted above, moral reasoning, rather than medical help is typically the appropriate means of enabling sane offenders to reform themselves. The trial process can serve as a vivid form of moral communication, which can help the offender to appreciate more fully the impact of her conduct on others and to resolve to change her behaviour (Duff 2001). It also shows respect for the offender's rationality and membership of the moral community to allow her to give an account of her conduct in court, before other members of the community (Duff 2001).

Actual Conduct and Standards of Proof

Before a sentence can be imposed on a sane offender, it must be proved beyond reasonable doubt that the person committed a crime. This principle can be justified on a non-retributive basis. It upholds the value of liberty by protecting the individual against the power of the state. The state also shows respect for citizens by having a very strong presumption that those citizens are non-dangerous. Past behaviour is one of the best guides to future behaviour.[12] It is therefore appropriate that proof that the individual has actually engaged in dangerous conduct should be a necessary condition of interfering with the freedom of sane individuals. The state also shows respect for citizens by having a very strong presumption that their conduct is guided by the fundamental moral values embodied in the criminal law.

However, proof beyond reasonable doubt of actual law-breaking is not a necessary condition for the detention of mentally disordered people who are judged to be dangerous. Can this distinction between sane offenders and the mentally disordered be justified? Well, there are actually some genuine worries about forcing a mentally ill person to undergo treatment and/or confinement, without strong evidence that the individual has actually engaged in dangerous conduct. Reconsider the case of Timothy. Now imagine he is given a routine brain scan and the doctors conclude that he has certain structures in his brain that are strongly correlated with extreme violence. Recall that Timothy has always been gentle and friendly, enjoys wandering round the town and wants to be liked by people. On the basis of the brain scan evidence, Timothy is confined in a secure mental hospital. This seems rather disturbing. Some people may feel that the risk to others outweighs Timothy's right to liberty. However, they may also feel that way about a sane person who was discovered to have the 'extreme violence' brain structure (particularly if that person was their neighbour, or their child's teacher or babysitter).

Conclusion

This article contested the claim that social-protection approaches to criminal justice that do not rely on the retributive conception of moral responsibility, remove offenders from the "sphere

[12] See J Callender, *Free Will and Responsibility: A Guide for Practitioners* (OUP, Oxford 2010), chapter 8.

of justice" and "rights" altogether. It argued that offenders would still be protected by non-retributive considerations of justice, in virtue of the fact that they are persons. It focused on an analogy with those who are non-responsible due to mental illness, as this example has been invoked by critics of moral responsibility scepticism, but, in fact, supports the claim that moral and legal rights should (and to some extent already are) accorded to non-responsible individuals, based on their personhood. Traditionally, punishment theorists have often wanted to draw a very sharp distinction between sane and insane law-breakers. This may have been motivated by the poor treatment that people with mental health problems have historically received. These theorists did not want sane offenders to be treated equally badly. However, the treatment of both types of offender would be improved if we focussed on the need to respect personhood and the principles of justice that apply to all law-breakers.

References

Caruso, G. (2019). The public health-quarantine. In Dana Nelkin and Derk Pereboom (eds.), Model Oxford Handbook of Moral Responsibility. New York: Oxford University Press (forthcoming).

_____ (2018). Skepticism about moral responsibility, The Stanford Encyclopedia of Philosophy (Spring 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/skepticism-moral-responsibility/>.

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

_____ (2001). The Abolition of Punishment. Suffolk University Law Review, 35(257): 1–15.Crone, D. L., & Levy, N. L. (2018). Are free will believers nicer people? (Four studies suggest not). Social Psychological and Personality Science. DOI: 1948550618780732.

D Dennett, 'My Brain Made Me Do It' (2011), *Max Weber Lecture Series.* Retrieved 30 September, 2011 from http://cadmus.eui.eu/bitstream/handle/1814/16895/MWP_LS_2011_01.pdf?sequence=1.

Dennett, D. and Caruso, G. (2018). Just Deserts: Can we be held morally responsible for our actions? Yes, says Daniel Dennett. No, says Gregg Caruso. Caruso, G. D., & Dennett, D. C. (2018). Just Deserts: Can we be held morally responsible for our actions? Yes, says Daniel Dennett. No, says Gregg Caruso.

Duff, A., & Duff, R. A. (2001). Punishment, communication, and community. Oxford University Press, USA.

Focquaert, F. (2014). Mandatory neurotechnological treatment: ethical issues. Theoretical medicine and bioethics, 35(1), 59-72.

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. Philosophical Transactions of the Royal Society B: Biological Sciences, 359(1451), 1775.
Kozlowski, L. T., & Sweanor, D. (2016). Withholding differential risk information on legal consumer nicotine/tobacco products: the public health ethics of health information quarantines. International Journal of Drug Policy, 32, 17-23.

T Honderich, T. (1984). Punishment: The Supposed Justifications.Middlesex: Penguin Books.

I Kant, H Paton (tr) (1948). The Moral Law: Groundwork of the Metaphysic of Morals. London: Routledge.

Levy, N. (2011). Hard luck: How luck undermines free will and moral responsibility. Oxford University Press on Demand.
Morris, S. (2018). The implications of rejecting free will: An empirical analysis. Philosophical Psychology, 31(2), 299-321.

Lewis, C. (1953). The humanitarian theory of punishment. Res Judicatae 6, 224.

McCloskey, H. (1972). A Non-Utilitarian Approach to Punishment. In G Ezorsky (ed), Philosophical Perspectives on Punishment. Albany: State University of New York Press.

Morse, S (2008). Thoroughly modern: Sir James Fitzjames Stephen on criminal responsibility. Ohio State Journal of Criminal Law 5, 505.

Moore, M. S. (2010). Placing blame: A theory of the criminal law. Oxford University Press, USA.

Pereboom, D. (2018). Incapacitation, Reintegration, and Limited General Deterrence. Neuroethics, 1-11.

_____ (2014). Free will, agency, and meaning in life. Oxford: Oxford University Press.

_____ (2013). Free will skepticism and criminal punishment. In Thomas Nadelhoffer (ed.), The Future of Punishment, pp.49-78. New York: Oxford University Press.

_____ (2006). Living without free will. Cambridge University Press.

_____ and Caruso, G. (2002). Hard-incompatibilist existentialism: Neuroscience, punishment, and meaning in life.

Nadelhoffer, T and Wright, J. (2017). Humility, free will beliefs and existential angst. In Caruso, G., & Flanagan, O. (Eds.). (2017). Neuroexistentialism: Meaning, morals, and purpose in the age of neuroscience .Oxford: Oxford University Press.

J Rawls. (1999) A theory of justice (2nd Ed) Oxford: Oxford University Press.

Shaw, E. (2014). Direct brain interventions and responsibility enhancement. Criminal Law and Philosophy, 8(1), 1-20.

Smilansky, S. (2011). Free will, fundamental dualism, and the centrality of illusion. In R. Kane (Ed.), *The Oxford handbook of free will* (pp.425-441). Oxford: Oxford University Press.

_____ (2001). From nature to illusion. *Proceedings of the Aristotelian Society*, 101, 71-95.

_____ (2000). *Free will and illusion.* Oxford: Oxford University Press.

Strawson, P. (1962). 'Freedom and Resentment' (1962) 48 Proceedings of the British Academy, 187.

Tadros, V. (2011). The ends of harm: The moral foundations of criminal law. OUP Oxford.

Vilhauer, B. (2013). Persons, punishment, and free will skepticism. Philosophical Studies, 162(2), 143-163.

Vincent, N, (2011). Capacitarianism, Responsibility and Restored Mental Capacities. In in B van den Berg and L Klaming (eds). Technologies on the Stand. Legal and Ethical Questions in Neuroscience and Robotics Wolf Legal Publishers, Nijmegen).

Waller, B. N. (2017). The Injustice of Punishment. Routledge.

Wood, A. and O'Neill, O. (1998). Kant on Duties Regarding Non-Rational Nature. In Proceedings of the Aristotelian Society 72(1), 211.

Brief Biography

Elizabeth Shaw is a lecturer in criminal law and criminology at the University of Aberdeen (UK), and a director of the *Justice Without Retribution Network*. Her primary research interests are criminal responsibility, penal theory, and moral uncertainty. She has written on the moral enhancement of offenders, psychopathy and free will.

Contact Details:

Dr Elizabeth Shaw
University of Aberdeen
School of Law
Taylor Building
Old Aberdeen
AB24 3UB
+44 (0)1224 272417
eshaw@abdn.ac.uk